Game-theoretic Empathetic Parameter Estimation in Two-Vehicle Interaction

by

Yi Chen

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved April 2021 by the
Graduate Supervisory Committee:

Max Yi Ren, Chair
Wenlong Zhang
Sze Zheng Yong

ARIZONA STATE UNIVERSITY

May 2021

ABSTRACT

Recent years, there has been many attempts with different approaches to the human-robot interaction (HRI) problems. In this paper, the multi-agent interaction is formulated as a differential game with incomplete information. To tackle this problem, the parameter estimation method is utilized to obtain the approximated solution in a real time basis. Previous studies in the parameter estimation made the assumption that the human parameters are known by the robot; but such may not be the case and there exists uncertainty in the modeling of the human rewards as well as human's modeling of the robot's rewards. The proposed method, empathetic estimation, is tested and compared with the "non-empathetic" estimation from the existing works. The case studies are conducted in an uncontrolled intersection with two agents attempting to pass efficiently. Results have shown that in the case of both agents having inconsistent belief of the other agent's parameters, the empathetic agent performs better at estimating the parameters and has higher reward values, which indicates the scenarios when empathy is essential: when agent's initial belief is mismatched from the true parameters/intent of the agents.

ACKNOWLEDGEMENTS

Special thanks to Dr. Yi Ren and Dr. Wenlong Zhang for mentoring and providing resources for this research. This research is done in conjunction with Lei Zhang, Tanner Merry, Sunny Amatya and Varun Jammula.

This paper is made possible by the National Science Foundation under grant CMMI-1925403.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE Page

Chapter 1

INTRODUCTION

As robotics technology advances, Human Robot Interactions (HRI) has become an unavoidably common problem, as engineers make attempts to integrate robots into human life. In many scenario, the intent and strategies of the human and the robot may not be understood by one another during a collaborate task such as manufacturing and transportation, which leads to low efficiency, or even accidents, both of which are highly undesirable. Particularly, one of the problem that was examined that motivated this research is the problem that many autonomous vehicle researchers including Waymo encounters: when the machine agent is unable to cooperate with human drivers through a stop sign efficiently. To tackle this problem, we consider the HRI as a general-sum dynamical game with incomplete information, where each agent is uncertain about the other agent's parameter and consequently their underlying payoff functions.

There has been research devoted to find the Perfect-Bayesian Equilibrium (PBE) of such game as in Buckdahn *et al.* (2011), but it has scalability limitation with the dimensions of the state, action, reward parameter space (Sinha and Anastasopoulos (2016)), which may be a fatal flaw as interactions such as traffic is highly complex; other methods to the problem includes simplifying to complete-information game including work by Foerster *et al.* (2017); Sadigh *et al.* (2018); Kwon *et al.* (2020); Schwarting *et al.* (2019), or using the discrete parameter estimation and motion planning steps in Sun *et al.* (2018a), Nikolaidis *et al.* (2017), Peng and Tomizuka (2019) and Fridovich-Keil *et al.* (2020). In this paper, we explore the parameter estimation method, where the robot and human holds a belief on other's parameters that dictates their reward and consequently, their strategies, for its tractability. By performing the parameter estimation, both agents (human and robot) can make decision on their motion planning based on their belief on how the others might act in the interaction.

We argue that such interaction in an traffic intersection should be formulated as a game, because intuitively when an agent is making a decision, it inevitably has to take the other agent's possible action into consideration, which involves the attempts to model the other agent's payoff function; and since the payoff function is unknown, it is considered an incomplete information game. Related research in such approach about autonomous vehicle interactions can be found in Li *et al.* (2018).

Most of the existing approaches to the HRI studies can be divided into two types: empathetic and non-empathetic. The non-empathetic parameter estimation methods make the assumption that the ego agent's parameters are known by the other agent, and then the ego agent performs inference on the other agent's parameter. On the other hand, the empathetic agent allows the inconsistency between the true parameters and the estimated parameter from the other agent. We argue that the assumption behind the non-empathetic model leads to undesired outcome, and by introducing the new method "empathetic inference parameter estimation," the HRI interaction will have a better reward value, and the parameter estimation will have higher accuracy. This paper makes attempt to answer the question:

When is empathy important in a HRI?

The structuring of the empathetic model begins with the robot and the human holding a belief over what the other agent's parameter is like (e.g. aggressive or non-aggressive), then over time the belief is updated with the observed action chosen by the other agent. The key difference to the existing models[Wang *et al.* (2020)](we will refer to as non-empathetic model) is that the empathetic robot is aware of the change in human agent's belief of itself, which helps with better understanding of the human decision making by including the effect of the agents in the same environment.

Parameter estimation requires the agent to observe other's action, and update its belief based on the observation. Boltzmann distribution is often used, such as in the work by Fridovich-Keil *et al.* (2020), to describe the noisy rationality of the agent's action space, which is determined by the Q-value (cumulative reward) of the action.

The approach for obtaining the estimated parameter in our studies is to perform so called "point estimates" based on the common belief, which has the flaw of not reflecting the significance of difference in the probability distribution: when one parameter is very slightly more likely than the other, versus when one parameter dominates another; it remains to be a more cost-effective way to obtain the best course of action, as considering the uncertainty in the belief increases the dimensions of the BVP.

1.0.1 Related Work

Several studies have similar interests of studying the uncertainty within the human-robot interaction, with several different approaches:

Empathy of agent

The newly proposed game-theoretic uncertainty estimation model is based on the work on empathetic intent inference algorithm by Wang *et al.* (2020), where the double-blindness problem is addressed, and intent parameter is introduced. In this paper, we extended the research by modeling agents such that it consists of two parameters: intent (aggressiveness) and uncertainty (confidence). The comparison is then made between the empathetic and non-empathetic agent based on the inference method, whereas the previous work's baseline comparison uses a different loss function. On the side note, the action-value function and intersection problem formulation in this paper differs from the previous work, which utilized the reward value resulting from fictitious self plays.

uncertainty estimation

To address the difficulties of consideration of movements of agent when performing motion planning, Fridovich-Keil *et al.* (2020) proposed model confidence inference to incorporate a degree of confidence of a robot's modeling of other agents. It is accomplished by keeping a Bayesian belief over a parameter, which dictates the other agent's motion. The idea of modeling the uncertainty in the robot's inference is incorporated as part of our algorithm (rational/confidence parameter λ) to accommodate the issue of inaccuracy in modeling of other agents' intent.

Policy-aware Interaction

Sun *et al.* (2018b) proposed a strategy-aware interaction algorithm, with Bayesian inference to estimate the human driver's policy in the game theoretic setting, then, motion planning is performed to generate a safe action. The resulting data is then compared with real traffic data-set to evaluate data such as inter-vehicle distance, switching frequency and mean-square error between the ground-truth trajectory. The strength of this study lies in the comparison with the collected real traffic data; however, the possibility of uncertainty in the modeling of the human is not taken into consideration.

1.0.2 Main Contribution

The main contribution of this paper can be categorized into several sections:

Problem formulation

We developed an interaction determined by initial states, agent's parameter, empathy of agent and belief about the parameters, in order to systematically evaluate the strength and weakness of agents having empathy.

Such interaction is built modularly with Python code, with agent's inference, decision and dynamical models easily extendable for future development. The repository is attached in the supplementary content section.

Empathy

We define the agent's empathy in contrast to the non-empathetic approaches where the ego agent's parameter is assumed to be a common knowledge. The proposed empathetic agent is tested and verified in an uncontrolled intersection multi-agent interaction. Through the case studies where the agents' initial common belief are inconsistent from the actual true parameters/intent, it is shown that empathy leads to significantly higher reward values.

Mapping of Intent and Action-value Function

We developed a method for updating the belief with a set of action-value functions, using an algorithm based on Bayesian update. Since the set of parameters/intent may not map to action-value function in an one-to-one fashion, this algorithm is helpful for obtaining the Bayesian updates of the common belief in a fast-paced interaction.

The paper is divided into the following: Methods portion addresses the mathematics modeling that are used in our empathy studies; Implementation documents the detail of the algorithm of how the mathematics are used as well as the selection of hyper-parameters; Case studies go through the hypothesis of such scenarios that empathy plays an important role in the defined intersection interaction; and finally, results and analysis showcase the trajectory and the agents' reward value from the experiments addressed in the case study for evaluating the strength of our proposed modeling of empathy.

Chapter 2

METHODS

The methods to the parameter estimation along with the motion planning is documented in this section. Details of how the algorithm incorporates the newly proposed methods is also explained.

2.0.1 Notation

We denote number of agents as N, and each agent has the same set of action U, reward parameter set Θ, noise parameter set Λ, initial belief choice β_0 and empathy choice l_i. Each agent can be modeled using the combination of reward parameter(intent) and noise parameter, denoted as $\beta_i =< \theta_i, \lambda_i >$, for agent i. Since in this paper, we have two agent in the interaction, the parameter pair at time $t = k$ can be denoted as $\beta(k) =< \beta_i, \beta_{-i} >_k$, where $\lambda_i \in \Lambda$ and $\theta_i \in \Theta$. Note that u, Θ, Λ and in term B are discrete sets, for the efficiency of Bayesian update, where it is done by iterating through the parameter sets and action sets. The agents also share instantaneous reward function f, terminal reward function c, dynamic model h and finite time horizon [0, T]. The interaction is assumed to be discrete-time due to the consideration of computation speed. We then express the multi-agent interaction parameterized by $s_i =< x_0, p_0, \theta^*, l >_i$, where x_0 is the initial state, θ^* is the agent's true parameter and p_0 is the initial common belief matrix.

In our studies, the value of Λ set and Θ set are selected to be (0.5, 0.1) and (5, 1) respectively, based on the formulation of the agent's loss function and the resulting action-value, which is explained in later part of this section.

2.0.2 Bayesian Inference

At the start of every interaction scenario, each agent holds a belief about other agent's parameter. Over time, the belief is updated with Bayesian update:

$$Pr(\beta = \beta_i | D(k)) = \frac{Pr(u(k)|x(k);\beta)Pr(\beta|D(k-1))}{\sum_{\beta' \in \mathcal{B}} Pr(u(k)|x(k);\beta')Pr(\beta'|D(k-1))}, \tag{2.1}$$

where k is the the time step, and the action probability $Pr(u(k)|x(k)$ is modeled as Boltzmann's distribution for modeling the noisy-rational agents

$$Pr(u_i|x;\beta) = \frac{\exp(\lambda_i Q_i(x,(u_i,u_{-i});\theta))}{\sum_{u' \in \mathcal{U}} \exp(\lambda_i Q_i(x,(u'_i,u_{-i});\theta))}. \tag{2.2}$$

Here, Q_i is the action value for agent i at state x with intent θ while holding the other agent's action u_{-i} fixed, and λ_i is the noise parameter for agent i. Note that as λ approaches 1, the probability of action with highest Q value approaches 1. Example of action probability distribution with different λ values is shown in figure 2.1.

To address the potential issue when the prior becomes zero which causes the probability to get stuck at zero, we use the re-sampling equation:

$$Pr_k(\beta) = (1-\epsilon)Pr_k(\beta) + \epsilon p_0(\beta), \tag{2.3}$$

where $p_0(\beta)$ is the initial belief over the set of parameters.

Decoupled intent and Q-value

If the action-value function and the parameter arre decoupled (asymmetric; not one-to-one), in such case a parameter pair (β, $\hat{\beta}$) could map to multiple action-values Q, then the belief updates can be formulated by starting with the Q function:

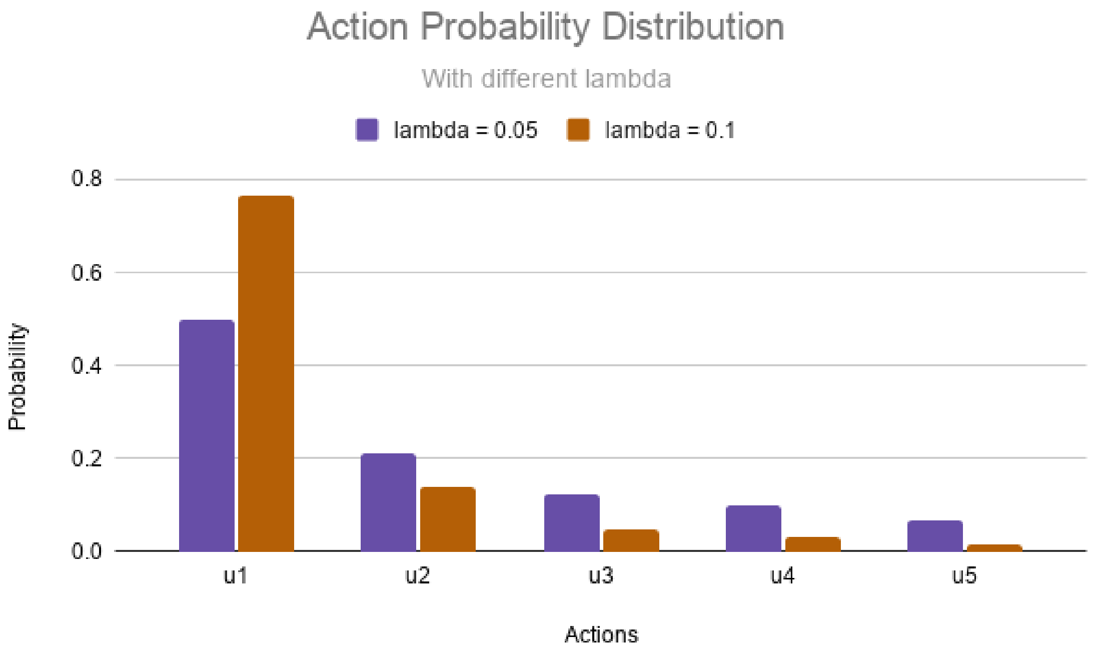

Figure 2.1: Example of Boltzmann Action Probability Distribution Given Different λ, the Noise Parameter; the Q Values to Calculate the Distributions are Taken Example from the Simulation.

$$Pr(Q_1, Q_2|D(k)) = \frac{Pr(u_1(k), u_2(k)|x(k); Q_1, Q_2)Pr(Q_1, Q_2|D(k-1))}{\sum_{(Q_2', Q_1') \in \mathcal{Q}^2} Pr(u_1(k), u_2(k)|x(k); Q_1', Q_2')Pr(Q_1', Q_2'|D(k-1))}, \tag{2.4}$$

which is the probability of the Q-function pair given the observation. Then we have the Bayesian updates for the common belief over agent's parameters:

$$Pr(\beta_1, \hat{\beta}_2|D(k)) = \sum_{(Q_2', Q_1') \in \mathcal{Q}^2} Pr(\beta_1, \hat{\beta}_2|Q_1', Q_2')Pr(Q_1', Q_2'|D(k)), \tag{2.5}$$

where

$$Pr(\beta_1, \hat{\beta}_2|Q^2) = \frac{P(Q^2|\hat{\beta}_2, \beta_1)P(\beta_1, \hat{\beta}_2|D(k-1))}{\sum_{(\beta_1', \hat{\beta}_2') \in \Lambda^2 \times \Theta^2} P(Q^2|\hat{\beta}_2', \beta_1')P(\beta_1', \hat{\beta}_2'|D(k-1))} \tag{2.6}$$

with Q^2 representing $(Q_1,\, Q_2)$ and $Pr(Q_1, Q_2|\beta_1, \hat{\beta}_2)$ is given by one divided by number of Q function pair given the β pair.

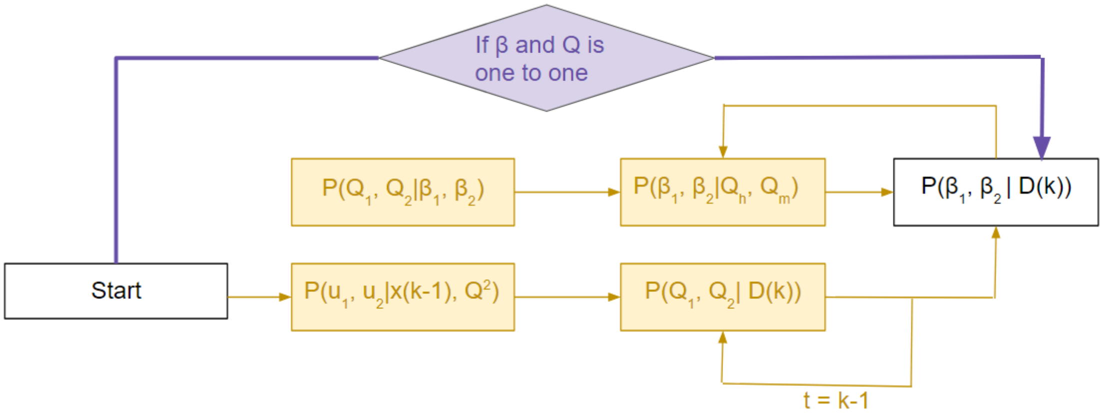
If β and Q is one to one
$P(Q_1, Q_2|\beta_1, \beta_2)$
$P(\beta_1, \beta_2|Q_h, Q_m)$
$P(\beta_1, \beta_2 | D(k))$
Start
$P(u_1, u_2|x(k-1), Q^2)$
$P(Q_1, Q_2| D(k))$
t = k-1

Figure 2.2: Flow of the Inference Algorithm in Different Settings: Whether Q and β are One-to-one. Starting Point is When Agents Receive the Latest Observation on Fellow Agents' Action and Ready to Evaluate and Update Its Belief.

The algorithm with the above equations implemented can be represented as shown in Fig. 2.2. Notice that if parameters map to single Q function, then the enclosed calculation portion can be simplified to Eq. 2.1. In our experiments, we assume that the parameter only maps to one action-value function, but the math is implemented for ease of extension in the future.

Motion prediction

The motion prediction is built for analysis purposes, to visualize the future trajectories of the vehicle based on its possible actions. It can be used to calculate the probability of an agent at certain future position and speed at $t = k+1, k+2, ...$ Note that, however, the space of the trajectory grows with the number of actions to the power of number of future time steps, $(length(U)^T)$, where T is the future look-ahead horizon. The probability of agent's being at state x at time $k+1$ is given by:

$$
\begin{aligned}
Pr(x(k+1)|Q_1,Q_2) = \sum_{x(k)\in\mathcal{X},u(k)\in\mathcal{U}} & Pr(x(k+1)|x(k),u(k)) \\
& Pr(u_1(k),u_2(k)|x(k);Q_1,Q_2) \\
& Pr(x(k)|Q_1,Q_2),
\end{aligned} \tag{2.7}
$$

where $x(k+1)$ is the state of the agents after 1 second of using certain action u_1 and u_2, and

$$
Pr(u_1(k),u_2(k)|x(k);Q_1,Q_2) = Pr(u_1(k)|x(k);Q_1)Pr(u_2(k)|x(k);Q_2), \tag{2.8}
$$

is the joint probability of actions that leads to the state $x(k+1)$, and $(Q_H, Q_M) \in \mathcal{Q}^2$. Then

$$
Pr(x(k+1)|D(k)) = \sum_{(Q_1,Q_2)\in\mathcal{Q}^2} Pr(x(k+1)|Q_1,Q_2)Pr(Q_1,Q_2|D(k)). \tag{2.9}
$$

is the probability of agents being at state x at time $k+1$ given the past observation $D(k-1)$.

The future states $x(k+1)$ are derived using the simple 1D dynamical model:

$$
\begin{bmatrix} \dot{d}_i(t) \\ \dot{v}_i(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} d_i(t) \\ v_i(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u_i(t). \tag{2.10}
$$

To summarize the Bayesian parameter inference, the agent's belief is updated starting with obtaining the Q value of the action taken from the last time step $k-1$ (Eq. 2.2), which is then used for evaluating the agent's likelihood of using each of the possible parameters (Eq. 2.1). Decoupled action-value and parameter version of the update as well as the motion prediction formulation are addressed. Next up we take a look at the different formulations of the belief matrix.

2.0.3 Action-value approximation

To obtain the action-value on the fly in time-sensitive interaction, solving a time consuming optimization problem online is not very effective, therefore a method to quickly obtain the value given a state-action pair is needed. To accomplish this, deep neural network (DNN) is used to approximate solutions to Hamilton-Jacobi-Issac (HJI) equation formulated for the boundary-value problem(BVP) in the general-sum dynamic game as addressed in the paper Chen *et al.* (2020), and from there we can estimate the action-value at any state within the intersection state space.

Boundary-value Problem (BVP)

While we construct the interaction to be discrete-time dynamic games, the boundary-value problem (BVP) is formulated from the HJI equations with Pontryagin's maximum principle (PMP) for a continuous differential game to better capture the dynamics of the states. The value function for the continuous differential game is given by

$$V_i(X_1, X_2, t) = F_i(T) - \int_t^T L_i(X_1, X_2, u_i, \theta_1, \theta_2)\, dt, \tag{2.11}$$

where the dynamic loss is

$$L_i(X_1, X_2, u_i, \theta_1, \theta_2) = u_i^2 + f_{collision}(x_1, x_2, \theta_1, \theta_2) \tag{2.12}$$

and the final loss

$$F_i(T) = \alpha x_i(T) - (v_i(T)) - v_i(t = 0))^2 \tag{2.13}$$

Notice that the final loss is designed such that the vehicle has incentives to keep its initial speed ($v(t = 0)$, with α being the hyperparameter. Then, the Humailton-

Jacobi-Issac (HJI) equation can be derived from the value function (eq. 2.11) as follows:

$$H_i = \left[\frac{\partial V_i}{\partial X_1}\right]^T f_1(X_1, u_1, t) + \left[\frac{\partial V_i}{\partial X_2}\right]^T f_2(X_2, u_2, t) - L_i(X_1, X_2, u_i) \tag{2.14}$$

Finding the optimal control input u^* maximizing the Hamiltonian:

$$\begin{aligned} 0 &= \left[\frac{\partial V_1}{\partial X_1}\right]^T f_1(X_1, u_1^*, t) + \left[\frac{\partial V_1}{\partial X_2}\right]^T f_2(X_2, u_2^*, t) - L_i(X_1, X_2, u_1^*) \\ 0 &= \left[\frac{\partial V_2}{\partial X_1}\right]^T f_1(X_1, u_1^*, t) + \left[\frac{\partial V_2}{\partial X_2}\right]^T f_2(X_2, u_2^*, t) - L_i(X_1, X_2, u_2^*) \end{aligned} \tag{2.15}$$

where $\frac{\partial V_i}{\partial X_j}$ can be represented as "costates" γ_{ij}, i.e. $\frac{\partial V_1}{\partial X_2} = \gamma_{12}$. Finally, by utilizing the Pontryagins Maximum Principle (PMP), the above HJI equations are solved as a boundary value problem (BVP) yielding V^* and ∇V^* for value approximation.

$$\begin{aligned} &\dot{x^*} = h(x^*(t), u^*(t)) \\ &x^*(0) = x_0 \\ &\dot{\gamma^*}_i = -\nabla_x H_i(x^*, u^*, \gamma_i^*(t); \theta) \\ &\gamma_i^*(T) = -\nabla_x F_i(x^*(T); \theta) \\ &u_i^*(t) = \arg\max_{u_i \in \mathcal{U}} H_i(x^*, u_i, \gamma_i^*(t); \theta), \\ &\dot{V}^*(x^*, t; \theta) = L(x^*, u^*; \theta), \\ &V^*(x^*, T; \theta) = F_i(x^*(T); \theta)\ \forall i = 1, ..., N, \end{aligned} \tag{2.16}$$

where γ is the costate, h is the dynamical model and u^* is the optimal solution, and V is the value. The Eq. (2.16) is then solved using a standard BVP solver from Kierzenka and Shampine (2001).

Value Approximation

Since the space spanned by the states and parameters can be very large, it is impractical to solve for every possible point in an multi-agent interaction, which makes value approximation a more efficient technique for obtaining the action-value given any state during the simulation. After solving the BVP given x_0 and θ, we obtain V^* and $\nabla_x V^*$ for a given combinations of parameters $\theta = (\theta_1, \theta_2) =$ (NA, NA), (NA, A), (A, NA), (A, A), which are then used for approximating values for the inference portion of the multi-agent interaction by solving the learning problem:

$$\min_w \sum_{(x,t,V^*,\nabla V^*)\in\mathcal{D}_v} (||\hat{V}(x,t;\theta,w) - V^*||^2 + C||\nabla_x \hat{V}(x,t;\theta,w) - \nabla_x V^*||^2). \tag{2.17}$$

Which provides the action value function:

$$Q_i(x,t,\theta,u_i) = \hat{V}(x,t) - L_i(x,u_i,\theta)\delta t \tag{2.18}$$

In theory, when defining action-value, it should be parameterized by both the physical state and the belief state. For simplification of the optimal control problem, the belief state is taken as the approximated parameter using the point estimates, instead of considering the probability distribution from the Bayesian inference algorithm.

It is noted that since the BVP is set up as a game, the other agent's action is needed when obtaining an action-value from the value network. To make the interaction possible in a real-time manner, the other agent's action is taken from the observed action from time $t = k - 1$. Intuitively, this approach is plausible since human's instantaneous decisions are made based on what is observed in the past and from experience.

More BVP formulation and value network details can be found in Chen *et al.* (2020), while this paper focuses more on analysis of the empathetic inference.

Chapter 3

IMPLEMENTATION

This section goes through the implementation of the equations addressed in the previous section for the inference and decision/motion planning algorithms, as well as the setup for the simulation for experimentation and case studies.

3.0.1 Belief matrix and update

Over the duration of the interaction, the agents keep a belief over the parameters of the other agent in the form of a matrix. The formulation of the belief matrix differs depending on the type of agent, but they are updated based on the observed state and action using equation 2.1 with the same method. At the start of an interaction, the initial belief $p_0(\beta)$ is generated from the initial belief as follows:

Algorithm 1: Algorithm for initializing belief matrix

```
Given belief parameters $p_0 = (\hat{\beta}_1, \hat{\beta}_2)$;
Define weight $w$ ;
Create a matrix with entries of ones;
For ($\beta_1$, $\beta_2$) in ($B_1$, $B_2$):
if parameter ($\theta$ or $\lambda$) matches with belief then
    Multiply the entry by $w$
else
    Multiply the entry by $(1 - w)/(len(B) - 1)$
end
Assert summation of all entries equals to one;
Result: A matrix with probability distribution of belief in parameter pair
```

To give an example on the structure of the common belief, in the case of empathetic agent, given initial parameters ($\beta_1 = (NA, NN), \beta_2 = (NA, NN)$) and weight $w = 0.8$, set of 2 theta and 2 lambda ($\beta = \Theta X \Lambda$), the resulting initial common belief would look like:

Example 1 *Example of initial belief matrix given the parameters (rows: agent 1, columns: agent 2):*

(θ, λ)	*(NA, NN)*	*(NA, N)*	*(A, NN)*	*(A, N)*
(NA, NN)	*0.4069*	*0.1024*	*0.1024*	*0.0256*
(NA, N)	*0.1024*	*0.0256*	*0.0256*	*0.0064*
(A, NN)	*0.1024*	*0.0256*	*0.0256*	*0.0064*
(A, N)	*0.0256*	*0.0064*	*0.0064*	*0.0016*

Each entry in the matrix stands for the belief over the likelihood of agents' parameter being (β_1, β_2) given the past observations, denoted as $P((\beta_1, \beta_2)|D(k-1))$. When updating the belief over the likelihood each parameter pair at each time step, each parameter pair is analyzed given the observed action and state, using Eq. 2.2.

3.0.2 Parameter estimation and Motion planning

At each time step, after the parameter estimation step (Bayes' update/inference), each agent performs motion planning based on the information in hand including the approximated parameters and the observed state. The point-estimate method extracts the parameter with the highest probability mass from the belief matrix in order to evaluate the best course of action.

The noisy-rational action distribution model, or Boltzmann's distribution, is used to depict the agent's decision making model. When choosing an action u_k from a discrete set of actions U, each action $u \in U$ is evaluated given the state of the game and parameter pair $(\beta_i,\, \hat{\beta}_{-i})$, while fixing the other agent's action to be the observed action $u_{-i,k-1}$. The method to obtain the parameter pair from the common belief that is used for evaluating actions defers for empathetic and non-empathetic agent in the following way:

Empathetic estimation

Empathetic agents use the point-estimated parameter pair $(\hat{\beta}_i, \hat{\beta}_{-i})_k$, or simplified as $\hat{\beta}(k)$, from the entire belief matrix such as the one shown in table 1 to evaluate the actions by enumerating over the action set, which can be expressed as

$$\hat{\beta}(k) = \arg\max_{\beta \in \mathcal{B}} p_k(\beta), \tag{3.1}$$

notice that the estimates $\hat{\beta}(k)$ is obtained from the argmax of the entire common belief matrix.

Non-empathetic estimation

For non-empathetic agent, the agent's estimates of the parameter of the other agent is obtained by looking at the partial common belief matrix by holding the agent's own true parameter is fixed. In other words, the estimation is done by extracting the β from the common belief conditioned on the ego parameter's true parameter:

$$\tilde{\beta}_{-i}(k) = \arg\max_{\beta_{-i} \in \mathcal{B}_{-i}} p_k(\beta_{-i}|\beta_i^*), \tag{3.2}$$

where $\tilde{\beta}_{-i}(k)$ is the non-empathetic agent i's estimate of other agent's (agent -i) parameter at time k, and β_i^* is the agent i's true parameter.

Motion planning

Following the parameter estimation step, the agents perform motion planning using the information in hand. Since the action-value for the 2 agents are coupled together as V1 and V2, we fix the other agent's action as $u_{other} = u_{k-1,other}$ when enumerating over the action set. Each resulting action pair $(u_{i,ego}, u_{k-1,other})$is then used as an input for the Q function to obtain the action-value and subsequently the Boltzmann distribution. Then, the best action is obtained using

$$u_i^* = \arg\max_{u_i \in \mathcal{U}} Q_i(x, (u_i, u_{-i}^\dagger); \hat{\theta}). \tag{3.3}$$

To summarize, the two types agents can have very distinct action-value for a given state, due to the difference in parameter belief, thus resulting in choosing different actions. Note that the action is chosen based on the highest probability mass calculated using eq. 2.2 (Boltzmann's distribution), for the ability to replicate the simulation results, as opposed to choosing action randomly according to the probability distribution.

3.0.3 Uncontrolled Intersection

The environment that the agents will be tested in is a single-lane, one-way uncontrolled intersection, meaning there is no stop signs or traffic light, where the total length of the lane is 70m, with the zone of the intersection zone spanning from 34.25 to 35.75m. This design is chosen that it can be fitted to many different scenarios in the traffic scenes by simply changing the coordinate of the space.

Each agent have a starting position x_0 with the range $d = [15, 20]$, which is the distance from the start of the road, i.e. the larger the x_0 the closer the agent is to the intersection. The starting speed v_0 has the range of $v_0 = [18, 25]$.

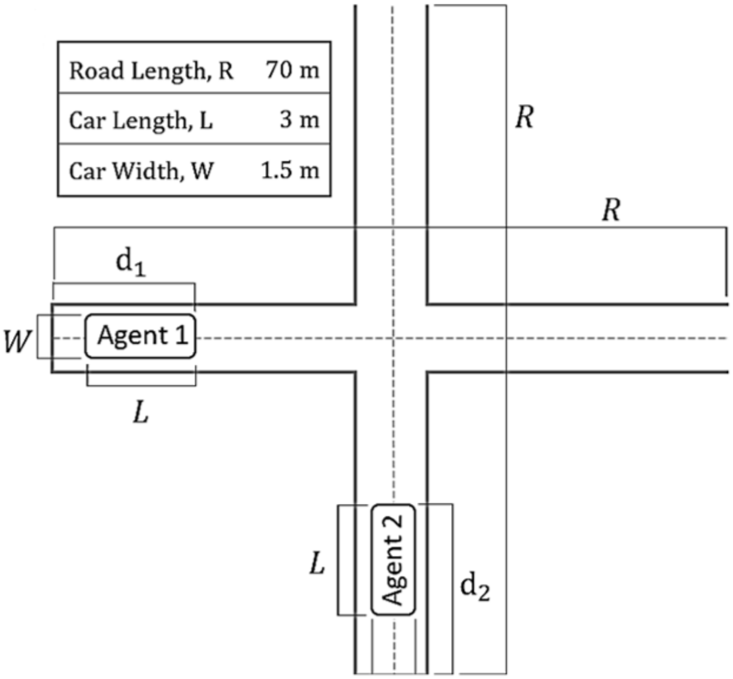

Figure 3.1: Uncontrolled Intersection Setup: Single Lane, One Way, No Stop Sign nor Traffic Light. d is the Initial Positions of the Agents.

The agent's loss/reward function is modeled differently around the intersection, according to their types: aggressive and non-aggressive. The non-aggressive agent has a larger "collision zone", starting from 31.25 to 38.75m, while the aggressive agent has a smaller zone from 34.25 to 38.75m; notice that the car has to fully exit the intersection to avoid collision. This is to model the comfort level of the agent when they are approaching another vehicle. The loss function is shown as in the figure 3.2.

Other parameters of the intersection include car size, vehicle speed and acceleration/deceleration capabilities, which are taken example from the real world average: 3 by 1.5 meter for the car, 0.1 to 40 m/s for the vehicle speed and -5 to 15 m/s^2 for the vehicle inputs. Since there is a limitation of how the actions are evaluated by their action-value function, the vehicle inputs are given as a discrete set U, called "action set," with entries (-5, 0, 3, 7, 10), to make it computationally possible to make decisions on the fly.

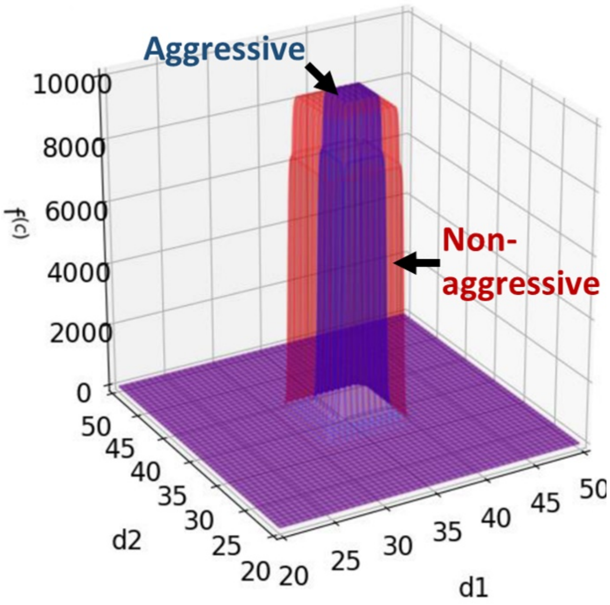

Figure 3.2: Collision Loss Function Construction for NA and A Agents: NA Agent's Loss is Pictured in Red, While A Agent's Loss is in Purple.

3.0.4 Simulation

In the simulation, the belief updates, motion planning addressed above are run through iteratively. The simulation workflow is demonstrated in Fig. 3.3: From the initialization with the given initial conditions, to termination upon satisfaction of terminal conditions (vehicle passing intersection).

Note that the simulation is in discrete time and the agents are given discrete set of actions for implementing the Bayes' update, and due to the limitation on finding a simultaneous Nash Equilibrium in a large action space, initial action is assumed to be zero. With the intersection case implemented, next up the case studies for testing the strength of our empathetic model are addressed.

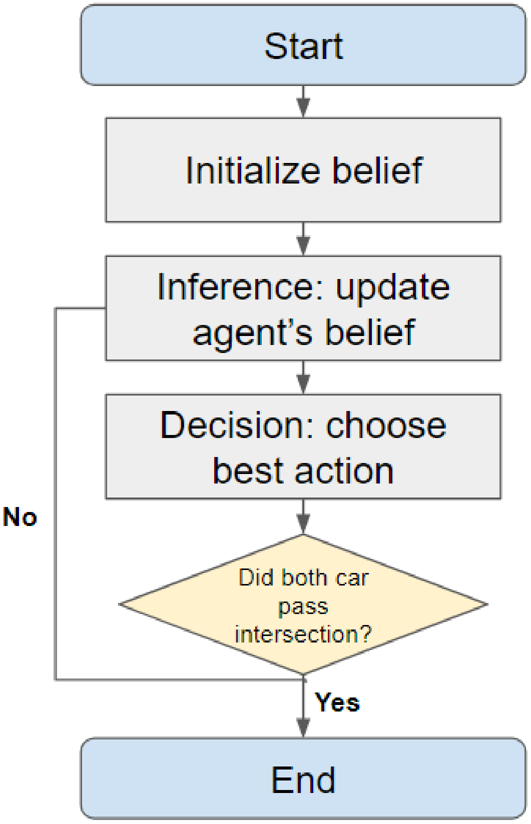
Start
Initialize belief
Inference: update agent's belief
Decision: choose best action
Did both car pass intersection?
No
Yes
End

Figure 3.3: Simulation Work Flow: Agents Update Belief and Make Decision Based on the Belief Iteratively Until the Interaction Ends.

Chapter 4

CASE STUDIES

From the method and implementation sections, it is easy to see that the empathetic inference is more computation heavy; therefore, an effective way for verifying whether such higher cost is worthwhile is needed. In this section, we propose some difficult scenarios to see whether our empathetic modeling improves the outcome of the interactions. The experimental interaction variables include changing of the initial beliefs, starting position and speed.

4.0.1 Starting Position and Speed

As discussed in the implementation section, each agent has a range of starting position x_i and speed v_0. The interaction is tested for agents starting from the grid of X_1 x X_2 while fixing the starting speed to reduce the dimension of the experiment space. Larger x_i (closer to collision zone) will put the agent's ability to quickly infer the agent's parameter to test, as there is less time to converge to the right policy before coming into contact with the other agent.

4.0.2 Initial conditions: beliefs

The parameter set of agents $\beta^* \in B$ is composed of combination of intent and noise parameters, creating parameters combinations (NA, NN), (NA, N), (A, NN), (A, N) for each agent, where "A" stands for aggressive, "N" stands for noisy and so on. Each of them represent a variable that affects the equations introduced in the method, for instance, NA (non-aggressive) represents "5" that gets plugged into the value function in Eq. 2.11, and noise parameter "NN" (non-noisy) represents "0.1" which gets plugged into the λ in Boltzmann equation in Eq. 2.2.

To test the strength of our proposed parameter estimation model, we let the agents start with incorrect initial belief over the other agent's parameter.

NA agents with A beliefs

At the beginning of an interaction, each agent believes the other agent to be aggressive ($\hat{\theta}$ = A), while their own ground truth parameter being non-aggressive (θ^* = NA). Intuitively, this false believe if not corrected, can lead to inefficient interaction, where both agents are hesitant to pass the intersection first, since they assume the aggressive agent will take the chance as they care less about close encounters.

A agents with NA beliefs

Oppose to the above setting, the agents in this case believe that the other agent is non-aggressive, while their own parameters are aggressive. When agents unknowingly assume others to be non-aggressive, intuitively, the agent may make decision that dangers both parties since it believes that it can pass the intersection first.

4.0.3 Agents with consistent initial belief

Similar settings to the inconsistent initial beliefs scenarios, here the agents are given an initial belief over the other agent's parameters but is consistent with the true parameter. The agent's ability to correctly model an agent is tested in this case, which also provides a baseline comparison to the above scenario, showing the difference in rewards when agents have correct versus incorrect beliefs.

The scenarios in this section lead us to designing the following hypotheses:

Hypotheses

- Empathy in the uncontrolled intersection leads to higher reward when agents have incorrect initial belief. We let agents' choice of empathy be l, and e, ne stands for empathetic and non-empathetic, p_0 be the initial belief, v be the reward of the agents, $s =< x_0, p_0, \theta, l >$ be the set of parameters,then if we have $l_1 = (e, e)$ and $l_2 = (ne, ne)$, $\theta_1^* = \theta_2^* \neq p_0$, there exists $x_0 \in X_0$ such that $v(s_1) > v_2(s_2)$.

- Empathy leads to higher reward when agents have consistent initial belief; Same formulation applies except for $\theta_1^* = \theta_2^* = p_0$.

The hypothesis is verified in a discrete-time simulation of an uncontrolled intersection with two agents, where the two agents attempt to infer other agent's parameter and score the highest reward. Results of the case studies including reward value, evolution of parameter belief and trajectories are shown and discussed in the next section.

Chapter 5

RESULTS AND ANALYSIS

In this section, the results from the settings discussed in the case study portion is shown and analyzed: the reward comparison yields a more case by case study between the initial conditions, while the trajectories present a more macroscopic view of the performance of the agent type. Policy choices and belief dynamics add another layer to help analyzing the results from reward and trajectories.

5.0.1 Reward comparison

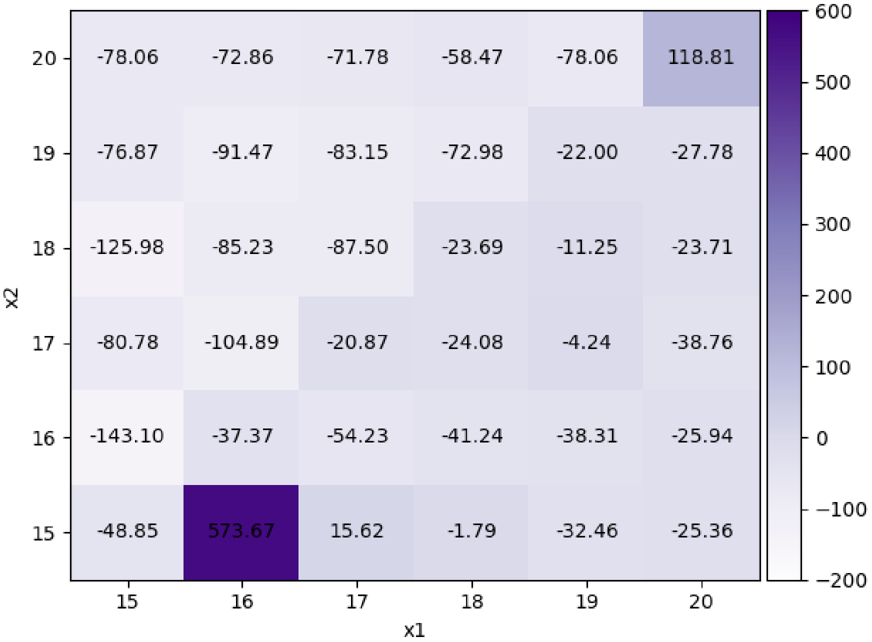

Figure 5.1: A.E.inconsistent vs A.base: Reward Difference between the Aggressive Empathetic Interaction with Inconsistent Initial Belief and Baseline Interaction ($Reward_{base} - Reward_E$). Lighter Color Represents Empathetic Agent Having Close or Higher Reward Than Baseline Performance and Vice Versa.

As addressed in the implementation section, we design a starting zone for the two agents. We run through the combination of starting positions, which is a total of 6 by 6 cases for each case study to thoroughly test all the trajectories for the social values. The velocity is fixed to 18m/s to reduce the dimension of the experiment variables. By having different starting position, we examine how NA and A agents interact in different scenarios: when the agents with empathetic and non-empathetic models start with the incorrect belief and when the agents start with the correct belief (we refer to this as "baseline"). The difference in the accumulative reward values will reflect on how well the type of agent is able to quickly update the belief on the other agent's parameters.

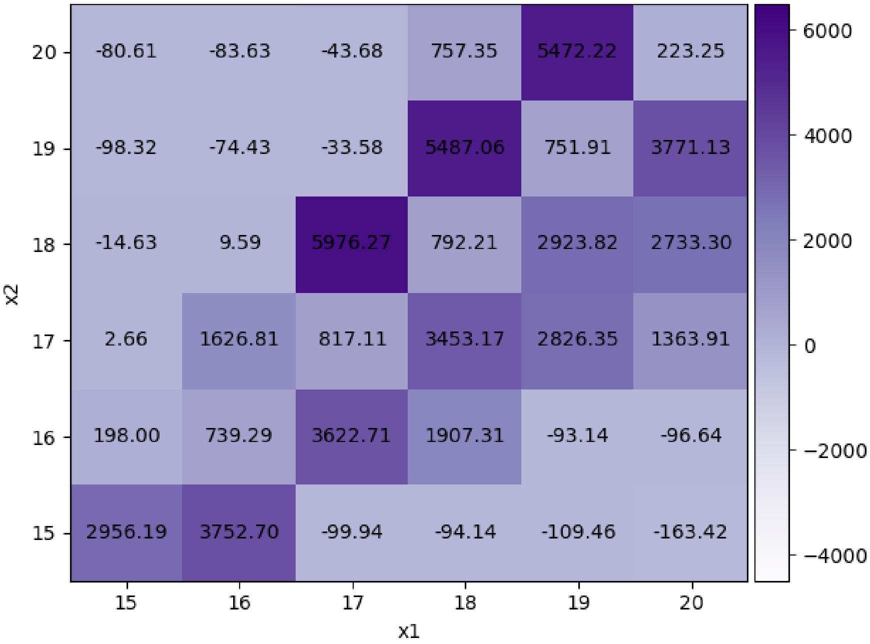

Figure 5.2: NA.E.inconsistent vs NA.base: Reward Difference between the Non-aggressive Empathetic Interaction and Baseline Interaction ($Reward_{base}$ - $Reward_E$). Lighter Color Represents Empathetic Agent Having Close or Higher Reward Than Baseline Performance and Vice Versa.

Fig. 5.1 and Fig. 5.2 compares the difference in reward values between the empathetic agents with inconsistent initial belief and the baseline interaction, where each entry is calculated as $Reward_{baseline} - Reward_E$. The empathetic interaction with inconsistent initial belief on average has higher reward by 30.694 for aggressive and deficit of 1418.85 for non-aggressive when compared to neural network baseline (with no inference).

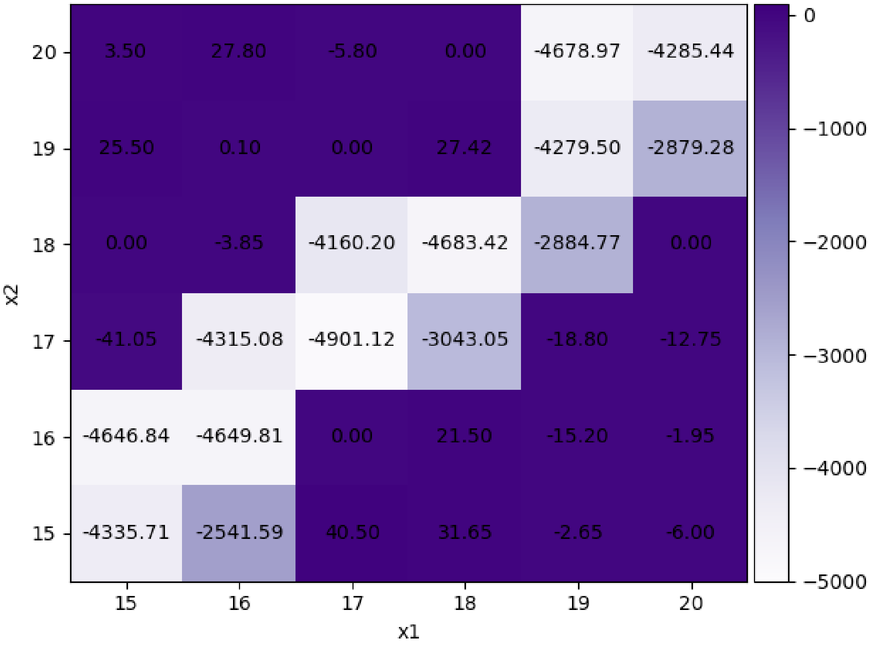

Figure 5.3: A.E vs A.NE (inconsistent): Reward Difference between the Aggressive Empathetic Interaction and Non-empathetic Interaction, with Inconsistent Initial Belief($Reward_{NE} - Reward_E$). Lighter Color Represents Empathetic Agent Having Significantly Higher Reward Than Non-empathetic Agent and Vice Versa.

Overall, the empathetic agents perform on par with the baseline cases except in the non-aggressive setting when the initial states of the agents are close together. On the diagonal part of the non-aggressive cases, having a fixed policy helped with safely navigating to avoid close encounters; however, the lower reward in baseline is due to the optimizer finding global solution outside of the action bounds then are normalized to fit in the boundaries, which makes the agents to spend slightly higher effort. It is also noted that the smaller difference in reward is indicative of the difference in effort resulted from the u^2 term in the reward value function, but the larger difference (higher than 1000) is a result from having (pseudo) collision, which yields more than -1000 reward for every time step spent in the collision scenario.

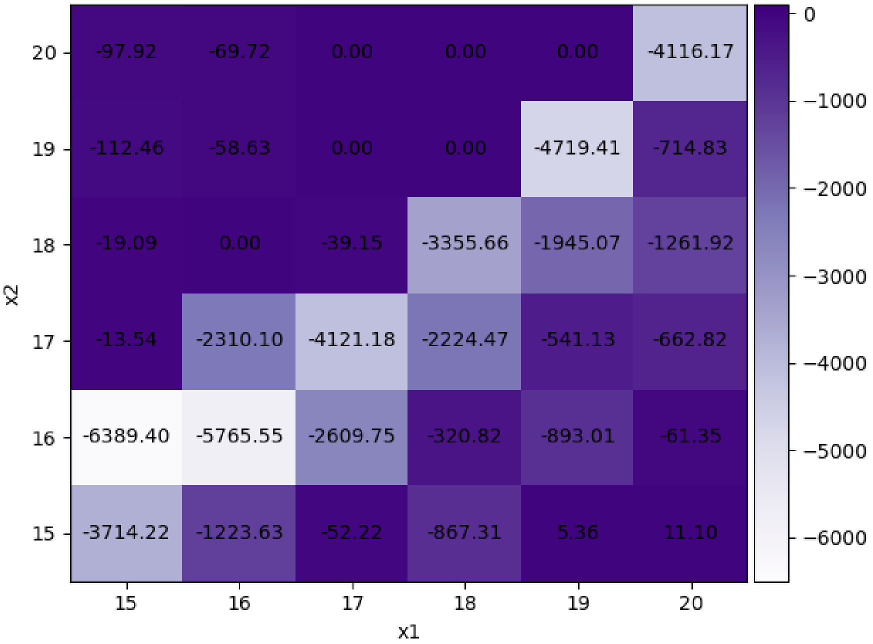

Figure 5.4: NA.E vs NA.NE (inconsistent): Reward Difference between the Non-aggressive Empathetic Interaction and Non-empathetic Interaction, with Inconsistent Initial Belief ($Reward_{NE} - Reward_{E}$). Lighter Color Represents Empathetic Agent Having Significantly Higher Reward Than Non-empathetic Agent and Vice Versa.

Fig. 5.3 and Fig. 5.4 shows the difference in accumulative reward between empathetic and non-empathetic (E vs NE) agents with inconsistent initial belief, in the case of aggressive (A) and non-aggressive (NA) interaction respectively. The negative values show the cases where empathetic agents outperform non-empathetic ones. On average, empathetic agents perform better than the non-empathetic ones by 1561.52 for aggressive, and 1340.67 for non-aggressive across the starting positions in terms of accumulative reward. In both aggressive and non-aggressive settings, the empathetic agents show significantly higher reward when the initial states are close for both agents, which may lead to a more obscure interaction.

By combining the above observations from the figures showing the difference in accumulative reward values, it is noted that empathetic interaction generally leads to better outcome in terms of the agent's reward than the non-empathetic ones, while performing on par with the baseline. To further prove the effectiveness of empathy, discussion on trajectories and policy choices are addressed up next.

5.0.2 Trajectories and policies

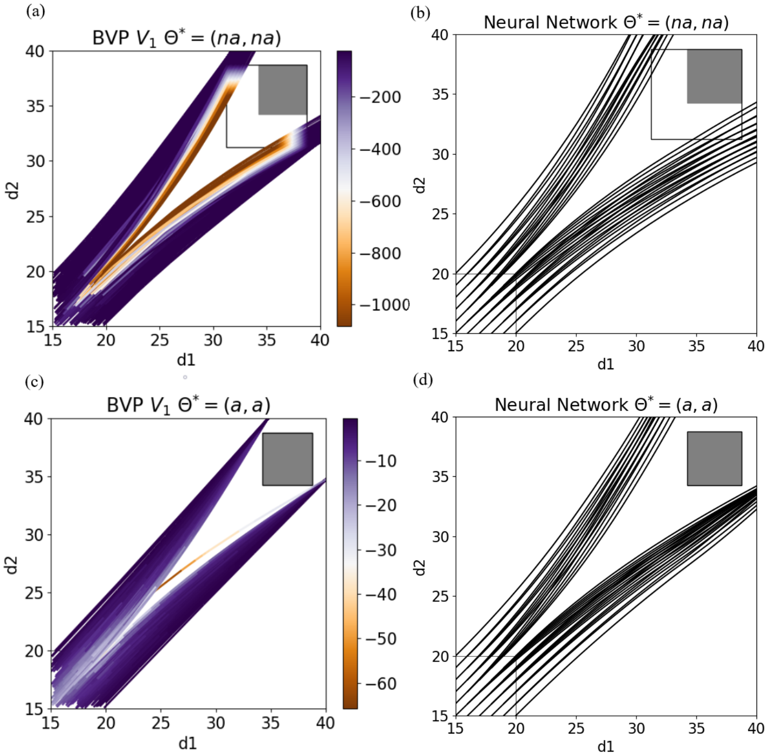

Figure 5.5: BVP Trajectory and NN Simulation Results: (a,b,c,d) are the Trajectories from BVP Solver, and (e,f) are the Trajectories from the Simulator Using the Trained Neural Network. The Color Represents the Instantaneous Rewards for the Two Agents in That Trajectory.

The resulting trajectories and the agent's policy choice in different initial belief are examined in this section. First, the trajectories from BVP solver and the trajectories by optimizing the action-value from using the trained value network (neural network, NN) without any inference updating the agents' belief over the parameters are compared, as shown in Fig. 5.5; the collision boxes and pseudo-collision boxes (for non-aggressive agents) are represented in the figures in grey and white colors respectively. Through this we validate the effectiveness of the value network in reflecting the action values given any state during the simulation. Additional trajectories in Fig 5.6 are included as baseline as part of the policy that will be used in the consistent and inconsistent tests. Note that few trajectories as shown in the figures where lines cross the dark rectangular box (collision zone), it is due to the limitation of samples used for training the network: there exist some states that are not trespassed by the BVP solver, or they are difficult to solve (i.e. when both have identical initial conditions); for the most part, neural network result performs equally as well as the BVP solutions, with the only major difference being that the NN trajectory is a one-shot whereas BVP can iteratively improve the trajectories to minimize effort versus reward, and that difference can be seen where NN trajectories are more spread out from the collision zone in order to avoid unwanted loss.

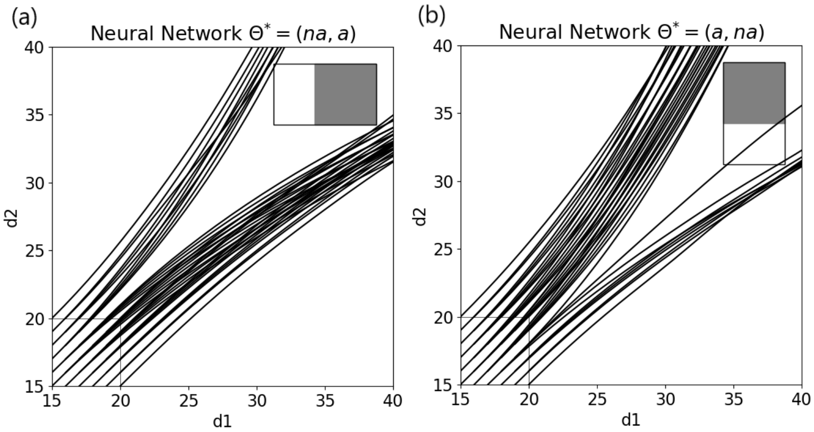

Figure 5.6: Neural Network Trajectories: Resulting Trajectories from Using Trained Value Network to Approximate the Action Values in (a) Non-aggressive Versus Aggressive and (b) Aggressive Versus Non-aggressive Settings.

Consistent Belief

The consistent belief cases test the capability for each type of agent to infer and stick to the correct belief over the fellow agent's parameters, and consequently the correct policy against the fellow agent. The capability for inferring the correct belief while avoiding collision is evaluated by presenting and analyzing the resulting trajectories and chosen policies on top of their rewards.

By comparing the consistent belief simulation trajectories shown in Fig. 5.7 to the BVP solution trajectories in Fig. 5.5, it is shown that the inference algorithm in such cases is able to correctly choose the right policy accordingly based on the value network with high resemblance in comparison to the baseline trajectories, regardless of agent types. This result provide a baseline understanding of the performance of the discrete-time inference algorithm using Bayes' updates.

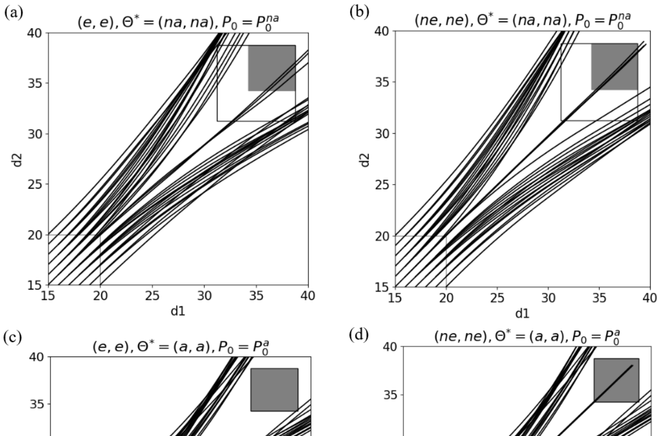

Figure 5.7: Consistent Initial Belief: When Belief Matches with Ground-true Reward Parameters. (a,b) Unknowingly Non-aggressive, (c,d) Unknowingly Aggressive. (a,c) Empathetic, (b,d) Non-empathetic.

Inconsistent Initial Belief

In the case when agents are given inconsistent initial belief, aggressive agents start with beliefs that the fellow agents are non-aggressive, and vice versa. In order for agents to efficiently pass the intersection and avoid any collision (or pseudo-collision for non-aggressive agents), having the correct belief and in turn correct policies is essential; if the agents have incorrect belief, i.e. aggressive agents believe each others to be non-aggressive, it is more likely to falsely assume that the fellow agent will yield or pass and thus leading to undesired outcome. Therefore, having inconsistent initial beliefs further put each type of agent's ability to converge to the right belief to test.

The figure 5.8 shows the trajectories of the agents with inconsistent initial belief. It is noted from the figure that when the agents are empathetic, the agents have better ability to avoid the collision zone except for the scenarios where the two agents start from close initial states. In contrast, non-empathetic agents have a difficult time correcting to the right policies, thus creating a lower reward interaction.

Policy choices

At every time step k, each agent chooses a policy based on the inferred parameter $\hat{\beta}_{-i}(k)$ and agent's own ground truth parameter β_i^*. Keep in mind that empathetic and non-empathetic obtains their parameter estimation differently as discussed in the implementation section, which produces the possible difference in policy choice that are studied here. The choice of policy over the trajectory are shown in this section: having value equals to 1 represents the correct choice of policy ($\hat{\beta} = \beta^*$, colored in purple) and vice versa.

Fig. 5.9 and 5.10 show that the empathetic agents have better ability at choosing the right policy (purple lines), whereas the non-empathetic ones have fewer instances

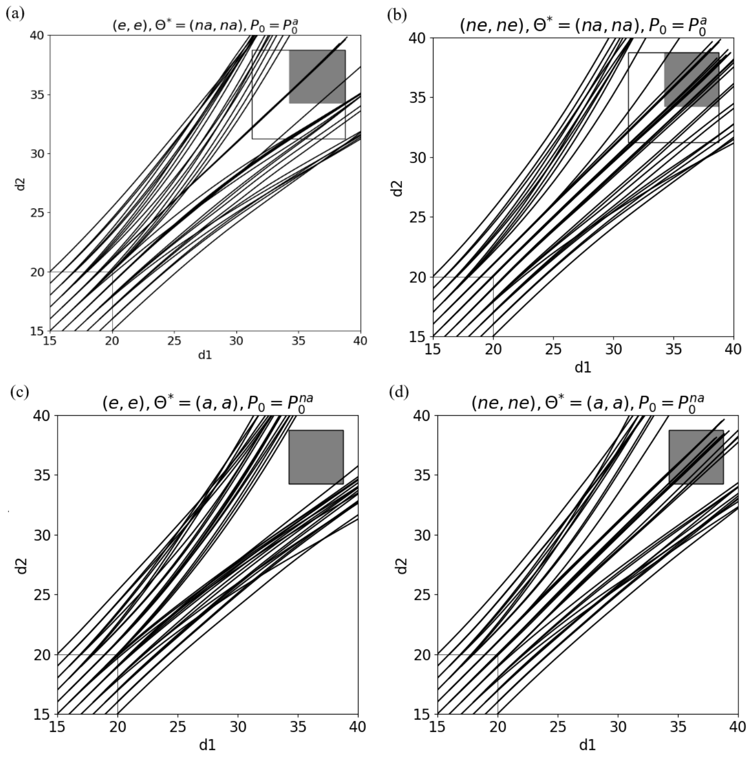

Figure 5.8: Inconsistent Initial Belief: When Belief Mismatches with Ground-true Reward Parameters. (a,b) Unknowingly Non-aggressive, (c,d) Unknowingly Aggressive. (a,c) Empathetic, (b,d) Non-empathetic.

selecting the right one. With the policy choice figures, additional observations can be made: while there are cases of empathetic agents having (pseudo) collisions, it can be seen that they are also the fewer cases where the correct policy is not followed through by both agents and creating an undesired situation. Thus, we draw the conclusion that empathetic interactions result in higher likelihood of agents correctly inferring the fellow agent's parameter and choosing the right policy, even when given the incorrect initial guesses.

5.0.3 Belief dynamics

To further investigate in detail of the change in belief, the way beliefs evolve over time is presented in the form of discrete-time belief dynamics. In order to

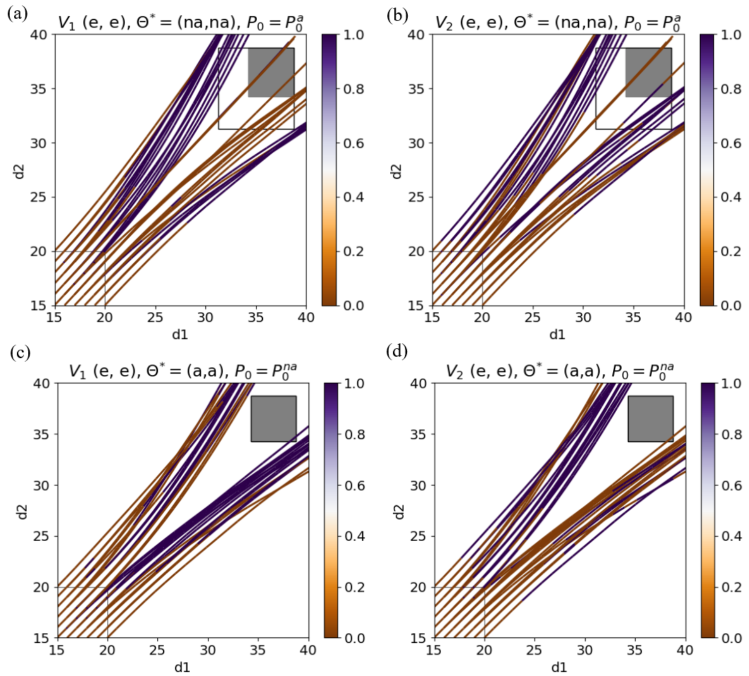

Figure 5.9: Policy Choice of E Agents: Color Represents the Policy Choices By Each Empathetic Agent, (a,c) are Agent 1, (b,d) are Agent 2.

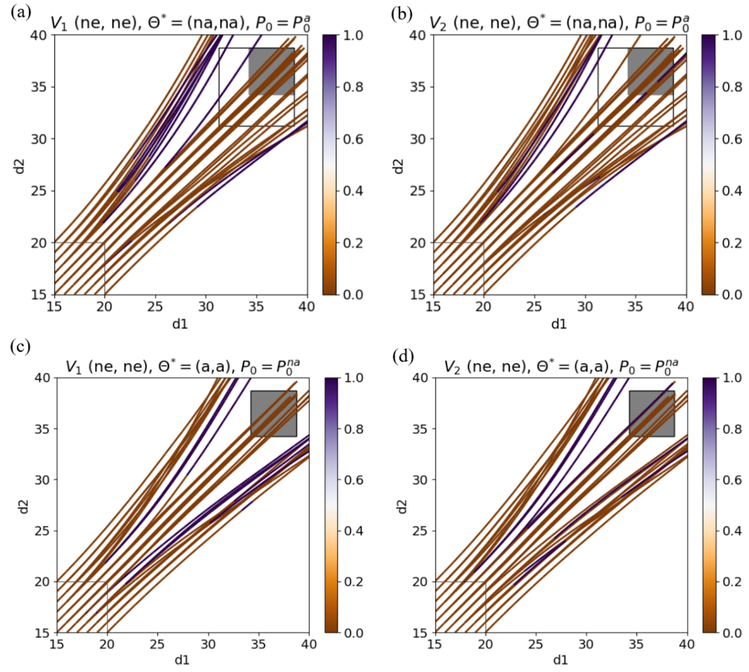

Figure 5.10: Policy Choice of NE Agents: Color Represents the Policy Choices By Each Non-empathetic Agent, (a,c) are Agent 1, (b,d) are Agent 2.

have a better look at the difference in belief dynamics, we select the cases from the inconsistent initial belief setting where the agent's reward is overwhelmingly better in the case of the empathetic agent.

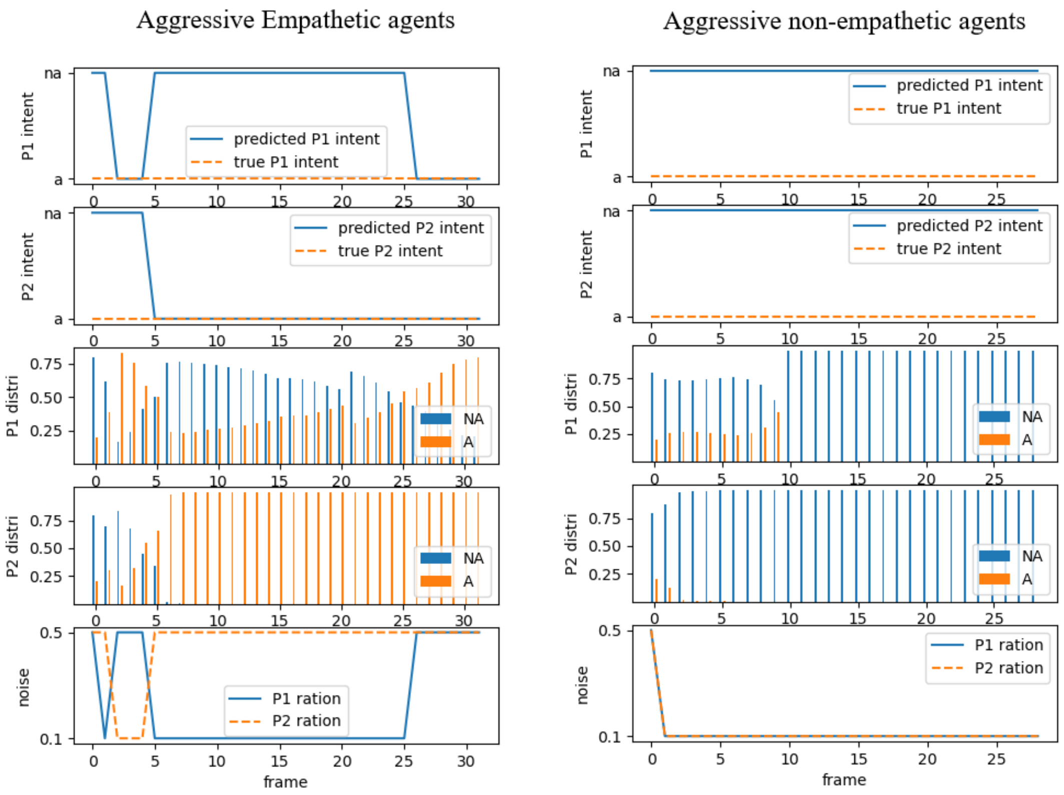

Figure 5.11: Belief Dynamics of E V.s.NE Agent: An Example of Evolution of Belief over Time of Empathetic Agents (left) Versus Non-empathetic Agents (right) with Inconsistent Initial Beliefs, Starting from $x1$=16 and $x2$=15. First Two Rows are Predicted Parameter By the Other Agent, 3rd and 4th are the Probability Distribution of Belief over the Simulation and the Last Row is the Belief over the Noise Parameter λ.

In the scenario of fig. 5.11 , the aggressive agents start from $x1 = 16$, $x2 = 15$ with inconsistent initial belief. The empathetic case result in the accumulated reward of -39.9, whereas the non-empathetic case has reward of -4686.74. The figure illustrates the marginal probability of other agent being each of the reward/intent parameter θ.

From the figure, it is trivial that in the empathetic interaction, the belief over the

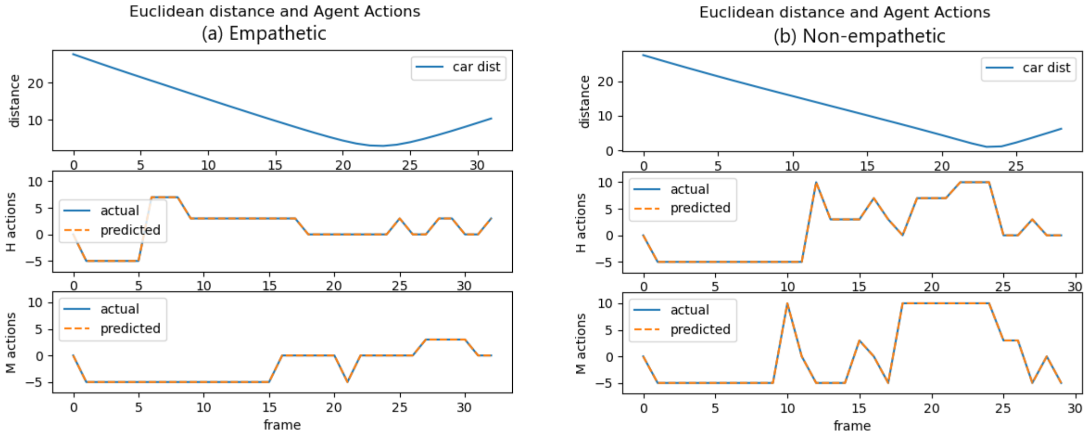

Figure 5.12: Actions Taken of E Vs. NE: Examples of Difference between Actions Taken between (a) Empathetic and (b) Non-empathetic, Starting from $x1$=16 and $x2$=15.

other agent quickly converged to the correct parameter which is aggressive (Aggressive, A), in contrast the non-empathetic interaction never converged, resulting in the agents taking the incorrect policy against the fellow agent. The actions taken by both types of agent are shown in Fig. 5.12, reflecting the difference in actions resulted from difference in belief: in the empathetic case car 1 (the car in front) sped up and car 2 slowed down to avoid close encounters; whereas the non-empathetic agents struggled to make the right move. Base on the belief dynamics figure, we make the relation that empathetic agent outperforms the non-empathetic ones by correctly inferring the parameters and choosing the right policy.

Chapter 6

CONCLUSION

In this paper, the advantage of "empathy" of agents in a multi-agent Interaction is studied in an uncontrolled intersection. We modeled the interaction as a multi-agent playing a incomplete differential game, where the two agents play the Nash Equilibrium based on their belief on the other agent's parameter. The difference between the empathetic and non-empathetic agents lies in the parameter estimation. From the simulation results, the reward values and policy choices show that empathy does indeed lead to better outcome.

Some open challenges still remain in Human-robot interactions such as our case studies, including finding perfect bayesian equilibrium, limitation on softmax calculation to be discrete (Boltzmann's distribution), etc. By overcoming some of the mathematical modeling difficulties in the future, the interaction can be further improved and be better at reflecting and dealing with the interactions in the real world.

The work can be extended by incorporating the probability distribution of the belief in the action-value, which helps differentiating the cases when the agent is highly confident versus when the agent is indecisive between the parameters. Studies can also be made based on the proposed modeling of the agents for more traffic scenarios such as roundabouts and lane-changing, to further validate the utility of empathy of agents. Lastly, while some notes on the belief dynamics are made, it is important that such dynamics is further scrutinized for a guarantee on the convergence of the correct parameter.

REFERENCES

Buckdahn, R., P. Cardaliaguet and M. Quincampoix, “Some recent aspects of differential game theory”, Dynamic Games and Applications **1**, 1, 74–114 (2011).

Chen, Y., L. Zhang, T. Merry, S. Amatya, W. L. Zhang and Y. Ren, “When shall i be empathetic? the utility of empathetic parameter estimation in multi-agent interactions”, (2020).

Foerster, J. N., R. Y. Chen, M. Al-Shedivat, S. Whiteson, P. Abbeel and I. Mordatch, “Learning with Opponent-Learning Awareness”, arXiv:1709.04326 [cs] ArXiv: 1709.04326 (2017).

Fridovich-Keil, D., A. Bajcsy, J. F. Fisac, S. L. Herbert, S. Wang, A. D. Dragan and C. J. Tomlin, “Confidence-aware motion prediction for real-time collision avoidance1”, The International Journal of Robotics Research **39**, 2-3, 250–265 (2020).

Kierzenka, J. and L. F. Shampine, “A bvp solver based on residual control and the maltab pse”, ACM Transactions on Mathematical Software (TOMS) **27**, 3, 299–316 (2001).

Kwon, M., E. Biyik, A. Talati, K. Bhasin, D. P. Losey and D. Sadigh, “When humans aren’t optimal: Robots that collaborate with risk-aware humans”, in “Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction”, pp. 43–52 (2020).

Li, N., D. W. Oyler, M. Zhang, Y. Yildiz, I. Kolmanovsky and A. R. Girard, “Game theoretic modeling of driver and vehicle interactions for verification and validation of autonomous vehicle control systems”, IEEE Transactions on control systems technology **26**, 5, 1782–1797 (2018).

Nikolaidis, S., D. Hsu and S. Srinivasa, “Human-robot mutual adaptation in collaborative tasks: Models and experiments”, The International Journal of Robotics Research **36**, 5, 618–634 (2017).

Peng, C. and M. Tomizuka, “Bayesian persuasive driving”, in “2019 American Control Conference (ACC)”, pp. 723–729 (IEEE, 2019).

Sadigh, D., N. Landolfi, S. S. Sastry, S. A. Seshia and A. D. Dragan, “Planning for cars that coordinate with people: leveraging effects on human actions for planning and active information gathering over human internal state”, Autonomous Robots **42**, 7, 1405–1426 (2018).

Schwarting, W., A. Pierson, J. Alonso-Mora, S. Karaman and D. Rus, “Social behavior for autonomous vehicles”, Proceedings of the National Academy of Sciences **116**, 50, 24972–24978 (2019).

Sinha, A. and A. Anastasopoulos, “Structured perfect bayesian equilibrium in infinite horizon dynamic games with asymmetric information”, in “2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)”, pp. 256–263 (IEEE, 2016).

Sun, L., W. Zhan and M. Tomizuka, “Probabilistic prediction of interactive driving behavior via hierarchical inverse reinforcement learning”, in “2018 21st International Conference on Intelligent Transportation Systems (ITSC)”, pp. 2111–2117 (IEEE, 2018a).

Sun, L., W. Zhan, M. Tomizuka and A. D. Dragan, “Courteous autonomous cars”, in “2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)”, pp. 663–670 (IEEE, 2018b).

Wang, Y., Y. Ren, S. Elliott and W. Zhang, “Enabling courteous vehicle interactions through game-based and dynamics-aware intent inference”, IEEE Transactions on Intelligent Vehicles **5**, 2, 217–228 (2020).