Improving Dialog Systems for Negotiation with Personality Modeling

Runzhe Yang[*]
Princeton University
runzhey@princeton.edu

Jingxiao Chen[*]
Shanghai Jiao Tong University
timemachine@sjtu.edu.cn

Karthik Narasimhan
Princeton University
karthikn@princeton.edu

Abstract

In this paper, we explore the ability to model and infer personality types of opponents, predict their responses, and use this information to adapt a dialog agent's high-level strategy in negotiation tasks. Inspired by the idea of incorporating a theory of mind (ToM) into machines, we introduce a probabilistic formulation to encapsulate the opponent's personality type during both learning and inference. We test our approach on the CRAIGSLISTBARGAIN dataset (He et al., 2018) and show that our method using ToM inference achieves a 20% higher dialog agreement rate compared to baselines on a mixed population of opponents. We also find that our model displays diverse negotiation behavior with different types of opponents.[1]

1 Introduction

Developing dialog systems for negotiation is challenging since the task requires a combination of good communication skills and strategic reasoning capabilities (Traum et al., 2008; Young et al., 2013; Keizer et al., 2017). While recent neural models (Wen et al., 2017; Dhingra et al., 2017; Zhou et al., 2019; He et al., 2018) have shown that useful dialogue strategies can be learned from offline corpora, they do not explicitly model the mental state of other agents, which can make it challenging to generate tailored strategies and utterances for different types of opponents.

In this paper, we introduce a new framework for generating strategic dialog inspired by the idea of Theory of Mind (ToM) from cognitive science (Premack and Woodruff, 1978; Bruner, 1981; Wimmer and Perner, 1983). When negotiating with others, humans innately infer the intention of the other party, and guess how their own utterances would affect the opponent's mental state. To emulate this capability in machines, we train a first-order ToM model to predict an opponent's response given the current state and the agent's own possible utterances. This first-order ToM model can then be incorporated into dialog agents to enable one-step lookaheads during inference.

In order to predict future responses, we model the opponent's personality type as a intermediate variable (z), which can be predicted using the dialogue history. We use this predicted personality, along with the previous state and utterance to calculate the likelihood of the opponent's next state for all possible actions that our agent can take in the current state. This allows us to compute an expected value of return for each action, which is subsequently used to produce a policy for our agent. We propose two variants of our ToM-based dialog agent – an **explicit** version that outputs the opponent type as an intermediate prediction, and an **implicit** version that models the opponent type as a latent variable. Both models can be instantiated as end-to-end neural networks and can be trained using reinforcement learning.

Our approach differs from existing opponent modeling work (Lee et al., 2018; Hadjinikolis et al., 2013; Oren and Norman, 2009; Rienstra et al., 2013; He and Boyd-Graber, 2016) in three aspects: 1) it provides strategic benefit during *inference* which leads to more successful negotiations, 2) it can flexibly adjust the degree of dependence on ToM predictions by changing a temperature parameter, and 3) it utilizes text utterances to infer types of opponents, thereby capturing side information (*e.g.*, emotion) that is useful yet absent from standard dialog state transitions.

We perform experiments on a modified version of the CRAIGSLISTBARGAIN negotiation task (He et al., 2018), where the agent is matched with dif-

---

[*]Authors contributed equally.
[1]Code and data available at https://github.com/princeton-nlp/NegotiationToM

ferent opponents from diverse populations (e.g., cooperative, competitive, and aggressive negotiators), without being provided information about their identity. Empirically, our method outperforms several baselines on the task by completing more deals and achieving higher utility. For instance, our model achieves about 20% higher dialog agreement rate and utility than a baseline dialog manager trained with reinforcement learning. Our analysis reveals that the agent demonstrates diverse negotiation behavior and adapts well to different types of opponents.

2 Related Work

Speaker-follower models and rational speech acts. Our work is related to recent papers using the Rational Speech Acts (RSA) model for natural language (Goodman and Stuhlmüller, 2013; Monroe and Potts, 2015; Goodman and Frank, 2016; Shen et al., 2019). RSA has also been applied to language grounding (Andreas and Klein, 2016) and vision-language navigation (Fried et al., 2018). Our first-order theory of mind modeling is different since we learn how the speaker's intent and utterance affect the opponent's reaction, instead of assuming the optimality of the listener in the speaker's mind. Recent RSA model (White et al., 2020) considers speakers and listeners in resource-constrained settings, while we do not enforce constraints on opponents.

Our approach with explicit characteristic modeling is also similar to the ToMnet (Rabinowitz et al., 2018), which uses a multi-agent reinforcement learning setting to learn identity embeddings of populations from past trajectories, and predict the mental state of an agent using the current trajectory. However, our first-order ToM models for negotiation also take utterances into account, which makes improving upon a base RL policy non-trivial.

Theory of Mind in dialog systems. Theory of mind for modeling user personality types and predicting responses has been studied in the context of building user simulators (Georgila et al., 2006; Rieser and Lemon, 2006) for training RL-based dialog systems, and to make dialog systems explainable (Chandrasekaran et al., 2017). Recent work on dialog policy learning has employed theory of mind with a focus on specific domains. The Recursive Mental Model (RMM) (Roman et al., 2020) was proposed for navigation settings, where questions and answers are generated between a

navigating agent and a guiding agent. Another approach – Answerer in Questioner's Mind (AQM) (Lee et al., 2018) – tackled an answer guessing game with information-theoretic methods. In these domains, the opponents are assumed to be cooperative, while our method is applicable for interacting with both cooperative and competitive opponents. Recently, Jang et al. (2020) employed Bayesian-optimal Monte-Carlo planning for end-to-end dialog generation at the utterance level. However, their method only models the latent goal of the opponent instead of potential responses like we do.

Opponent modeling in RL. Apart from dialog systems, opponent modeling has been explored in other multi-agent reinforcement learning settings (Wen et al., 2019; von der Osten et al., 2017; He and Boyd-Graber, 2016; Hadjinikolis et al., 2013; Rienstra et al., 2013). Our approach differs from these works by: 1) providing strategic benefit during real-time inference, 2) adjusting the degree of dependence on the ToM predictions through a temperature parameter, and 3) utilizing text utterances in the dialog to infer types of opponents, thereby capturing side information that is useful yet absent from standard state transitions.

3 Framework

Task. We consider a task-oriented dialog setting where there are two agents, a buyer and a seller. The buyer's goal is to purchase the listed item with minimum cost, and the seller's goal is to sell the item at a price as high as possible. The item description is public for both agents, while the target prices are private for both buyer and seller. Two agents negotiate in alternating turns until they conclude with an agreement or disagreement.

MDP Formulation. We formulate the negotiation process between two agents as a *multi-agent Markov Decision Process* (MAMDP), $\langle \mathcal{N}, \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \Pi, n \rangle$. $\mathcal{N} = \{-1, 1\}$ is the set indicating two agents (*buyer=-1 / seller = 1*). $\mathcal{A}$ is the action space consisting of *dialog acts*. For example, a valid dialog act $a_t^i \in \mathcal{A}$ can encode the *intent* (*inform*, *propose*, *counter*, etc.) and *price* that the agent i tries to express in the t-th round. Two agents act alternatively, i.e., if at the round t only the agent i moves, then at the round $t + 1$ only the agent $-i$ moves.

$\mathcal{S}$ is the state space consisting of the negotiation status. We define $s_0 \in \mathcal{S}$ as the initial status of the dialog, which contains the information about items
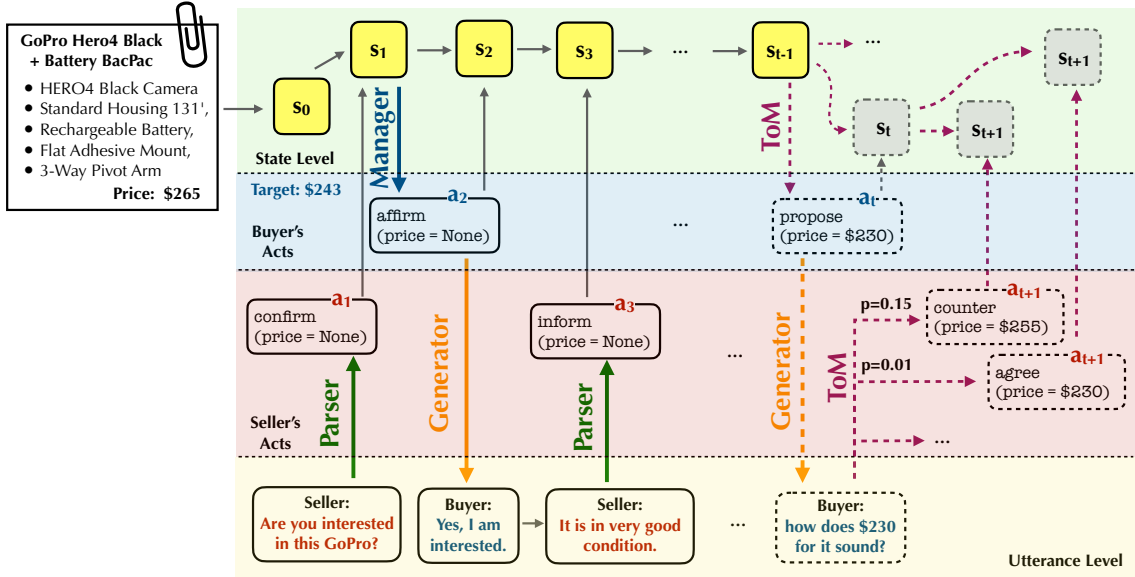
Figure 1: Our Theory of Mind (ToM) framework of negotiation systems. The interaction between a *buyer* and a *seller* can be divided into three levels: The utterance level, dialog act level, and state level. The parser extracts an intent and key information (*e.g.*, price) from an input utterance as a dialog act. Both intents and key information, along with the context (*e.g.* description about the item), contribute to the state of dialog. The traditional RL-based dialog manager decides a dialog act based on the current state. And the generator converts the abstract dialog act back to a natural language utterance, also based on the previous state. The first-order ToM model explicitly predicts the response of the opponent and the state transition, which supports more strategic negotiation.

to be negotiated (e.g., initial price, description). We also define $s_t = (s_0, a_1^i, a_2^{-i}, \ldots, a_{t-1}^i, a_t^{-i})$. In this way, the only randomness of the environment comes from the opponents policy ($s_{t-1} \rightarrow a_t^{-i}$), i.e., $s_{t-1} \rightarrow s_t$ is stochastic, while $(s_{t-1}, a_t^{-i}) \rightarrow s_t$ is deterministic. Note that the state s_t is only partially observable in reality, since one can only infer the true intent from the corresponding utterance. We provide a summary of all the symbols used in Table 1.

3.1 Negotiation Systems

As illustrated in Figure 1, our negotiation system encapsulates three important modules following traditional goal-oriented dialog systems (Young et al., 2013):

- A **parser** that converts the opponent's utterance u_{t-1}^{-i} to dialog act a_{t-1}^{-i} (e.g., *"Are you interested in this GoPro"* → confirm(price=None)). Since the dialog acts in our system do not intend to capture the complete semantics of a sentence, a simple rule-based parser is effective;

- A **manager** that decides the responding dialog act a_t^i according to the current dialog state $s_{t-1} = (s_0, \ldots, a_{t-1}^{-i})$. Our ToM model is applied to this component of the system;

- A **generator** that produces natural language response u_t^i based on the current dialog act a_t^i and the dialog state s_{t-1}, or equivalently s_t (e.g., the previous dialog state + propose(price=$230) → *"How does $230 for the GoPro sound?"*). It can be either deterministic to reduce computational cost or probabilistic to encourage diversity in language.

Following (He et al., 2018), the parser and the generator modules are obtained by rule-based method or supervised learning in advance, and fixed when training the dialog manager using supervised learning (SL) or fine turning using reinforcement learning (RL). The SL dialog manager employs a neural network to model state transitions $P(s_t|s_{t-1})$ (or equivalently, $\pi(a_t^i|s_{t-1})$) of the training corpus by minimizing the cross entropy loss. The RL dialog manager further fine tunes the SL model by maximizing a composed reward function with reinforcement learning. The learned dialog policy $\pi(a_t^i|s_{t-1})$ can be further improved by enforcing some hand-craft rules.

There are two main problems with the SL or RL manager. First, the policy learned by an RL-based dialog manager produces reactive responses (Tamar et al., 2016), which are usually inadequate in a long term planning problem requiring more

Symbol	Definition
$\mathcal{N} = \{-1, 1\}$	Identities of the two players (*buyer = -1 / seller = 1*)
$s_0 \in \mathcal{S}$	Initial state of the dialog (e.g., list price, description).
$a_t^i \in \mathcal{A}$	dialog act of agent i. If $-i$, it denotes the opponent.
$s_t \in \mathcal{S}$	State at the end of round t, $s_t := (s_0, a_1^i, a_2^{-i}, \ldots, a_{t-1}^i, a_t^{-i})$
$\mathcal{P}(s_t\|s_{t-1})$	Transition probability, associated with agent policies
$\pi^i(a_t^i\|s_{t-1})$	Probability of the agent i choosing a_t^i given the previous dialog state.
$\mathcal{R} = \{r^{-i}, r^i\}$	Reward functions for agents i and $-i$.
n	Maximum length of dialogs.
u_t^i	Utterance of agent i. If $-i$, it denotes the opponent.
z^{-i}	Type the opponent. Annotations are available in the corpus.

Table 1: Notation for our MAMDP formulation of task-oriented dialog for negotiation.

strategic thinking, such as negotiation. Second, it does not take the effect of the agent's generated utterances on opponents' reactions into account. To address these problems, we propose an approach to incorporate the theory of mind (ToM) (Premack and Woodruff, 1978) into the inference process. This enables one-step looking ahead to consider the effect of the agent's utterances and generate more thoughtful strategies.

4   First-Order Theory of Mind for dialog

The goal of the first-order theory of mind is to predict how a dialog act and an utterance generated by us would affect the reaction of the opponent. As illustrated in Figure 1, suppose that our current dialog state is s_{t-1}, which consists of the history of past dialog acts and the initial information, as well as the current utterance u_{t-1}^{-i} from the opponent. The ToM model simulates the situations where we take dialog act a_t^i (*e.g.*, propose(price=\$230)) and utter the sentence u_t^i (*"how does \$ 230 for it sound"*), and estimates the probability distribution of the opponents response a_{t+1}^i. By combining actions and states by definition, our first-order ToM model estimates the transition probability $T(s_{t+1}\|u_{t-1}^{-i}, s_t, u_t^i)$.

In practice, the opponent may have different language preferences (*e.g.,* using more aggressive or

mild words when countering) and strategies (*e.g.,* tend to insist on their target price or agree to a compromise). The first-order ToM can either **implicitly** capture these personalities by learning the transition $T(s_{t+1}\|u_{t-1}^{-i}, s_t, u_t^i)$, or **explicitly** infer the type of the opponent's personalities z^{-i} first, from the past interaction and the opponent's utterance, i.e., learning an identifier $z_{t-1}^{-i} = f(s_{t-1}, u_{t-1}^{-i})$, and then learns the transition based on that information, i.e., $T(s_{t+1}\|z_{t-1}^{-i}, s_t, u_t^i)$, to make accurate prediction about opponents reaction.

4.1   First-order ToM Policies with Explicit Personality Modeling

We introduce a policy with an explicit first-order ToM model $T(s_{t+1}\|z^{-i}, s_t, u_t^i)$, where the opponent's personality z^{-i} can be estimated from partial dialog. During training, the ground truth of the type of opponents personalities, z, is given. Therefore we can train an **identifier** $z_{t-1}^{-i} = f(s_{t-1}, u_{t-1}^{-i})$ with extra supervision to predict the opponents type every round. During the inference process, the probability of taking action a_t^i, i.e., a policy $\pi_{\texttt{ToM}}(a_t^i\|s_{t-1}, z_{t-1}^{-i})$, is proportional to

$$\exp\left\{\frac{1}{\beta}\sum_{u_t^i}\underbrace{G(u_t^i\|s_t, z_{t-1}^{-i})}_{\text{Generator}}\sum_{s_{t+1}}\underbrace{T(s_{t+1}\|z_{t-1}^{-i}, s_t, u_t^i)}_{\text{1}^{\text{st}}\text{-order ToM}}\underbrace{V(s_{t+1})}_{\text{Value Fn.}}\right\},$$

where the exponent can be interpreted as the expected best return over opponent's next moves, after taking action a_t^i at state s_{t-1} (compressed as s_t). In the above expression, $T(s_{t+1}\|z_{t-1}^{-i}, s_t, u_t^i)$ is the **explicit first-order ToM model**, which can be trained by supervised learning from the corpus; $G(u_t^i\|s_t, z_{t-1}^{-i})$ is the **generator** which renders utterance conditioned on the current state and the personality of the opponent; $V(s_{t+1})$, is the **value function** estimated by the RL-based dialog manager, which gives the best future return estimation supposing the current state is s_{t+1}. It approximates $V(s_{t+1}, z_{t-1}^{-i})$ when it is nearly optimal. β is the **temperature parameter**. Since $\pi_{\texttt{ToM}}$ is normalized as a Boltzmann distribution, when temperature $\beta \to \infty$, $\pi_{\texttt{ToM}}$ is a uniform distribution over the next states; when $\beta \to 0$, $\pi_{\texttt{ToM}}$ is nearly deterministic assigning most probability mass to the s_t with the largest expected value after one-step ToM looking ahead.

4.2 First-Order ToM Policies with Implicit Personality Modeling

We also introduce first-order ToM policy with implicit personality modeling, where we do not have a module explicitly which explicitly predicts the opponent identity z. Instead, we combine the identifier and ToM model in the explicit version, to directly learn $T(s_{t+1}|u_{t-1}^{-i}, s_t, u_t^i)$ without extra supervision. In this case, $\pi_{\texttt{ToM}}(a_t^i|s_{t-1}, u_{t-1}^{-i})$ is proportional to

$$
\exp\left\{ \frac{1}{\beta} \sum_{u_t^i} \underbrace{G(u_t^i|s_t)}_{\text{Generator}} \sum_{s_{t+1}} \underbrace{T(s_{t+1}|u_{t-1}^{-i}, s_t, u_t^i)}_{\text{1st-order ToM}} \underbrace{V(s_{t+1})}_{\text{Value Fn.}} \right\},
$$

where $T(s_{t+1}|u_{t-1}^{-i}, s_t, u_t^i)$ is called the **implicit first-order ToM model**, and the rest of components are similar to the explicit version.

We call $\pi_{\texttt{ToM}}$ a first-order ToM policy, because it utilizes the first-order transition of the opponent, and estimates the expected outcome of performing a certain action which leads to state s_t. The personalities of the opponent are implicitly inferred from the previous utterance u_{t-1}^{-i} and the history s_t. In practice, the summation (expectation) is approximated by Monte Carlo sampling.

Implicit vs Explicit model. We expect both explicit and implicit ToM models to provide several unique benefits. First, co-training the identifier $f(s_{t-1}, u_{t-1}^{-i})$ and the explicit first-order ToM model $T(s_{t+1}|z^{-i}, s_t, u_t^i)$ is expected to have better sample efficiency than the implicit ToM model $T(s_{t+1}|u_{t-1}^{-i}, s_t, u_t^i)$ since it utilizes the prior knowledge that personality identity affects state transition, and is trained with more supervision. Besides, with the personality z^{-i}, the generator and the value functions can also adapt to different populations of opponents. However, the annotations for opponent types are not available for all corpora, therefore the implicit model would be a more general approach.

4.3 Combining the RL Policy as a Prior

After learning the above two ToM models from the corpus, we leverage the pre-trained RL policy as a prior with the 1st-order ToM policy to perform the **inference**. The final policy is given by

$$
\pi(a_t^i|s_{t-1}, z_{t-1}^{-i}) \propto \pi_{\texttt{rl}}(a_t^i|s_{t-1}) \cdot \pi_{\texttt{ToM}}(a_t^i|s_{t-1}, z_{t-1}^{-i}),
$$

where $\pi_{\texttt{rl}}$ is a policy obtained in a previous RL training process (see Section 5).

From a Bayesian point of view, $\pi_{\texttt{rl}}$ can be seen as a prior $\mathbb{P}(a_t^i|s_{t-1})$, and the $\pi_{\texttt{ToM}}$ is analog to the likelihood $\mathbb{P}(\text{best return}|a_t^i, s_{t-1})$ by its definition (not strictly true since it has to be summed up to one) which modifies the probability assignment in $\pi_{\texttt{rl}}$, i.e., the posterior $\mathbb{P}(a_t^i|\text{best return}, s_{t-1})$. This gives the probability that the current agent should move to s_t in order to reach the highest return in the end. $\pi_{\texttt{ToM}}$ modifies the probability assignment in $\pi_{\texttt{rl}}$, when $\beta \to \infty$ in $\pi_{\texttt{ToM}}$, it is equivalent to the original RL policy $\pi_{\texttt{rl}}$.

5 Dialog Managers

We compare three hybrid **dialog managers** combining neural networks and rules to control the flow of dialog:

(1) The **SL+rule** manager employs a LSTM-based network to learn the transitions from s_{t-1} to s_t from corpus. Rules ensure that only deals meeting 70% target are acceptable.

(2) The **RL** manager uses an actor-critic method (Mnih et al., 2016), which contains a policy network with the same neural network architecture as the SL manager, and a value network predicts the future returns given states.

(3) The **ToM** manager uses the first-order ToM policy as described in Section 4. to learn the best response policy $\pi_{\texttt{ToM}}(a_t^i|s_{t-1}, u_{t-1}^{-i})$ which is aware of the opponent's personalities and mental state.

An extra LSTM model is used to encode u_{t-1}^{-1} in both explicit and implicit ToM models, and learn the personality $z_{t-1}^{-i} = \texttt{LSTM}(u_{t-1}^{-1}, s_{t-1})$ in explicit ToM models which encodes a distribution. Note that for all three managers, we applied reasonable hand-crafted rules to prevent unreasonable policies. Specifically, the agent will never offer a price below its bottom line and will reject the opponent's offer if it is worse than its bottom line.

Training and Fine Tuning. We first train the supervised learning (**SL**) manager to minimize a loss function for the dialog act predictions

$$
\mathcal{L}_{SL} = \texttt{CE}_{\text{intent}} + \alpha \cdot \texttt{MSE}_{\text{price}},
$$

which is a linear combination of the cross entropy loss between the predicted intent and the ground truth intent, and the mean squared error between the predicted price and the ground truth price. The

Dialog Act	Definition	Example
greet	say hello or chat randomly.	"Hello I am interested in buying."
inquire	ask any question about product, year, price, usage, etc.	"How long have you had it? Does it come with any of the accessories?"
inform	provide information about the product, year, usage, etc.	"It has the capability to offer great support for those over 6 foot."
propose(price=)	initiate a price or a price range for the product.	"The list price is 600 but am willing to negotiate."
counter(price=)	propose a new price or a new price range (can be the same).	"I'm sorry, I find homeopathy and any other pseudo-science to be profoundly upsetting as well. I'd be willing to go as high as $5100."
counter-noprice	want to propose a new price but do specifically mention a new price.	"I'm sorry, that's far too low. Since you know EC's reputation for quality, you know it's worth more than that."
confirm	ask question about with information to be confirmed.	"Will the chair work for someone who is under 6 feet tall?"
affirm	give an affirmative response to a confirm/propose.	"Yes absolutely the interface is quick and the phone is up to date as far the updates go."
deny	give a negative response to a confirm/propose.	"No, I would expect that you would pick it up."
agree(price=)	make a deal.	"Fair enough. $30 it is!"
disagree(price=)	cannot make a deal.	"I cannot take $65 for something that is worth almost twice that. Sorry, but no deal."
offer(price=)	final offer with price, no utterance	OFFER($65)
accept	final acceptance, no utterance	ACCEPT
reject	final rejection, no utterance	REJECT
quit	leave the negotiation, no utterance	QUIT

Table 2: Our redesigned dialog acts based on the CRAIGSLISTBARGIN dataset, where propose, counter, agree, disagree are four intents must be followed by a price slot, offer, accept, reject, and quit are four terminal dialog acts with no corresponding natural language.

reinforcement learning (**RL**) manager is then fined tuned from the **SL** manager to maximize a reward function described in Section 6, with the actor-critic methods (Mnih et al., 2016). The actor network is initialized as the **SL** manager's LSTM-based network, and the critic network is partially initialized with the same network, followed by a MLP to predict the value.

For the **ToM** manager, we reuse $V(s_{t+1})$ from a well trained **RL** manager's critic network, and fix it during inference. The **implicit first-order ToM model** $T(s_{t+1}|u_{t-1}^{-i}, s_t, u_t^i)$ is directly trained via supervised learning to minimize the same loss $\mathcal{L}_{SL}$. For the **explicit first-order ToM model**, $T(s_{t+1}|z_{t-1}^{-i}, s_t, u_t^i)$, we first train a LSTM-based identifier $z_{t-1}^{-i} = f(s_{t-1}, u_{t-1}^{-i})$, which receives ground truth opponent personality z^{-i} from the corpus during training. $T(s_{t+1}|z_{t-1}^{-i}, s_t, u_t^i)$ is learned with the input from the well-trained identifier.

To obtain the **1st-order ToM policy** for the inference, we approximate the sum (expectation) in π_{ToM} by Monte Carlo sampling with the generator, and discretize the price in a normalized price range. In practice, we found quantizing the price range with 100 units is a good balance between time consumption and the quality of approximation.

6 Experimental Setup

We test our ToM negotiation framework on the CRAIGSLISTBARGAIN (He et al., 2018), which contains 6682 human-human dialogs between a buyer and a seller alternately bargaining for the price of an item on Craigslist.

Ontology. We redesign the ontology of the CRAIGSLISTBARGAIN dataset to support a more diverse dialog act than the original coarse dialog acts (He et al., 2018), which can reflect more ways of mental state change in a negotiation. We used the Microsoft Language Understanding Intelligent Service (LUIS) to relabel the dataset , and merged some similar label types, such as insist and vague-price into counter-noprice, and intro and great into greet. All fifteen dialog acts after our modifications are in Table 2. There are four intents propose, counter, agree, disagree that must be followed by a price slot, and four terminal acts accept, reject, and quit. When an agent takes an offer action, the other agent has to respond with accept or reject. Note that the function of this dialog act is not to capture the full semantic meaning of one utterance, but to serve as a logical skeleton for the dialog.

686

Reward function design. We set the reward r^i for the agent i to be a linear function of the final price, such that the buyer achieves maximal reward of 1 at its target price, the seller achieves maximal reward of 1 at the listing price, and both agents receive zero rewards at the midpoint of the listing price and the target price. When there is no deal, both agents receives equivalent penalty.

Diverse opponent populations. All our negotiation experiments are conducted against variations of the **SL+rule** manager as the opponent. For the variations, we create 7 different opponent populations (id=0∼6) by injecting different rules for changing prices and rendering utterances. Price changing rules are functions of the number of sentences in the conversation history, which model the agreeability and the flexibility of a person. When rendering utterances, we use a template-based language generator as in (He et al., 2018), and insert population-specific tokens in utterances by sampling according to different opponent types.

The **cooperative** population (id=5) will gradually compromise and move its price from the midpoint. The utterances of this population also contain more polite and mild words indicating its negotiable position. The **most aggressive** population (id=0) will insist its price until the end, and utters more stubborn words. The **competitive** population (id=6) compromises from target price slower than the **cooperative**. The other populations will follow price changing curves in between these two extremes, and also have different language properties. The population types are accessible during training as ground truth values of z_i to provide supervision (see Appendix A for details).

Models. The **dialog managers** we compare are described in Section 5. For the utterance **parser**, we use Microsoft Language Understanding Intelligent Service (LUIS) (Williams et al., 2015) with 10 annotated training examples for each dialog act. For the **Generator**, we use a retrieval-based model similar to He et al., 2018 which samples an utterance from the top 10 matched templates.

Evaluation Metrics. We evaluate generated dialogs across four aspects:

1. **Agreement rate** (Ag), which is the percentage of dialogs that reach agreements.

2. **Objective utility** (Ut), which is given by

$$\texttt{Ut}^i = \begin{cases} (P_{deal} - P_{target}^{-i})/\Delta P, & \text{deal;} \\ 0, & \text{no deal} \end{cases}$$

where P_{deal} is the final deal price, the and total price range $\Delta P = P_{target}^i - P_{target}^{-i}$, where P_{target}^i, and P_{target}^{-i} are the extreme target prices of the two agents. Note that this is different from the *subjective utility* of each agent based on only its own price range, which may result in utilities > 1 or < 0 more often.

3. **Deal fairness** (Fa), which is only for completed deals, as $\texttt{Fa}^i = 1 - 2 * |\texttt{Ut}^i - 0.5|$.

4. **Dialog length** (Len), which is the average turns of sample dialogs.

7 Results

Improvement of dialog policy. We evaluate **SL+rule**, **RL**, and our **ToM** model on a mixed population for 4352 dialogs, which contains about 630 dialogs for each population. As shown in Table 3, our explicit **ToM** model consistently achieves the highest agreement rate (Ag), with 56%, 4%, and 20% improvements compared to vanilla RL against cooperative, competitive, and mixed populations, respectively. Though deal agreement is hard for competitive opponents, our explicit **ToM** model achieves more than 30% improvement on the deal utility when interacting with this population. On the mixed population, the reward (Re) for **SL+rule** agent is low, as it is not directly optimized for better reward. **RL** agent improves the Re a lot compared with the **SL+rule** baseline. However, both **ToM** agents achieve better reward even when compared with **RL** agent, which shows the advantage of strategic planning. Besides, unlike the **SL+rule** only pursues high utility when there is a deal, but ends with every low Ag, our **ToM** models best balance both the agreement rate and agent utility of each dialog, and outperforms **SL+rule** and **RL** for all populations.

Implicit vs. explicit models. We found that the implicit **ToM** model can also achieve better Ag and Ut than the baselines for all populations. But the overall performance is slightly worse than the explicit **ToM** model. This can be explained by the fact that the explicit model has more information about the population type during training. One may worry about the potential error cascade issue the explicit ToM models, as we see in Figure 2,

Method	Cooperative Opponents (id=5)				Competitive Opponents (id=6)				Mixed Population (id=0∼6)				
	Ag ↑	Ut ↑	Fa ↑	Len ↓	Ag ↑	Ut ↑	Fa ↑	Len ↓	Ag ↑	Ut ↑	Fa ↑	Len ↓	Re ↑
SL+rule	0.006	0.006	0.00	**10.51**	0.005	0.005	0.00	**10.64**	0.009	0.008	0.00	**10.59**	−0.48
RL	0.57	0.57	0.00	15.48	0.42	0.18	0.32	16.10	0.47	0.38	0.00	15.79	0.00
ToM (implicit)	0.76	0.72	0.00	13.14	0.34	0.20	0.45	14.26	0.48	0.44	0.00	13.87	0.15
ToM (explicit)	**0.88**	**0.78**	**0.03**	11.34	**0.44**	**0.24**	**0.55**	12.10	**0.56**	**0.47**	**0.10**	11.74	**0.16**

Table 3: Agreement rate (Ag), agent utility (Ut), deal fairness (Fa), dialog length (Len), and reward (Re) for dialog managers (SL, RL, implicit and explicit ToM, with the best $\beta = 0.05$) playing against cooperative, competitive, and mixed populations. Small negative fairness scores are truncated as zeros in the table.

Item: Phone, Listing price: $50 **A: ToM(buyer), bottom: $50, top $38** B: Cooperative(seller), bottom: $35, top: $50	*Item: Furniture, Listing price: $50* **A: ToM(buyer), bottom: $50, top: $38** B: Competitive(seller), bottom: $35, top: $50
A: Hello are you selling the blackberry? B: Sure thing! What price are you thinking? **A: Absolutely, but the price is a little higher than I would like** B: My phone is in excellent condition and priced to sell quickly. **A: Would you be willing to take $38.** B: Yea I can do $44 for it. You come pick it up from me though. in person. Thanks **A: I think so, too! Sounds great!** B: OFFER($43) **A: ACCEPT**	**A: Hi. How are you?** B: maybe. what do you have in mind? **A: How about $38 dollars** B: Look, this is an expensive chair if you are willing to buy it somewhere else. I-m only selling it cause I'm moving out of the country.Otherwise I'd keep it but I don't want to just give it away, you know? How about $49? **A: ok. i can do $46 if you split the shipping in half** B: If you come pick it up today, that sounds fair. **A: Sounds fair to me, thank you** B: OFFER($48) **A: ACCEPT**

Table 4: Example dialogs generated by ToM (explicit) against Cooperative (left) / Competitive (right) opponents.

the top 1 accuracy of the identifier in the explicit model is only 69%, though it is significantly above the chance. Our experiment show that even with an imperfect identifier, the explicit model can still outperform an implicit model, which is directly optimized for better performance.

Population-aware strategies. As Table 3 shows, the ToM model can provide more deal fairness (Fa, normalized price difference to the midpoint) to competitive opponents, since they rarely compromise, meanwhile reaching higher Ag and Ut. When opponents are cooperative and easy to negotiate with, our ToM model can achieve much better agent utility by taking advantage of losing some dialog fairness. This implies our ToM model is able to utilize different characteristics of the opponents in the strategy generation.

We provide some sample dialogs from the explicit ToM model in Table 4. When the seller is competitive, the buyer can adaptively raise its price and exchange for additional benefits, e.g., "ok. i can do $46 if you split the shipping in half", to make the deal happen. We note that sometimes the offer prices slightly deviate from the agreed prince in negotiation but the ToM agent still accepts. This may be because the deflects of SL-based opponents is predictable to the ToM agent.
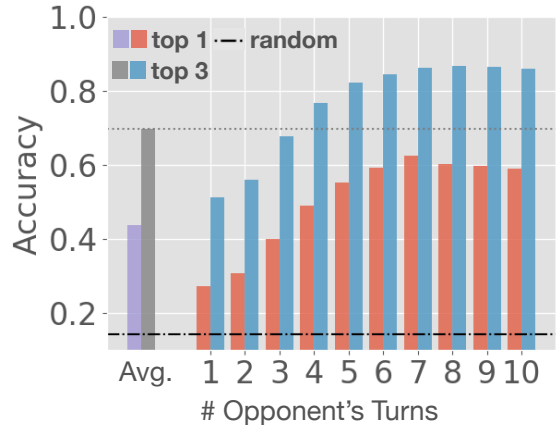

Figure 2: Top 1 and top 3 accuracy of the characteristic identifier (z) during interaction with opponents.

Effectiveness of the opponent identifier. Figure 2 shows the identifier can capture the opponent identities well during interaction. The accuracy of the identifier increases as the dialog progresses. The top 1 accuracy after 6 opponent's turns is above 69%, and the top 3 accuracy is above 84%, where the chance is only 14.2%. The average top 1 accuracy is 43.8% for all turns in 5000 dialogs of different lengths. We also find the explicit ToM models can better prevent overfitting than implicit models. More details are in appendix B.

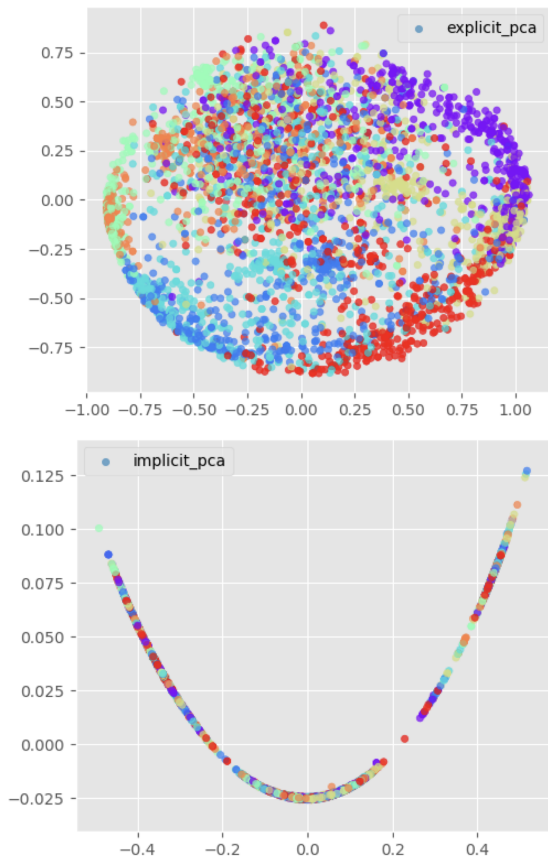**Visualization of population embeddings.** In Figure 3, we show the PCA visualization of the

Figure 3: PCA visualization of latent variables in the explicit (upper) and implicit (lower) ToM models. Colors indicate different opponent populations.

normalized latent variables in both explicit and implicit ToM models. The latent variables are extracted from one layer before the output of the identifier or its equivalence in the implicit model. The explicit ToM model learns embeddings encoding different opponent populations, as the major variances of variable are captured by the difference of opponent populations. However, without extra supervision, the extraction of the population identity is difficult in the implicit ToM model. Further analysis shows that the variances of the latent variables in the implicit ToM model are mainly explained by intent types. We include more detailed analysis and t-SNE visualization in appendix B.

8 Conclusion

In this work, we proposed a novel framework to integrate the concept of Theory of Mind (ToM) into generating task-oriented dialogs. Our approach provides the ability to model and infer personality types of opponents, predict changes in their mental state, and use this information to adapt the agent's high-level strategy in negotiation tasks. We in-

troduced a probabilistic formulation for first-order ToM and introduce two ways to incorporate it into a dialog agent, by 1) explicitly and 2) implicitly modeling the personality of the opponent. We tested our approach on a modified version of the CRAIGSLIST-BARGAIN dataset (He et al., 2018) with diverse opponents. Our experiments show that our method using ToM inference achieves about **20%** higher dialog agreement rate and utility compared to baselines on a mixed population of opponents. When negotiating with the cooperative opponents, the improvement of agreement rate is **54%**. Some directions for future work include developing efficient schemes to approximate the value computation for future states, exploring higher orders of ToM, as well as a tighter integration of ToM into utterance generation and processing.

Ethical Considerations

Our dataset is modified from the open-sourced CRAIGSLISTBARGAIN dataset (He et al., 2018), which consists of negotiation dialogs between sellers and buyers on items from the Craigslist website. The initial dataset was collected using crowd workers on Amazon Mechanical Turk (AMT) playing the role of buyers and sellers. We redesigned the ontology to support more diverse dialog acts than the original coarse dialog acts. We manually labeled 10 examples for each intent, and used the Microsoft Language Understanding Intelligent Service to relabel the whole dataset. We create seven different populations by injecting different rules about changing prices and rendering utterances.

Our paper involves an NLP application that can negotiate with people to reach agreement on deals. It is still at an early exploration stage so we do not expect it will currently cause any negative social impact such as massive job loss. If a mature version of such a system is deployed in the future, it may lead to less fair deals between the AI system and humans, as the system is optimized to find the best strategy that maximizes its own utility. But overall, we believe it will encourage market efficiency.

References

Jacob Andreas and Dan Klein. 2016. Reasoning about pragmatics with neural listeners and speakers. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1173–1182.

Jerome S Bruner. 1981. Intention in the structure of action and interaction. *Advances in infancy research*.

Arjun Chandrasekaran, Deshraj Yadav, Prithvijit Chattopadhyay, Viraj Prabhu, and Devi Parikh. 2017. It takes two to tango: Towards theory of ai's mind. *CoRR*, abs/1704.00717.

Bhuwan Dhingra, Lihong Li, Xiujun Li, Jianfeng Gao, Yun-Nung Chen, Faisal Ahmed, and Li Deng. 2017. Towards end-to-end reinforcement learning of dialogue agents for information access. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 484–495.

Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. 2018. Speaker-follower models for vision-and-language navigation. In *Advances in Neural Information Processing Systems*, pages 3314–3325.

Kallirroi Georgila, James Henderson, and Oliver Lemon. 2006. User simulation for spoken dialogue systems: learning and evaluation. In *INTERSPEECH 2006 - ICSLP, Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA, September 17-21, 2006*. ISCA.

Noah D Goodman and Michael C Frank. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, 20(11):818–829.

Noah D Goodman and Andreas Stuhlmüller. 2013. Knowledge and implicature: Modeling language understanding as social cognition. *Topics in cognitive science*, 5(1):173–184.

Christos Hadjinikolis, Yiannis Siantos, Sanjay Modgil, Elizabeth Black, and Peter McBurney. 2013. Opponent modelling in persuasion dialogues. In *Twenty-Third International Joint Conference on Artificial Intelligence*.

He He and Jordan L. Boyd-Graber. 2016. Opponent modeling in deep reinforcement learning. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pages 1804–1813.

He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. 2018. Decoupling strategy and generation in negotiation dialogues. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2333–2343.

Youngsoo Jang, Jongmin Lee, and Kee-Eung Kim. 2020. Bayes-adaptive monte-carlo planning and learning for goal-oriented dialogues. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7994–8001.

Simon Keizer, Markus Guhe, Heriberto Cuayáhuitl, Ioannis Efstathiou, Klaus-Peter Engelbrecht, Mihai Sorin Dobre, Alex Lascarides, and Oliver Lemon. 2017. Evaluating persuasion strategies and deep reinforcement learning methods for negotiation dialogue agents. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*, pages 480–484.

Sang-Woo Lee, Yu-Jung Heo, and Byoung-Tak Zhang. 2018. Answerer in questioner's mind: information theoretic approach to goal-oriented visual dialog. In *Advances in Neural Information Processing Systems*, pages 2579–2589.

Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. 2016. Asynchronous methods for deep reinforcement learning. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pages 1928–1937.

Will Monroe and Christopher Potts. 2015. Learning in the rational speech acts model. *arXiv preprint arXiv:1510.06807*.

Nir Oren and Timothy J Norman. 2009. Arguing using opponent models. In *International Workshop on Argumentation in Multi-Agent Systems*, pages 160–174. Springer.

Friedrich Burkhard von der Osten, Michael Kirley, and Tim Miller. 2017. The minds of many: Opponent modeling in a stochastic game. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 3845–3851.

David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526.

Neil C. Rabinowitz, Frank Perbet, H. Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew Botvinick. 2018. Machine theory of mind. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 4215–4224.

Tjitze Rienstra, Matthias Thimm, and Nir Oren. 2013. Opponent models with uncertainty for strategic argumentation. In *Twenty-Third International Joint Conference on Artificial Intelligence*.

Verena Rieser and Oliver Lemon. 2006. Cluster-based user simulations for learning dialogue strategies. In *INTERSPEECH 2006 - ICSLP, Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA, September 17-21, 2006*. ISCA.

Homero Roman Roman, Yonatan Bisk, Jesse Thomason, Asli Çelikyilmaz, and Jianfeng Gao. 2020. RMM: A recursive mental model for dialog navigation. *CoRR*, abs/2005.00728.

Sheng Shen, Daniel Fried, Jacob Andreas, and Dan Klein. 2019. Pragmatically informative text generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4060–4067.

Aviv Tamar, Sergey Levine, Pieter Abbeel, Yi Wu, and Garrett Thomas. 2016. Value iteration networks. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 2146–2154.

David Traum, Stacy C Marsella, Jonathan Gratch, Jina Lee, and Arno Hartholt. 2008. Multi-party, multi-issue, multi-strategy negotiation for multi-modal virtual agents. In *International Workshop on Intelligent Virtual Agents*, pages 117–130. Springer.

Tsung-Hsien Wen, David Vandyke, Nikola Mrksic, Milica Gasic, Lina Maria Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve J. Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers*, pages 438–449.

Ying Wen, Yaodong Yang, Rui Luo, Jun Wang, and Wei Pan. 2019. Probabilistic recursive reasoning for multi-agent reinforcement learning. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

Julia White, Jesse Mu, and Noah D. Goodman. 2020. Learning to refer informatively by amortizing pragmatic reasoning. *CoRR*, abs/2006.00418.

Jason D. Williams, Eslam Kamal, Mokhtar Ashour, Hani Amr, Jessica Miller, and Geoffrey Zweig. 2015. Fast and easy language understanding for dialog systems with microsoft language understanding intelligent service (LUIS). In *Proceedings of the SIGDIAL 2015 Conference, The 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2-4 September 2015, Prague, Czech Republic*, pages 159–161.

Heinz Wimmer and Josef Perner. 1983. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103–128.

Steve J. Young, Milica Gasic, Blaise Thomson, and Jason D. Williams. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.

Yiheng Zhou, He He, Alan W. Black, and Yulia Tsvetkov. 2019. A dynamic strategy coach for effective negotiation. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue, SIGdial 2019, Stockholm, Sweden, September 11-13, 2019*, pages 367–378.

Appendix

A   Experimental Setup

To test our proposed framework in a realistic persuasive negotiation setting, we use the CRAIGSLISTBARGAIN dataset (He et al., 2018), which contains 6682 human-human dialogs between a buyer and a seller alternately bargaining for the price of an item on Craigslist. The listed price and a description is presented to both agents, and a private price is assigned to the buyer as the target. We set the reward r^i to be a linear function of the final price, such that the buyer achieves maximal reward of 1 at its target price, the seller achieves maximal reward of 1 at the listing price, and both agents receive zero rewards at the midpoint of the listing price and the target price. When there is no deal, both agents receives equivalent penalty of -0.5.

Ontology   We redesign the ontology of the CRAIGSLISTBARGAIN dataset to support a more diverse dialog act than the original coarse dialog acts (He et al., 2018), which can reflect more ways of mental state change in a negotiation. A dialog act consists of intent and a set of arguments. In our experiments, we only focus on the price as it is the most important goal of this task. All fifteen dialog acts are listed in Table 2. There are four intents propose, counter, agree, disagree that must be followed by a price slot, and accept, reject, and quit are four terminal dialog acts with no utterance. When an agent takes an offer, the other agent has to respond with accept or reject. Note that the function of this dialog act is not to capture the full semantic meaning of one utterance, but to serve as a logic skeleton of the dialog.

System Design   Parser: We use Microsoft Language Understanding Intelligent Service (LUIS) (Williams et al., 2015) with 10 starting training examples for each dialog act in our experiment. **Generator:** We use a retrieval-based model similar to He et al., 2018 which samples an utterance from the top 10 matched templates. We compared three hybrid **dialog managers** combining neural nets and rules to control the flow of the dialog. (1) The **SL** manager employs a neural network to learn the transitions from s_{t-1} to s_t from dataset. We use a sequence model with two-layer LSTM with 300 hidden units for both the encoder and the decoder. (2) The **RL** manager uses an actor-critic method (Mnih et al., 2016), which contains a policy network with the same neural network architecture as the SL manager, and a value network predicts the cumulative reward given input states. The **RL** manager also learns $\pi^i(s_t|s_{t-1})$ but with the goal of maximizing the total reward. (3) The **ToM** manager uses the first-order ToM policy as described in 4 to learn the best responding policy $\pi_{\mathrm{ToM}}(s_t|s_{t-1}, u_{t-1}^{-i})$ with the awareness of the opponent's characteristics and mental state change. An extra LSTM model is used to learn the characteristic identity $z_{t-1}^{-i} = f(s_{t-1}, u_{t-1}^{-i})$ in the explicit ToM model. For all three managers, we improve the learned policy by enforcing hand-craft rules. For example, the agent should never offer price below its bottom line and reject the opponent's offer if it is worse than its bottom line.

Populations of Opponents   When playing against with a **SL** manager, we create 7 different populations of opponents by injecting rules for changing the price and rendering utterance. Price changing rules are functions of the number of sentences in the conversation history, which model the agreeability and the flexibility of a person. The agreeability of a person is reflected in the range of relative prices (utility) at which a deal could be made. For example, a competitive opponent has a higher lower bound on the price, while a cooperative opponent has a lower initial price. The flexibility of a person is reflected in the slope and convexity of the price-changing rules. The price changing function for the most aggressive and stubborn opponent has a zero slope, encouraging them to insist on their initial price until the end of the dialog. The more determined a seller is, the more concave the price changing function becomes.

When rendering utterances, we use a template-based language generator as in (He et al., 2018), and insert population-specific tokens in utterances by sampling according to different opponent types. For example, in the utterances from a competitive opponent, words like "afraid" or "unfortunately" appear more often, while words like "great" or "ok" will appear more frequently in the utterances from a cooperative opponent. Utterances of different populations should follow different distributions, and these sets of tokens are designed for this purpose.

We vary the price range, slope, and convexity to

obtain the different behaviors for the seven different opponent types. The **mildest** population will gradually compromise and lower its price (or raise its price if it is buyer). The utterances of this population also contain more polite and mild words indicating its negotiable position. And the **most aggressive** population will insist its price until the end, and utters more stubborn words. The other five populations will follow different price changing curves in between these two extremes, and also have different language properties. All of these populations will deal at a certain price range, which depends on latest proposal price and current dialog length in different ways.

Training and Fine Tuning We train the **SL** manager on 5000 dialogs for 20 epochs and choose the model with the lowest validation loss. The **RL** manager is fine-tuned from a well-trained **SL** agent by playing against itself. We choose the model with the highest reward.

For the **ToM** manager, the value function $V(s_{t+1})$ is borrowed from a well trained **RL** manager, and fixed during inference. The implicit **first-order ToM model** $T(s_{t+1}|u_{t-1}^{-i}, s_t, u_t^i)$ is trained in a similar way as the **SL** manager. To obtain the explicit first-order ToM model $T(s_{t+1}|z_{t-1}^{-i}, s_t, u_t^i)$, we co-trained it with a LSTM-based identifier $z_{t-1}^{-i} = f(s_{t-1}, u_{t-1}^{-i})$ for 2,000 episodes. Each run on training each manager was performed using a single NVIDIA GTX 2080 Ti GPU with 16GB RAM in approximate 2 hours. All the managers were trained using Adam with learning rate 0.001. For the **ToM** manager, hyperparameter β was randomly searched in range of $0.05, 0.1, 1, 10$, and the setting with the best results was $\beta = 0.05$.
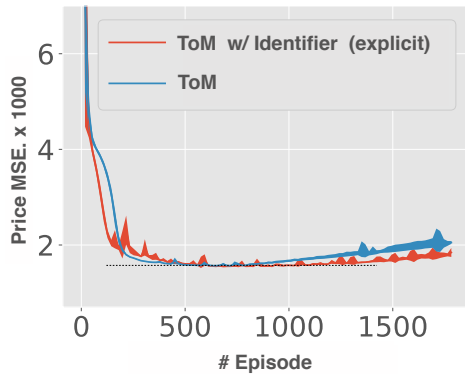
B Additional Experimental Results

Figure 4: The validation price MSE loss curves of implicit and explicit first-order ToM models. The explicit model with an identifier has slightly better sample efficiency and better prevents overfitting.

B.1 Comparison of Implicit and Explicit ToM models

We compared the implicit and the explicit ToM models as described in Appendix Section 4 in the main article. Here additional Figure 4 shows the validation mean squared error between the predicted price and the ground truth price of the opponent in the next turn (if exists) over 3,584 dialogs. Two models can both be trained well to perform this one-step prediction, while the explicit model with an identifier has slightly better sample efficiency and better prevents overfitting. This supports our hypothesise that the prior of different types of opponents is important.

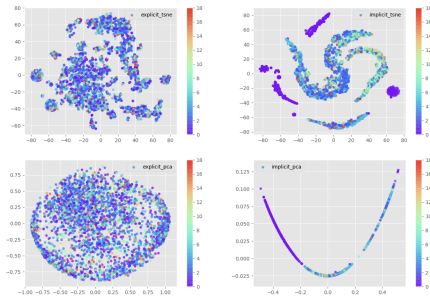## B.2 Visualization of Latent Variables

Figure 5: t-SNE and PCA visualization of latent variables in the explicit ToM model (left) and the implicit ToM model (right). Colors indicate different intents taken by the the agent.
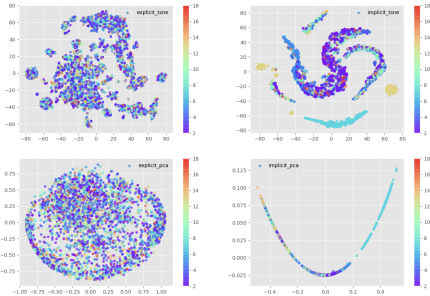

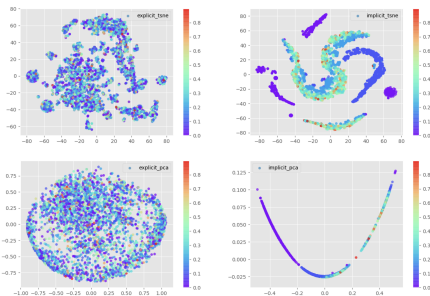Figure 6: The same plot as Figure 5. Colors indicate different intents taken by the the opponent.

Figure 7: The same plot as Figure 5. Colors indicate numbers of rounds of interaction.