Multi-Agent Online Coordination via Bayesian Inverse Planning

Meiming Wang\textsuperscript{1,a}, Long Lan\textsuperscript{2,b}, Wenjing Yang\textsuperscript{3,c}\textsuperscript{*}

\textsuperscript{1,2,3}Institute for Quantum Information \& State Key Laboratory of High Performance Computing, College of Computer National University of Defense Technology Changsha, China

\textsuperscript{a}wangmeiming@nudt.edu.cn, \textsuperscript{b}long.lan@nudt.edu.cn

\textsuperscript{c}\textsuperscript{*} Corresponding author: wjyang1088@163.com

Abstract—Multi-agent coordination is a classical applied engineering problem and has been widely used in daily life. To efficiently coordinate all agents in the environment, most traditional models rely on a robust communication system between agents. However, in the scenario of autonomous vehicles, building a reliable communication system is extremely expensive and is very difficult to achieve in the foreseeable future. In this paper, we propose an online multi-agent coordination mechanism for each agent in the absence of communication and explore the benefits of Bayesian inverse planning in achieving the Nash equilibrium which provides an optimal coordination solution for each participant in multi-agent environment. By deliberate design, the proposed Bayesian inverse planning enables each independent agent to infer goals and predict future actions of other agents only by simply observing the past actions of them. To make a timely response to the dynamic environment, the inferences and predictions can be flexibly adjusted accordingly with the continuously updated action information of agents. Based on future action predictions, agents take coordinated actions to achieve Nash equilibrium coordination. We designed several mixed driving scenarios in the real world to test our model, and simulation results show that our model is efficient and flexible, compared with an optimal model and a heuristic model.

1. INTRODUCTION

With the increasing popularity of autonomous driving technology, more and more autonomous vehicles will appear on the road. There have been many researches on autonomous driving coordination [1-3]. However, most vehicles on the road are currently driven by human. How to coordinate autonomous vehicles controlled by intelligent agents with human-driven vehicles is a key issue that needs to be solved urgently. Most researches on coordinated driving focus on explicit cooperation, which depends on communication between vehicles [1,4]. Nevertheless, the communication between autonomous vehicles is not easy to achieve in the near future [5], let alone the communication between human drivers and auto-driving agents, since the required support equipments and infrastructures to enable communication are impossible to be deployed in the short term. Hence, auto-driving agents should be able to coordinate online with human agent efficiently and safely in the absence of communication. Besides efficiency, the flexibility of coordination is also important. Because human drivers may suddenly change their previous goals and future driving route caused only by a phone call or text message, the inferences and prediction results of the coordination model should flexibly adjust in time.
to adapt to this unexpected situation, so as to prevent traffic collisions caused by untimely and wrong inference.

Inspired by the scenarios of [2], we design the mixed scenarios that simulate the real-life coordination between vehicles to measure the performance of the coordination between human driving agents and autonomous driving agents. Fig. 1 shows one of the scenarios designed in this paper. In Fig. 1(a), the yellow Vehicle 1 is driven by the auto-driving agent, coming from the right side. The red Vehicle 2 and the blue Vehicle 3 are driven by human drivers coming from the left side. All the agents have their goal exits: Exit 1, Exit 2, Exit 3 or Exit 4. The autonomous vehicle agent Vehicle 1 needs to coordinate with Vehicle 2 and Vehicle 3 to reach their respective goal exits while driving through the non-signalized intersection. In Fig. 1(c), Vehicle 1 takes appropriate cooperative action in advance to avoid collision with Vehicle 3, based on the goal inference and action prediction results of Vehicle 3. Considering time cost, the coordination between vehicles ought to be as efficient as possible under safety.

To address the challenges of efficiency and flexibility described above, this paper develops a multi-agent online coordination mechanism based on Bayesian inverse planning, which doesn't depend on communication between agents. Our model ensures that the coordination between agents converges to the Nash equilibrium, which is the most frequently used stable state of coordination, since it's optimal for each participant and therefore efficient for their goal realization. Bayesian inverse planning enables the agents to infer goals and predict future actions based on observed prior information of other agents, such as past action sequences, and to continuously update their inferences based on new information of other agents, achieving online learning. The main contributions of this paper are as follows:

---

Figure 1. One of the mixed driving scenarios that several vehicles coordinate to pass through the non-signalized intersection. The yellow Vehicle 1 driven by autonomous agent needs to coordinate with the human driving vehicles (the blue Vehicle 2 and the red Vehicle 3) by inferring their goals and predicting their actions without communication. (a) The initial state. (b) The vehicles are driving to their goal exits. (c) Vehicle 1 coordinate with Vehicle 2 to avoid collision. (d) The vehicles have passed the intersection and moving to goal exits. The black arrows indicate the directions of vehicles.
We present a multi-agent online coordination model based on Bayesian inverse planning to achieve coordinated Nash equilibrium. Our model can infer goals and predict the future actions of agents only by observing past sequence of actions, so that the agents can take appropriate actions to achieve efficient coordination, without any communication between agents.

We design several mixed and real driving scenarios to measure the coordination performance of vehicles passing through a traffic intersection, and then conduct a series of experiments to compare the efficiency of our model with a heuristic model and an optimal model which ensures adequate and timely communication between agents. The experimental results show that our model outperforms the heuristic model and performs almost as efficiently as the optimal model.

We simulate the situations that agents suddenly change their previous goals while moving and the results demonstrate the flexibility of our model compared with the optimal model and the heuristic model.

2. RELATED WORK
Communication between multiple agents enables the agents to share their own intentions, perceived information and action selection strategies [8]. Therefore, communication is very important for multi-agent coordination. Most of the popular multi-agent coordination mechanisms rely on the communication between agents [9]. With the wide application of neural networks [10–12], the models proposed in [13–15] can even learn the communication protocol between agents through deep neural networks. However, there are still many scenarios where communication between agents is impossible. In autonomous driving scenarios, building a global communication system is very costly and hard to achieve in a short time [3]. In order to solve the problem of coordination between autonomous vehicles which cannot communicate with each other, researchers of [5,16,17] use Monte Carlo Tree Search (MCTS) method to simulate the results of next several time steps, and then select the most favorable joint action to achieve efficient coordination. However, these MCTS methods require a great number of simulation data. The consumption of computing resources rises sharply when the scene becomes complex. Recently, some researches focused on multi-agent cooperation for ad-hoc teamwork [7,16,17], which enables the agents to learn the model of teammates by analyzing their action sequences. Only when teammates have clear and consistent goals can the learned models be accurate enough. The generalization and flexibility of these methods are poor when teammates change their previous goals during coordination.

Besides, some studies of the theory of mind in computational cognition science, inspired by the excellent cooperative ability of human beings, attempt to imitate human cognitive ability in coordination [7, 18–20]. The authors of [7] propose a computational theory for understanding human behavior and intention based on inverse planning. Based on the rationality of agents, they analyze the historical action sequences of agents and calculate the probability of potential goals of agents by Bayesian inference. The experimental results compared with human subjects show that the results of their model are quite similar to those of human subjects. Their purpose is to verify to what extent their model understands human actions by comparing the calculation of goal probability with human subjects’ inferences, during a single agent moving along special paths in a simple 2D environment. Our model is to solve the multi-agent coordination problem, which predicts future actions based on the calculations of the agents’ goal probability, and perform efficient coordination with other agents.

For multi-agent reinforcement learning, it is a basic stability requirement that the action policy of each agent converges to equilibrium state [21]. Nash equilibrium is most frequently used. In the coordination of multi-agents, the action selection policies adopted by rational agents are actions that are most beneficial to themselves respectively, such as actions that are most advantageous to their respective goal realization. When reaching coordinated Nash equilibrium, no participant will benefit if any one of them individually changes their own policies, that is, the current coordinated state is the best for all participants [22]. Thus, it is an optimal and efficient strategy to perform the action policy that satisfies Nash equilibrium in multi-agent coordination.
3. THE COORDINATION MODEL

In this section, we elaborate our model in detail. The proposed model is formalized based on the goal-oriented Markov Decision Process. Firstly, the model predicts future actions of agents through goal inference, then it achieves efficient coordination towards Nash Equilibrium.

3.1. Goal-Oriented Markov Decision Process

Our model achieves multi-agent coordination based on agent goal inference and action prediction. So we use the popular goal-oriented Markov Decision Process (MDP) [23] to formalize the relationship between the state, action and goal of the agent at each time step. Specifically, goal-oriented MDP is a tuple \( \langle S, A, P, R, \gamma, G \rangle \), where \( S \) is the set of states in the environment; \( A \) is the set of actions; \( P \) is \( P(s'|s, a) \), which is the probability that the state changes to \( s' \) from \( s \) \( (s, s' \in S) \) after the agent takes action \( a \); \( R(s, a) \) is the reward function that measures the reward of taking action \( a \) at state \( s \); \( \gamma \) is the discount factor measuring the impact of future reward at present state; \( G \) is the goal of the agent. Action policy \( \pi(a|s) \) represents the probability of taking action \( a \) at state \( s \). Given the action policy \( \pi \), the value of state \( s \) is computed as:

\[
V^*(s) = \sum_{a \in A} \pi(a|s) \sum_{s' \in S} P(s'|s, a) \left[ R(s, a) + \gamma V^*(s') \right].
\]

Similarly, the Q-value, which measures the reward of taking action \( a \) at state \( s \), is computed as:

\[
Q(s, a) = \sum_{s' \in S} P(s'|s, a) \left[ R(s, a) + \gamma V^*(s') \right].
\]

3.2. Prediction of Future Actions Based on Goal Inference

In goal-oriented MDP, the action of an agent is directly dependent on its goal and the environment. In general, the agent is considered to be rational, which means that the agent chooses proper actions to achieve its goal as efficiently as possible [7]. The goal of the agent depends on the environment and its prior preference. Fig. 2 shows the relationship of goal, environment, action and priori. The shaded nodes Environment and Action are variables that can be directly obtained through observation, while the unshaded nodes Goal and Priori are implicit variables that need to be inferred. The environment is fully observable and deterministic.

![Figure 2. The relationship of goal, environment, action and priori. Shaded nodes are variables that can be directly obtained through observation, while the unshaded nodes are implicit variables that need to be inferred.]
where \( P(G|E) \) is the prior probability of goal \( G \) in the environment \( E \). In this paper, we take it as a uniform distribution, which is the regular setting as the prior preference is unknown. The second term on the right hand side of (3), \( P(A|G,E) \), represents the probability of choosing action \( A \) in the given environment and goal. As a rational agent, it will choose the action that is more beneficial for the goal realization. In other words, the probability of the agent choosing action \( A \) is positively related to the benefit value of the action. After calculating the value of \( P(G|E) \) and \( P(A|G,E) \), the probability of goal, \( P(G|A,E) \), can be computed through (3).

After calculating the posterior goal probability of the agent, the probability of future actions \( A' \) can be calculated using Bayes Rule as follow:

\[
P(A'|A,E) \propto P(A'|G,E)P(G|A,E). \tag{4}
\]

The analysis above gives the overall mathematical relationship between the agent’s action, goal and environment. More specifically, the relationship of the action, goal and state of the agent is shown in Fig. 3, combining with goal-oriented MDP. Environment node is not added in the figure considering the readability, which is represented as \( w \) in equations. Similarly, the shaded nodes can be obtained through observation, while the unshaded nodes are latent variables to be inferred. The process of inference and prediction is to infer the goal first based on the previously observed state sequence \( s_1, \ldots, s_{t-1}, s_t \) and action sequence \( a_1, \ldots, a_{t-2}, a_{t-1}, a_t \) up to time \( t \), and then predict future action \( a_{t+1}, a_{t+2}, \ldots, a_{t+n} \) and state \( s_{t+1}, s_{t+2}, \ldots, s_{t+n} \) in the following \( n \) time steps.

Figure 3. Fine-grained relationship of agent’s action, goal and state. Shaded nodes are variables can be observed, while unshaded nodes are implicit variables that need to be inferred.

Similar to (3), the posterior probability of goal \( P(g|s_{t-1}, w) \), given the past state sequence up to current time and the goal, can be computed as:

\[
P(g|s_{t-1}, w) \propto P(s_t|s_{t-1}, g, w)P(g|s_{t-1}, w), \tag{5}
\]

where \( P(g|s_{t-1}, w) \) is the prior probability that the agent chooses goal \( g \), which recursively updates as the posterior probability of goal \( g \), given the environment and previously observed state sequence \( s_{1:t-1} \). \( P(g|s_t, w) \), the initial distribution of \( P(g|s_{t-1}, w) \), is set as uniform distribution in our model, because, without extra information about the agent’s goal preference, we can only assume that the agent has the same probability of selecting each potential goal in the initial state. The other term of (5) on the right hand side, \( P(s_t|s_{t-1}, g, w) \), which represents the probability of the state transition from \( s_{t-1} \) to \( s_t \), given the environment and the goal, can be computed by marginalizing over all the actions \( a_{t-1} \) that cause the state to change from \( s_{t-1} \) to \( s_t \):
\[ P(s_t|s_{t-1}, g, w) = \sum_{a_{t-1}} P(s_t|a_{t-1}, w) P(a_{t-1}|s_{t-1}, g, w), \]  \hspace{1cm} (6)

where \( P(s_t|s_{t-1}, a_{t-1}, w) = 1 \) since the environment is deterministic and therefore the state transition from \( s_{t-1} \) to \( s_t \) is deterministic after taking action \( a_{t-1} \). \( P(a_{t-1}|s_{t-1}, g, w) \), corresponding to \( P(A|G,E) \) in (3), represents the probability that the agent selects action \( a_{t-1} \) at state \( s_{t-1} \) given the environment \( w \) and the goal \( g \), where \( \pi \) is the action policy. For readability and clarity, \( t-1 \) is replaced by \( \tau \) and \( P(\tau|s_{t-1}, g, w) \) is written as \( P(\tau|s_{t-1}, g, w) \). It is proportional to the difference between the values of action \( a_\tau \) and optimal action \( a_\tau^* \) at state \( s_\tau \), which is the Q-value in (2), compared with the value of optimal action \( a_\tau^* \), so it can be computed as:

\[ P(\tau|s_{t-1}, g, w) = K^{-1}\exp\left[\beta(Q_{p,w}^\tau(s_{t-1}, a_\tau) - Q_{p,w}^\tau(s_{t-1}, a_{\tau})\right], \]  \hspace{1cm} (7)

where \( Q_{p,w}^\tau(s_{t-1}, a_\tau) \) represents the value of taking action \( a_\tau \) at state \( s_{t-1} \) and \( Q_{p,w}^\tau(s_{t-1}, a_{\tau}) \) represents the maximum value of taking optimal action \( a_{\tau}^* \) at state \( s_{t-1} \), given the goal and the environment. The parameter \( \beta \) measures the degree of the agent’s rationality. Larger \( \beta \) means the agent prefers to select optimal actions, while smaller \( \beta \) means the agent selects actions more randomly. The coefficient \( K \) is the normalizing constant and computed as:

\[ K = \sum_{a_\tau \in A_\tau} P(\tau|s_{t-1}, g, w), \]  \hspace{1cm} (8)

where \( a_\tau \in A_\tau \) is the set of all actions \( a_\tau \) that can be executed at state \( s_\tau \). \( Q_{p,w}^\tau(s_{t-1}, a_\tau) \) can be computed as:

\[ Q_{p,w}^\tau(s_{t-1}, a_\tau) = \sum_{s_{t+1}} P(s_{t+1}|s_{t-1}, a_\tau, w)\gamma V_{p,w}^\tau(s_{t+1}) + R(s_{t}, a_{\tau}), \]  \hspace{1cm} (9)

When the action selection policy is optimal, represented as \( \pi^* \), we can get Bellman equation as follow:

\[ V_{p,w}^\tau(s_{t-1}) = \sum_{a_\tau} P(\tau|s_{t-1}, g, w) Q_{p,w}^\tau(s_{t-1}, a_\tau), \]  \hspace{1cm} (10)

that is,

\[ V_{p,w}^\tau(s_{t-1}) = \sum_{a_{\tau}} P(\tau|s_{t-1}, g, w) \times\]

\[ \left[ R(s_{t-1}, a_{\tau}) + \sum_{s_{t+1}} P(s_{t+1}|s_{t-1}, a_{\tau}, g, w)\gamma V_{p,w}^\tau(s_{t+1}) \right]. \]  \hspace{1cm} (11)

Equation (11) uses value iteration algorithm [24] to get the convergent value of \( V_{p,w}^\tau(s_{t-1}) \), then substitute it into (9), (7), (6) and (5) to compute the probability of goal \( g \) at time \( t \), namely, \( P(g|s_{t-1}, w) \). Then this model can predict the future state by marginalizing the probability of state transition over goal \( g \):

\[ P(s_{t+1}|s_{t-1}, g, w) = \sum_{g_{t+1}} P(s_{t+1}|s_{t-1}, g_{t+1}, w) P(g_{t+1}|s_{t-1}, w). \]  \hspace{1cm} (12)

Equation (12) means that the probability of future state to be \( s_{t+1} \) at time \( t+1 \) is equal to the weighted sum of the probabilities of different potential goal \( g_{t+1} \) and the transition probability from state.
to \( s_{t+1} \), conditioned on the previously observed state sequence \( s_{1:t} \) and the given environment \( w \). The probability of future action \( a_{t+1} \) is easy to get after we get the probability of future state since the state transition from \( s_t \) to \( s_{t+1} \) caused by action \( a_{t+1} \) is deterministic.

Based on the calculation of probability distribution of the next state, we can get the probability distribution of future \( k \) states through recursively extrapolated future \( k \) time steps by marginalizing over all possible future states at each time step, as:

\[
P(s_{t+1:t+k}|s_{1:t}, w) = \sum_{s_{t+1}} \sum_{s_{t+2}} \cdots \sum_{s_{t+k}} P(s_{t+1:t+k}|s_{1:t}, w) \times P(s_{t+1}|s_{1:t}, w),
\]

where \( s_{t+1:t+k} \in S_{t+k} \) represents the adjacent states of state \( s_{t+1} \). Based on the prediction of the agents' future states or actions, other agents are able to take actions that coordinate with each other and achieve their goals as efficiently as possible.

### 3.3. Efficient Coordination Based on Nash Equilibrium

For the coordination between \( n \) agents, a Nash equilibrium is the joint action policy \( [\pi^*_1, \ldots, \pi^*_n] \) such that each individual policy \( \pi^*_i \) is the optimal policy. At Nash equilibrium, any individual agent that changes its strategy will not get better returns. Thus the expected reward is maximum given all participants' policies \( [\pi_1, \ldots, \pi_n] \):

\[
E(v_i|\pi_1, \ldots, \pi_n) \leq E(v_i|\pi^*_1, \ldots, \pi^*_n) \quad \forall \pi_i,
\]

where \( v_i \) represents the reward value of agent \( i \). In our model, \( v_i \) is the value of state of agent \( i \) in (1). At any state \( s_t \), the reward of agent \( i \), namely \( v^e_i \), is optimal, which can be computed as:

\[
v^e_i(s_t, \text{NE}(s, v^e_1, \ldots, v^e_n)) = \max \{v^e_i(s)|v^e_1, \ldots, v^e_n, j \neq i\},
\]

where \( \text{NE} \) represents Nash equilibrium. \( \text{NE}_i \) is the action policy of agent \( i \) in a Nash equilibrium, which is the optimal action policy \( \pi^*_i(a|s) \):

\[
\pi^*_i(a|s_t) = \text{NE}_i(s, v^e_1, \ldots, v^e_i, \ldots, v^e_n).
\]

Nash equilibrium encourages each agent to select the action following the action optimal policy \( \pi^*_i(a|s) \), which ensures that the value of next state is the highest among all the next states. Therefore, coordination will be efficient if the action policy agents satisfies Nash equilibrium.

### 4. Experiments and Results

In this section, we firstly elaborate the mixed driving scenario used as benchmark for different models. Then the test results are displayed and carefully analyzed to show the performance of our model.

#### 4.1. The Mixed Driving Scenarios

To simulate the situations of multi-agent coordination between human drivers and autonomous driving agents, we designed several mixed driving scenarios that vehicles driven by human agents and autodriving agents coordinate to pass through the intersection. As shown in Fig. 4, three vehicles are passing through the intersection in each map. The black arrows indicate the initial driving direction of the vehicles and we eliminate the special situations that the vehicles may return to the exits from which they came, since there is no coordination in such cases and it makes no sense to consider these cases in this paper. All the vehicles need to coordinate to pass through the intersection and reach respective goal exits, without any communication during the whole coordination. The vehicles are traveling at the same speed to arrive at their desired goal exits in the 2D grid map. The human-driving vehicles adopt greedy strategy to reach the goal exit, which means they drive along the shortest path to goal exits. The auto-
driving vehicles, however, need to coordinate online with the human drivers, while driving along the path as short as possible. The vehicles will stop to avoid collisions in an emergency that they are about to collide with other vehicles. They will get an expensive time cost if the emergency of collision happens. The coordination process is that the auto-driving agent firstly make goal inferences and action predictions of the other vehicles based on observation of their past actions up to the current time, and then take proper actions that satisfies Nash equilibrium to coordinate to travel through the intersection efficiently and reach respective goal exits.

Figure 4. The mixed driving scenarios in different maps. (a) In Map 1, Vehicle 1 is the auto-driving agent; Vehicle 2 and Vehicle 3 are human-driving agents. Vehicle 1 needs to coordinate with Vehicle 2 and Vehicle 3 to pass through the intersection (the red dotted rectangle area) to arrive at Exit 1, Exit 2, Exit 3, or Exit 4. (b) In Map 2, Vehicle 1 and Vehicle 2 are auto-driving agents; Vehicle 3 is the human-driving agent. All vehicles coordinate to pass through the intersection to arrive at their goal exits. (c) In Map 3, Vehicle 1 and Vehicle 2 are auto-driving agents; Vehicle 3 is the human-driving agent. The black arrows indicate the directions of the vehicles.

In order to comprehensively evaluate the performance of the proposed model, this paper designed a series of mixed driving scenarios varying in maps and types of vehicles showed in Fig. 4. In Map 1, Vehicle 1 is the auto-driving agent; Vehicle 2 and Vehicle 3 are human-driving agents. Vehicle 1 needs to coordinate with Vehicle 2 and Vehicle 3 to pass through the intersection (the red dotted rectangle area) to arrive at Exit 1, Exit 2, Exit 3, or Exit 4. In Map 2, Vehicle 1 and Vehicle 2 are auto-driving agents, while Vehicle 3 is the human-driving agent. The human-driving Vehicle 3 still has the highest priority over auto-driving vehicles in coordination, while Vehicle 2 has higher priority when coordinating with the other auto-driving agent, Vehicle 1. Goal exits in this map include Exit 1, Exit 2, Exit 3, Exit 4 and Exit 5. In Map 3, the types of agents are the same as Map 2, but there are 6 exits in total.

The scenario is formalized as: the position of the agent represents its state; the set of actions is \{up, down, left, right, stop\}, while vehicles only stop when they are about to collide; \(R(s,a)\) is the cost of action, which is -1 for each action but 0 when reaching its goal; \(\gamma\) is set as 1; \(G\) is the position of goal; \(\beta\) is set as 0.8. For simplicity, the time cost of all vehicles passing through the red dotted rectangle area in Fig. 4 is used as the indicator to measure the coordination performance, which is relatively the same as the performance measured by the time all vehicles arrive at their respective goal exits. The experiment terminates after either all of the vehicles passing through the intersection or 30 time steps elapse. When an emergency of collision is about to happen, the vehicles will stop immediately and the cost is set to be -10 as punishment. As a result, coordination strategy aimed at achieving Nash equilibrium encourages vehicles to avoid collisions and then drive along the shortest path to goal exits.

To measure the benefits of the proposed mechanism quantitatively and qualitatively, we introduce an optimal model as the baseline and a heuristic model inspired by [27–29] which infers goals based on intuitive inference. In the optimal model, vehicles are able to share their goals and future actions by full communication. The model is denoted as Full Communication Model (FCM). The heuristic model (HM) is inspired by the common situations that when we see a moving vehicle close to a goal exit, we will intuitively think that it is traveling towards the closest exit. So the heuristic model is formalized as: the
probability of the agent choosing a goal is inversely proportional to its distance from the goal, which means the closer the agent is to a goal, the greater the probability it chooses this goal. After inferring the other agents’ goals, the heuristic model uses the same action prediction mechanism as (13) and the same coordination method in Section III(C).

4.2. Experimental Results
A series of experiments are conducted to test the efficiency and flexibility of the proposed coordination mechanism. For efficiency analysis, we simulate all the shortest paths that the vehicles might move and compare the time cost to pass through the intersection. The results demonstrate the superiority of our model over HM. For flexibility analysis, we select several typical goal-changing scenarios that make vehicles change their goal exits at certain time step while moving on the road. Compared with FCM and HM, our model outperforms HM in terms of flexibility in these typical scenarios.

4.2.1. Quantitative Analysis: We assume that vehicles will drive along the shortest paths to their goal exits based on the principle of rationality mentioned in Section III(B), so we simulate all the shortest paths to all possible goal exits that the vehicles may choose, while setting Vehicle 1’s goal exit as the one in front of its initial direction for simplify (Exit 4 in Map 1, Exit 5 in Map 2, Exit 6 in Map 3). Fig. 5 shows the process of coordination in one of the scenarios of Map 1. We compared the performance of coordination by measuring the time cost that all vehicles pass through the intersection.

Fig. 6 shows the average time cost of the three models for all vehicles in Map 1 to pass through the intersection when the goals of Vehicle 2 and Vehicle 3 are Exit 1, Exit 2 or Exit 3 (the shorter the time, the better the performance of the model). Our model performs as well as the optimal model FCM on all cases and more efficiently than the heuristic model on some cases. For the situations that Vehicle 2 chooses Exit 1 and Exit 3, all the three models are comparable in time cost, because no collision happens in these cases and all of them can drive along the shortest paths to goal exits. However, when Vehicle 2 goes to Exit 2, our model takes significantly shorter time to pass through the intersection than HM no matter which exit Vehicle 3 goes to. In Map 2 and Map 3, we only consider the time cost of the paths that may cause collision to reflect the performance of the coordination models more significantly and set the average time steps of Vehicle 3 moving towards all exits as the value of the vertical axis. Fig. 7 shows the results in these two maps. Our model also performs as well as FCM and outperforms HM in Map 1 and Map 2. The key reason is that our model accurately predicts the future actions but HM performs poorly when the vehicles are about to collide while crossing the intersection. In our model, accurate goal inferences and action predictions lead to efficient online coordination, while HM makes wrong goal inference and action prediction resulting in vehicle collision and sharp increase in the time cost.
Figure 5. An example of the mixed driving scenarios that vehicles pass through the intersection in Map 1. We show all possible positions for the next time step of Vehicle 2 and Vehicle 3, and predict the most likely paths for the next 3 time steps based on the action predictions of our model. The transparency of the vehicle indicates the probability of the position that it will be in the next few time steps. (a) At time 0, initial state. (b) At time 4, vehicles moving on the road. (c) At time 8, the yellow Vehicle 1 driven by auto-driving agent take evasive action to avoid collision based on the action prediction of our model. (d) At time 13, all vehicles have passed through the intersection, which means the coordination has completed.

Figure 6. Performance of the three models in Map 1 in terms of time cost for all vehicles to pass through the intersection in all the 9 goal combinations of Vehicle 2 and Vehicle 3.

Figure 7. Performance of the three models in (a) Map 2, and (b) Map 3. We focus on the time cost of the paths that may cause collision in these 2 maps and take the average time steps of Vehicle 3 moving towards all exits as the value of the vertical axis. When no collision happens in certain goal combinations of Vehicle 2 and Vehicle 3, the time cost is the time steps of a random shortest path.
4.2.2 Qualitative Analysis: Efficiency of coordination is trivial to ensure if vehicles keep moving towards a fixed goal exit since correct prediction of actions is not difficult to make due to the trajectory with obvious goal-directing. However, when a vehicle suddenly changes its previous goal, other vehicles on the road should identify the change as soon as possible and take proper actions to coordinate with it, because inaccurate goal inferences may lead to incorrect prediction of action and result in collision. Therefore, the model should have enough flexibility to adapt to the sudden change of the agent’s goal. To test the flexibility of our model, we make the human-driving vehicles or one of the auto-driving vehicles change its previous goal while driving, and then see if the other auto-driving vehicles controlled by our coordination mechanism can identify the goal changes and how many time steps the coordination model needs, which guides other vehicles to take timely actions to avoid collision and coordinate as efficiently as possible. This paper designed 3 typical scenarios in every map where Vehicle 2 or Vehicle 3 suddenly changes their previous goal exits while driving on the road. The scenarios in the 3 maps are as follows:

Scenarios in Map 1,
- Scenario 1: Vehicle 3 changes its goal exit from Exit 1 to Exit 2 at the 5th time step.
- Scenario 2: Vehicle 3 changes its goal exit from Exit 3 to Exit 2 at the 5th time step.
- Scenario 3: Vehicle 2 changes its goal exit from Exit 1 to Exit 2 at the 8th time step.

Scenarios in Map 2,
- Scenario 1: Vehicle 3 changes its goal exit from Exit 4 to Exit 2 at the 5th time step.
- Scenario 2: Vehicle 2 changes its goal exit from Exit 5 to Exit 2 at the 3rd time step.
- Scenario 3: Vehicle 2 changes its goal exit from Exit 1 to Exit 2 at the 5th time step.

Scenarios in Map 3,
- Scenario 1: Vehicle 2 changes its goal exit from Exit 1 to Exit 3 at the 5th time step.
- Scenario 2: Vehicle 3 changes its goal exit from Exit 3 to Exit 2 at the 4th time step.
- Scenario 3: Vehicle 3 changes its goal exit from Exit 5 to Exit 2 at the 5th time step.

Fig. 8 shows the time cost for all vehicles to pass through the intersection in these three scenarios. In Map 1 we recorded the time cost of all possible shortest paths of the vehicles, while in Map 2 and Map 3, we only focused on the paths that vehicles might collide. Our model outperforms HM and has almost the same time cost as the optimal model FCM in all scenarios of every map, since HM makes wrong inferences about the change of the vehicle’s goal or identifies the change too late, which leads to wrong action predictions and results in emergent stop to avoid collision.
Figure 8. Performance of our model in goal-changing scenarios in (a) Map 1, (b) Map 2, and (c) Map 3. Our model outperforms HM and has almost the same performance as the optimal model FCM in all scenarios of each map.

To figure out the reason why our model takes shorter time than HM when the vehicle changes its goal while passing the intersection, in Fig. 9, we show the probability distributions calculated by the two models of the goal exits of Vehicle 3 in Scenario 1 of Map 1, where Vehicle 3 changes its goal exit from Exit 1 to Exit 2 at the 6th time step and leaves the intersection at the 12th time step. Fig. 9(a) shows the probabilities of vehicle 3’s goal exits computed by our model over time. At time step 0, the probability of Vehicle 3 choosing any one of the three potential goal exits is equal. As the vehicle moves to its’ goal exit, the most likely goal of Vehicle 3 becomes Exit 1, which is exactly the initial goal of Vehicle 3. But Vehicle 3 changes its goal to Exit 2 at the 6th time step, and our model senses the change only after 2 time steps according to the calculation result that the probability of Exit 2 becomes the highest one. So our model predicts Vehicle 3’s future actions accurately and coordinates with it to avoid collision. However, in Fig. 9(b), the heuristic model can’t figure out that the goal exit of Vehicle 3 has changed to Exit 2 until time step 10, which is too late because Vehicle 1 and Vehicle 3 have already stopped to avoid collision. Obviously, vehicles controlled by HM cost more time to pass through the intersection than our model. Fig. 10 shows the results of Scenario 1 in Map 2, which also support the same conclusion.

Figure 9. Probability distribution of Vehicle 3’s goal exits of calculated by (a) our model, and (b) HM in Scenario 1 of Map 1. The y-axis represents the probability of Vehicle 3’s potential goal exits and the x-axis represents time steps.
Figure 10. The probability distribution of goal exit of Vehicle 3 calculated by (a) our model, and (b) HM in Scenario 1 of Map 2.

Fig. 11, however, shows that our model costs the same time as HM to identify the goal change in scenario 1 of Map 1, but vehicles controlled by our model still take less time to pass the intersection than HM (Fig. 8(c)). Notice that the probability of Exit 3 at time step 11 in Fig. 11(a) is higher than the probability at the same time in Fig. 11(b). The reason is that our model still correctly predicts the future action of Vehicle 3, since the higher probability of the goal leads to the higher probability of the corresponding action in (12). In other words, the lower probability of the goal calculated by HM means HM suggests that the vehicle might choose other goals with a considerable probability, resulting in the wrong action predictions. Besides, Table 1 shows the time cost for our model and HM to identify goal changes in all scenarios of each map. Our model costs less time than HM in most of the scenarios. In summary, the results of goal-changing scenarios reveal that our model has stronger flexibility than HM and almost the same performance as the optimal model in terms of the time cost to pass the intersection in goal-changing scenarios.
Figure 11. The probability distribution of goal exit of Vehicle 2 calculated by (a) our model, and (b) HM in Scenario 1 of Map 3.

TABLE 1. Performance of flexibility in terms of time steps for our model and HM to identify the goal change of vehicles in all scenarios of each map (the less the time, the better the performance).

| Scenario | Map 1 | Scenario | Map 2 | Scenario | Map 3 |
|----------|-------|----------|-------|----------|-------|
| Scenario 1 | 5     | Scenario 1 | 10    | Scenario 1 | 6     |
| Scenario 2 | 3     | Scenario 2 | 12    | Scenario 2 | 4     |
| Scenario 3 | 5     | Scenario 3 | 8     | Scenario 3 | 3     |

5. CONCLUSIONS AND FUTURE WORK

We develop a multi-agent online coordination model which enables agents to infer goals and predict actions through Bayesian inverse planning and to achieve Nash equilibrium in coordination, without relying on communication between them. Our model solves the challenges of efficiency and flexibility in mixed driving scenarios where autonomous vehicles need to coordinate with human driving vehicles or other auto-driving vehicles. Experiments on the simulation of multiple vehicles passing through the intersection empirically demonstrate the efficiency and flexibility of our model.

This paper provides an enlightening research idea for the multi-vehicle coordination problem in the future when more and more autonomous vehicles are driving on the road. Yet the mixed driving scenario we designed is relatively simple. To make our idea clear, we only consider several vehicles on the road and the actions and environments in our case are also simple. In real world, they are more complex. Therefore, the focus of our future work lies in two parts: increase the scalability of the algorithm to ensure efficient coordination with a large number of vehicles, and improve the flexibility and stability of the model to deal with more complex traffic situations.

ACKNOWLEDGMENT

This work was partially supported by the National Natural Science Foundation of China (No. 91948303, 91648204, and 61803375), the National Key Research and Program of China (No. 2017YFB1001900 and 2017YFB1301104), and the National Science and Technology Major Project.

REFERENCES

[1] Englund, C.; Chen, L.; Ploeg, J.; Semsar-Kazerooni, E.; Voronov, A.; Bengtsson, H.H.; Didoff, J. The grand cooperative driving challenge 2016: boosting the introduction of cooperative automated vehicles. IEEE Wireless Communications 2016, 23, 146–152.
[2] Kessler, T.; Knoll, A. Cooperative Multi-Vehicle Behavior Coordination for Autonomous Driving. 2019 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2019, pp. 1953–1960.
[3] Rios-Torres, J.; Malikopoulos, A.A. A survey on the coordination of connected and automated vehicles at intersections and merging at highway on-ramps. IEEE Transactions on Intelligent Transportation Systems 2016, 18, 1066–1077.

[4] Duering, M.; Pascheka, P. Cooperative decentralized decision making for conflict resolution among autonomous agents. 2014 IEEE International Symposium on Innovations in Intelligent Systems and Applications (INISTA) Proceedings. IEEE, 2014, pp. 154–161.

[5] Kurzer, K.; Engelhorn, F.; Zöllner, J.M. Decentralized cooperative planning for automated vehicles with continuous monte carlo tree search. In Proceedings of the 21st International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2018, pp. 452–459.

[6] Bušoni, L.; Babuška, R.; De Schutter, B. Multi-agent reinforcement learning: An overview. In Innovations in multi-agent systems and applications-1; Springer, 2010; pp. 183–221.

[7] Baker, C.L.; Saxe, R.; Tenenbaum, J.B. Action understanding as inverse planning. Cognition 2009, 113, 329–349.

[8] Li, M.; Yang, W.; Cai, Z.; Yang, S.; Wang, J. Integrating Decision Sharing with Prediction in Decentralized Planning for Multi-Agent Coordination under Uncertainty. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), 2019, pp. 450–456.

[9] Dorri, A.; Kanhere, S.S.; Jurdak, R. Multi-agent systems: A survey. IEEE Access 2018, 6, 28573–28593.

[10] Junyang, Z.; Huili, W.; Yang, G.; Xiao, H. Review of deep learning. Application Research of Computers 2018, 7, 1921–1936.

[11] Lan, L.; Wang, X.; Hua, G.; Huang, T.S.; Tao, D. Semi-online Multi-people Tracking by Re-identification. International Journal of Computer Vision 2020, pp. 1–19.

[12] Sadeghian, A.; Kosaraju, V.; Sadeghian, A.; Hirose, N.; Rezatofighi, H.; Savarese, S. Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 1349–1358.

[13] Foerster, J.; Assael, I.A.; De Freitas, N.; Whiteson, S. Learning to communicate with deep multi-agent reinforcement learning. In Proceedings of the Neural Information Processing Systems (NeurIPS), 2016, pp. 2137–2145.

[14] Sukhbaatar, S.; Fergus, R.; others. Learning multiagent communication with backpropagation. In Proceedings of the Neural Information Processing Systems (NeurIPS), 2016, pp. 2244–2252.

[15] Jiang, J.; Lu, Z. Learning attentional communication for multi-agent cooperation. In Proceedings of the Neural Information Processing Systems (NeurIPS), 2018, pp. 7254–7264.

[16] Lenz, D.; Kessler, T.; Knoll, A. Tactical cooperative planning for autonomous highway driving using Monte-Carlo Tree Search. In Proceedings of the 2016 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2016, pp. 447–453.

[17] Kurzer, K.; Zhou, C.; Zöllner, J.M. Decentralized cooperative planning for automated vehicles with hierarchical monte carlo tree search. In Proceedings of the 2018 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2018, pp. 529–536.

[18] Shum, M.; Kleiman-Weiner, M.; Littman, M.L.; Tenenbaum, J.B. Theory of minds: Understanding behavior in groups through inverse planning. In Proceedings of the AAAI Conference on Artificial Intelligence, 2019, Vol. 33, pp. 6163–6170.

[19] Baker, C.; Saxe, R.; Tenenbaum, J.B. Bayesian models of human action understanding. In Proceedings of the Neural Information Processing Systems (NeurIPS), 2006, pp. 99–106.

[20] Ullman, T.; Baker, C.; Macindoe, O.; Evans, O.; Goodman, N.; Tenenbaum, J.B. Help or hinder: Bayesian models of social goal inference. In Proceedings of the Neural Information Processing Systems (NeurIPS), 2009, pp. 1874–1882.

[21] Hu, J.; Wellman, M.P. Nash Q-learning for general-sum stochastic games. Journal of machine learning research 2003, 4, 1039–1069.

[22] Gürkan, G.; Pang, J.S. Approximations of Nash equilibria. Mathematical Programming 2009, 117, 223–253.
[23] Teichteil-Königsbuch, F. Stochastic safest and shortest path problems. In the Proceedings of the AAAI Conference on Artificial Intelligence, 2012.
[24] Sutton, R.S.; Barto, A.G., Value Iteration. In Reinforcement learning: An introduction, Second ed.; MIT press, USA, 2018; pp. 98–101.
[25] Barrett, H.C.; Todd, P.M.; Miller, G.F.; Blythe, P.W. Accurate judgments of intention from motion cues alone: A cross-cultural study. Evolution and Human Behavior 2005, 26, 313–331.
[26] Gigerenzer, G.; Todd, P.M., Simple heuristics that make us smart. In How motion reveals intention: categorizing social interactions; Oxford University Press, USA, 1999; pp. 257–286.
[27] Zacks, J.M. Using movement and intentions to understand simple events. Cognitive Science 2004, 28, 979–1008.