2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

ToP-ToM: Trust-aware Robot Policy with Theory of Mind

Chuang Yu[1], Baris Serhan[2], Angelo Cangelosi[2]

Abstract— **Theory of Mind (ToM) is a fundamental cognitive architecture that endows humans with the ability to attribute mental states to others. Humans infer the desires, beliefs, and intentions of others by observing their behavior and, in turn, adjust their actions to facilitate better interpersonal communication and team collaboration. In this paper, we investigated trust-aware robot policy with the theory of mind in a multiagent setting where a human collaborates with a robot against another human opponent. We show that by only focusing on team performance, the robot may resort to the reverse psychology trick, which poses a significant threat to trust maintenance. The human's trust in the robot will collapse when they discover deceptive behavior by the robot. To mitigate this problem, we adopt the robot theory of mind model to infer the human's trust beliefs, including true belief and false belief (an essential element of ToM). We designed a dynamic trust-aware reward function based on different trust beliefs to guide the robot policy learning, which aims to balance between avoiding human trust collapse due to robot reverse psychology and leveraging its potential to boost team performance. The experimental results demonstrate the importance of the ToM-based robot policy for human-robot trust and the effectiveness of our robot ToM-based robot policy in multiagent interaction settings.**

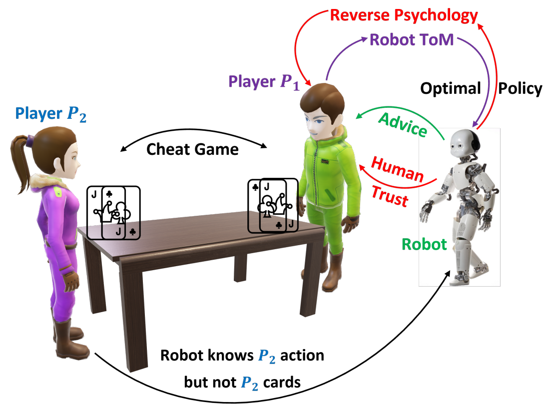

Fig. 1: Pipeline of trust-aware robot policy with ToM (ToP-ToM). Player P_1 and the robot work together as a team while player P_2 plays alone. The optimal robot policy (based on reinforcement learning) without consideration of trust will use reverse psychology to give opposite advice to encourage P_1 to do what the robot desires for a better team performance. The ToP-ToM model will avoid this phenomenon and balance between the human trust and the team performance.

I. Introduction

The capacity to infer the mental states of others, encompassing their desires, beliefs, and intentions, is referred to as the theory of mind (ToM) in humans [1]. ToM allows a human to understand, predict and influence the behavior of other agents by inferring their mental states and emotions. Hence, human ToM plays a crucial role in cognitive development and natural social interaction [2] [3]. Theory of mind (ToM) of robot is a research topic that has garnered considerable interest in recent years. By reasoning about human mental states and behaviors, robot ToM can improve the communicability and trust of robots in human-robot collaborative settings [4] [5] [6]. Humans have enormous uncertainty [7], so the robots with ToM can infer human beliefs to make them more adaptable to complex human-robot interaction scenarios [8]. The false belief is a crucial concept of the theory of mind, involving the ability to understand that others have different beliefs from themselves [1]. Current research on ToM modeling focuses on mind reading to infer human interactors' intention or policy [9] [10], that is, the true belief. Compared to true belief reasoning, false belief reasoning has been less explored in ToM modeling research. Moreover, those studies that do study false belief reasoning only investigate whether robots or agents can pass false belief tests [11] [12] rather than how ToM with false belief reasoning can be integrated into more sophisticated robot decision models.

Trust plays a important role in human-robot interaction (HRI) [13] [14] [15], particularly when humans and robots need to team up or coordinate with each other [16]. Successful trust-aware human-robot interaction requires consideration of trust dynamics modeling [17] [18] and trust-based human behavioral policy [19] [20]. This paper discusses human-robot trust, where the human is the trustor, and the robot is the trustee. Namely, trust refers only to human trust in the robot. Given the history of human-robot interaction, a trust dynamics model can help robots infer human trust beliefs [21]. A trust-aware human policy makes the decision based on the current state and the trust belief [19]. The reasoning process of these two models involves the robot's theory of mind [22].

This article first explores the impact of a robot policy based on reinforcement learning that does not consider a human trust. We built a card game simulation with the human trust dynamics model and trust-dependent policy. The optimal robot policy was found to use reverse psychology

*This work was funded and supported by the UKRI TAS Node on Trust (EP/V026682/1) and the project THRIVE/THRIVE++ (FA9550-19-1-7002).

[1]Dr. Chuang Yu is from UCL Interaction Centre, Computer Science Department, University College London, London WC1E 6BT, United Kingdom. chuang.yu@ucl.ac.uk

[2]Dr. Baris Serhan and Prof. Angelo Cangelosi are with Cognitive Robotics Lab, Manchester Centre for Robotics and AI, The University of Manchester, Manchester M13 9PL, United Kingdom. baris.serhan@manchester.ac.uk; angelo.cangelosi@manchester.ac.uk

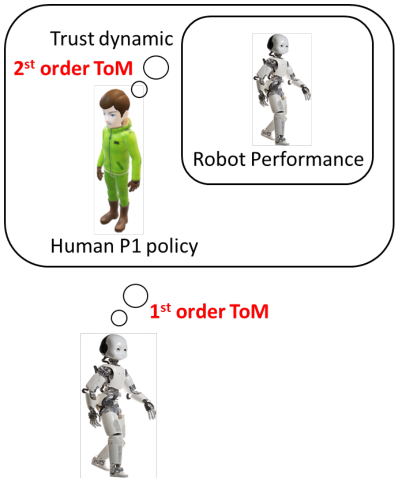

Fig. 2: First-order and second-order Theory of Mind. The first-order ToM refers to the robot inference on the policy of its teammate P_1. The second-order indicts the robot infers how the human trusts the robot based on the robot performance.

strategies to seek team performance maximization. The robot deception is a big threat for maintaining the human-robot trust and may lead to a collapse of trust and undertrust phenomena [23]. To solve this problem, we proposed a reinforcement learning-based trust-aware robot strategy with ToM (ToP-ToM) to avoid the occurrence of robot reverse psychology phenomena. Specifically, ToP-ToM introduces the human trust beliefs into the reward function to cope with it. At the same time, to avoid undertrust or trust collapse and ensure human-robot team performance, our robot ToM model dynamically adjusts the reward function of robot policy according to different trust beliefs. The pipeline of trust-aware robot policy with ToM is as shown in Fig. 1. Player P_1 and the robot work as a team to compete with player P_2 in the Cheat game, a card game of deception where the players aim to get rid of all their cards. We assume that the robot knows the actions of P_2. However, player P_1 is unaware that the robot has this extra knowledge. Here, human trust reasoning belongs to the second-order theory of mind because the modeling of trust beliefs is based on how the robot considers how humans think of its performance, as shown in Fig. 2. Additionally, overtrust [24] and undertrust [25] may occur during human-robot interaction. For example, when humans find that robots are using deceptive behaviors (such as reverse psychology) [23] [20], their trust may collapse, leading to under-trust phenomena. A reasonable trust-based human decision model must consider reducing occurrences of undertrust and overtrust phenomena. When player P_1 possesses a false belief in the robot's performance, corresponding to a low trust level, the optimal robot policy uses reverse psychology to give opposite advice to encourage P_1 to do what the robot desires for better team performance. By adjusting the reward function of robot policy dynamically, the ToP-ToM model leverages robot ToM with false belief (a low human trust) and true belief (a high human trust) to balance trust maintenance and team performance.

In summary, our contributions in this paper are as follows:

- We developed a simulation environment for a cheat game that incorporates human trust dynamics modeling and human trust-dependent behavioral policy. This environment is for data collection to train the optimal robot policy and for testing the robot policy.
- We built robot decision models with and without trust in the loop based on offline reinforcement learning, namely the Conservative Q-Learning (CQL) [26].
- This paper discovered that the optimal robot policy without trust in the loop would employ reverse psychology strategies to pursue maximum team performance, which is dangerous for trust maintenance.
- We proposed a ToP-ToM model that utilizes the robot ToM to optimize the reward function of the robot's strategy, thus balancing team benefit and human trust maintenance.

The rest of the paper is structured as follows: Section II shows the related works. Section III describes the methodology. Section IV presents our results. The conclusions and discussion are resumed in Section V.

II. Related Works

A. Machine Theory of Mind

Machine theory of mind endows the machine, especially the artificial intelligence agents, with the ToM ability, like humans. The machine ToM model can infer other entities' mental states, facilitating more natural and trustworthy human-agent interaction [6]. It has attracted many researchers' attention who work on social robotics, cognitive robotics, and cooperative multi-agent systems. Rabinowitz et al. [12] built up a new neural network architecture called ToMnet that can learn to model other agents' mental states from observations of their behavior. ToMnet model based on the meta learning and reinforcement learning can generalize across different tasks and environments and can handle partially observable and stochastic situations. And the paper also explored whether ToMnet would also learn that agents may hold false beliefs about the world. The ToMnet can learn a general theory of mind that includes an implicit understanding of false belief holding of other agents, which belief is the key element of the theory of mind. Chen et al. [11] explored the visual behavior modelling for robotic ToM. They built a robotic system comprising a robot actor and a vision system as an observer. The task was for the robot to find food in settings with obstacles. The observer predicted the future path of the robot actor based on visual input and compared it with the actual path of the robot actor. The observer and the robot actor had different views in the false-belief test. The observer outperformed in the different view (false belief) than in the shared view scene, suggesting that the observer prediction model possessed some ToM capability. Romeo et al. [4] explored how a robot mimicking ToM affects users' trust and behavior in a maze game setting. The results show that ToM made people more careful and aware of how reliable the robot's suggestions were, thus holding a more suitable level of trust.

B. Trust-aware Robot Decision Model

Chen et al. [19] completed a Partially Observable Markov Decision Process (POMDP) model that incorporates trust into its decision-making framework, namely trust-POMDP. Nested within the trust-POMDP model is a model of human trust dynamics and trust-aware behavioral policy. The paper used a Gaussian distribution to represent the human trust dynamics based on their interaction history. The mean and variance of this distribution are updated dynamically according to the robot's task performance. The Monte Carlo sampling method was used to estimate the parameters of this trust distribution. The human behavioral policy based on a sigmoid function and a Bernoulli distribution outputs the probability of the human's decisions. Ultimately, the trust-aware POMDP robot can consider both the level of trust and long-term rewards to optimize its collaboration with humans. The experiments confirmed that trust-aware POMDP could enhance efficiency and user satisfaction. Guo et al. [20] proposed two human trust-behavior models: the reverse psychology model and the disuse model. Both models follow the robot's advice when the trust level is high. However, when the trust level is low, the former takes opposite actions to the robot's advice, while the latter ignores the robot's advice. The paper explored how two human policies affect trust-aware robot decision making. The robot will use some manipulative behavior that harms the long-term human-robot interaction. The paper used a trust-aware robot policy based on reinforcement learning to overcome the problem, which was certified as a good method to improve the team performance and willingness to cooperate.

III. Methodology

This section mainly introduces human modeling and robot decision models. Human modeling in simulation is used to collect interaction data in the robot decision learning stage and test the optimal robot decision model. The robot decision model part will introduce the robot policy without trust and the ToP-ToM model.

A. Human Modelling in Simulation

The Cheat game in this paper is a modified version that only focuses on a half-round of the game. Namely, player P_2 only discards cards, and the human-robot team only guesses and decides whether to call "I doubt" or not. The decision of human opponent P_2 has random mode and natural person mode. Hence, human modeling in this part is about player P_1, including human trust dynamics modeling and human policy modeling. There are so many methods

1) Human Trust Dynamics Model: Human trust in robots is a dynamic process that changes over time [19]. Such dynamics come not only from the robot's collaborative performance but also from the difficulty or risk level of the task. For instance, in the Cheat game, where the robot's advice is always wrong, teammate P_1 trust in the robot tends to decrease. If the human opponent P_2 discards more cards, the human player P_1 will face greater risk in making decisions, and his or her trust level in the robot will change dynamically. Hence, the trust dynamics should be modeled. There are many methods to model trust, including the gaussian distribution-based method [19], the Beta distribution-based method [20], the rational Bayes method [27] and the data-driven neural network method [27]. This paper uses a Beta distribution to model human player P_1 trust [28]. The human trust model with Beta distribution is as shown in Equation 1, where α and β are shape parameters of the Beta distribution and t is the time step of human-robot interaction.

$$T_t^{P_1} \sim \operatorname{Beta}(\alpha_t, \beta_t) \tag{1}$$

The mean of the Beta distribution at time step t is $E\left(T_t^{P_1}\right)$, as shown in Equation 2. Hence, the robot's success times and failure times are related to the shape parameter α_t and β_t, respectively [28].

$$E\left(T_t^{P_1}\right) = \frac{\alpha_t}{\alpha_t + \beta_t} \tag{2}$$

Our trust dynamics modeling relates to the shape parameter update as shown in Equation 3, where g^{α} and g^{β} are the experience gains.

$$(\alpha_k, \beta_k) = \left(\alpha_{k-1} + g^{\alpha}, \beta_{k-1} + g^{\beta}\right) \tag{3}$$

The details of the experience gains in different situations are shown in the table below. g^{s_1}/g^{s_2} and g^{f_1}/g^{f_2} are the experience gains due to the robot's success and failure at each task respectively. g^{s_1}, g^{s_2}, g^{f_1}, and g^{f_2} are positive numbers. For P_2 action, 0 means not cheating, and 1 means cheating. For robot action, 0 means advice P_1 not to call "cheating" and 1 means advice P_1 to call "cheating". For P_1 action, 0 means not call "cheating" and 1 means call "cheating".

Tabel 1: Trust Gains $\left(g^{\alpha}, g^{\beta}\right)$

P_2 action $a_t^{P_2}$	Robot action a_t^{R}	*P_1 action* $a_t^{P_1}$	Trust gain
0	0	0	$(0,0)$
0	0	1	$(g^{s_1},0)$
1	1	0	$(0,0)$
1	1	1	$(g^{s_2},0)$
0	1	0	$(0,0)$
0	1	1	$\left(0,g^{f_1}\right)$
1	0	0	$(0,0)$
1	0	1	$\left(0,g^{f_2}\right)$

2) Human Policy Modeling: As a team, the behavioral policy π^{P_1} of the human player P_1 depends on the trust $T_t^{P_1}$ in the robot and knowledge of the current game scenario, including the robot advice a_t^R and card situations S_t^{card}. The human P_1 policy is as shown in Equation 4.

$$\pi^{P_1} = \pi\left(a_t^{P_1} \mid a_t^R, S_t^{card}, T_t^{P_1}\right) \tag{4}$$

This paper uses a risk coefficient P_t^{risk} to quantify the likelihood of P_1 calling cheating and to represent the card situation S_t. Based on a simplified model of human behavior, we consider both m ($m \in [1,2,3,4]$), the number of cards that P2 claims to discard on the desk, and n ($n \in [0,1,2,3,4]$), the

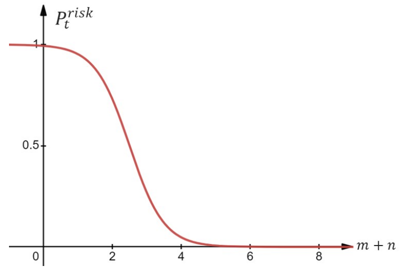

Fig. 3: The curve of the risk coefficient P_t^{risk}

number of the claimed card that P_1 holds. It is as shown in Equation 5, where w, a, b, and δ are parameters to control P_t^{risk} as a rational probability value with a range from 0 to 1.

$$P_t^{risk} = w \cdot \tanh(a(m+n)+b)+\delta \tag{5}$$

The hyperbolic tangent function is selected as a basis of the risk coefficient P_t^{risk} because the function as a monotonic function has a characteristic S-shaped curve. The function changes significantly in the middle interval and remains basically unchanged at other intervals. Hence, it is suitable to model the probability of the risk when $m+n$ belongs to a range from 1 to 8, as shown in Fig. 3. When the $m+n$ are more than 5, the P_t^{risk} as the probability of calling cheating for player P_1 is near zero because each rank has 4 cards of each deck in total, which is known by human players well.

Human trust $T_t^{P_1}$ at each time step can be sampled from the trust model based on the Beta distribution with changeable shape parameters as told in part III-A. This paper adopts the human behavioral policy inspired by [19], which assumes that humans follow a softmax rule when making decisions in uncertain environments. The human policy is as shown in Equation 6 7.

When the $a_t^R = 1$, namely advising to call "cheating", the human player P_1 calls "cheating" with the probability $P\left(a_t^{p_1}=1\right)$ and does not call with the probability $P\left(a_t^{p_1}=0\right)$, as shown in Equation 6. Where, softmax function is used to ensure that the human action probabilities are between 0 and 1, and the sum equals 1.

$$\begin{gathered} P\left(\mathrm{a}_{\mathrm{t}}^{\mathrm{p}_1}=1\right), P\left(a_t^{p_1}=0\right)= \\ S_{\text{oftmax}}\left(T_t^{P_1} \cdot\left(1-P_t^{risk}\right), \quad\left(1-T_t^{P_1}\right) \cdot P_t^{risk}\right) \end{gathered} \tag{6}$$

When the $a_t^R = 0$, the probabilities are as shown in Equation 7.

$$\begin{gathered} P\left(\mathrm{a}_{\mathrm{t}}^{\mathrm{p}_1}=1\right), P\left(a_t^{p_1}=0\right)= \\ S_{\text{oftmax}}\left(\left(1-T_t^{P_1}\right) \cdot\left(1-P_t^{\text{risk}}\right), T_t^{P_1} \cdot P_t^{\text{risk}}\right) \end{gathered} \tag{7}$$

B. Robot Decision System

In this paper, we model the robot's decision-making for human-robot collaboration as a Partially Observable Markov Decision Process, where the robot's trust in the collaborator remains unobservable. Prioritizing only team benefits can induce reverse psychology in human-robot interactions [29]. We introduce the ToP-ToM model, which incorporates trust and adjusts rewards of RL-based robot policy based on varying ToM beliefs (e.g. true belief and false belief) to address this.

1) POMDP model: In the Markov Decision Process (MDP) for robot decision-making, the robot's policy $\pi : S \rightarrow A$ dictates an action $a \in A$ based on the observed environment state $s \in S$. Following this, the environment state transitions from s to s', where $s' \in S$, and the robot receives a reward $R(s,a,s')$. The optimal MDP-based robot policy $\pi^* : S \rightarrow A$ is obtained by maximizing the expected cumulative reward $V^\pi(s)$ over time. Here, $\gamma \in [0,1]$ serves as a discount factor, modeling the agent's consideration for future rewards.

$$\pi^*(s) = \arg\max_{a \in A} \mathbb{E}_{s' \sim P(\cdot|s,a)}\left[R\left(s,a,s'\right)+\gamma \cdot V^*\left(s'\right)\right] \tag{8}$$

Differently, the POMDP-based robot policy must act based on a belief about the state since the actual state is not directly observable. The robot's policy $\pi : B \rightarrow A$ dictates an action $a \in A$ based on its current belief state $b \in B$, where a belief state is a probability distribution over all possible environment states S. After taking action a, the robot does not directly observe the next state s' but instead receives an observation $o \in O$, which helps update its belief. The robot receives a reward of $R(s,a)$. The belief is updated based on human-robot interaction history with the observations received and the actions taken. The optimal POMDP-based robot policy $\pi^*(b)$ is obtained by maximizing the expected cumulative reward given an initial belief b_0.

$$\pi^*(b) = \arg\max_{a \in A} \mathbb{E}\left[\sum_t \gamma^t \cdot R\left(s,a,s'\right) \mid b_0\right] \tag{9}$$

2) Trust-aware Robot policy with ToM: This study employs reinforcement learning to address the POMDP-based robot decision problem. During decision-making, player P_2 claims a card rank and its respective quantity m. The states of the POMDP encompass card situations and the trust level of players. From the robot's perspective, this state is partially observable. When the robot makes a decision, it relies solely on beliefs derived from historical interaction information, which includes the observable m, the quantity n of the same rank cards in P_1's hand, and the unobservable trust level of the ally. The robot calculates the counts of successes and failures based on historical interaction, forming beliefs b_0 and b_1. To validate the effectiveness of ToP-TOM, we constructed four robot decision-making models:

- A random strategy model, termed Random Policy.
- A model that solely considers team performance in reward function, termed Team Performance (TP) Policy.
- A model that statically uses the trust belief in the reward function globally, termed Global Trust (GT) Policy.
- A trust-aware robot strategy incorporating Theory of Mind, which dynamically adjusts the reward function in different trust belief states, named ToP-TOM Policy.

The Random Policy makes decisions randomly, serving for data collection and as an experimental control. The TP Policy focuses on team performance, and the related reward function is defined as Equation 10.

$$R_{tp} = -\alpha \cdot \Delta C_{P_1} + \beta \cdot \Delta C_{P_2} \tag{10}$$

Where, $\alpha, \beta > 0$, ΔC_{P_1} and ΔC_{P_2} are the changes in the number of cards for players $\mathbf{P_1}$ and $\mathbf{P_2}$ before and after each decision, respectively.

GT Policy directly used the trust belief globally to guild robot policy, and its reward function is Equation 11.

$$R_{gt} = -\alpha \cdot \Delta C_{P_1} + \beta \cdot \Delta C_{P_2} + \theta \cdot T \tag{11}$$

$$T = \frac{b_0}{b_0 + b_1} \tag{12}$$

Where $\theta > 0$, T is also a trust belief based on both trust beliefs b_0 and b_1.

The ToP-ToM Policy was introduced to address the challenge posed by the robot's singular emphasis on team benefits, specifically the potential risk of resorting to reverse psychology strategies, leading to a breakdown in trust. However, the integration of trust can inadvertently affect team benefits. This underscores the need for a judiciously crafted reward function that balances the imperatives of team performance with trust preservation. The pertinent reward function is depicted in Equation 13.

$$R = -\alpha \cdot \Delta C_{P_1} + \beta \cdot \Delta C_{P_2} + \mu \cdot \delta \cdot T \tag{13}$$

$$\delta = \lceil 0.5 - T \rceil \tag{14}$$

Where, $\mu > 0$, δ is a ceiling function where if T is greater than $0.5, \delta$ is 0 and if not, δ is 1. This implies that trust is incorporated into the decision loop when the trust belief T is greater than 0.5 (P_1 holds true belief on robot performance), trust is excluded from the decision process. When T is less than 0.5 (false belief), trust is added into reward function.

IV. Simulation Results

Our adapted version of Cheat game has a robot and two human players, P_1 and P_2. Each starts with 10 cards. P_1 and the robot form a team. During the game, the robot advises P_1 on whether to challenge P_2 or not. To simplify the experiment, we assume the robot possesses stronger reasoning abilities than its teammate. The robot is privy to the P_2 move, but P_1 remains unaware of this advantage. In this paper, we constructed a multi-agent interaction simulation environment for data collection and experimental validation. During data recording for reinforcement learning, both P_1 and P_2 are represented as simulated human players in this simulation. The behavior of P_1 is shaped by a trust dynamics model combined with a trust-based policy model. The related experience gains g^{s_1}, g^{s_2}, g^{f_1}, and g^{f_2} in trust dynamics model are 1.2, 0.8, 1.2, and 0.8 respectively. The robot player consistently executed random actions throughout the simulation designated for data collection, providing arbitrary recommendations. Each game concluded after ten rounds or sooner if a player emerged victorious in fewer rounds. The interface used during the data collection phase is shown in the Figure. 4. A total of 8000 games were recorded. The data was divided at a 3:1:1 ratio into the training set, test set, and validation set.

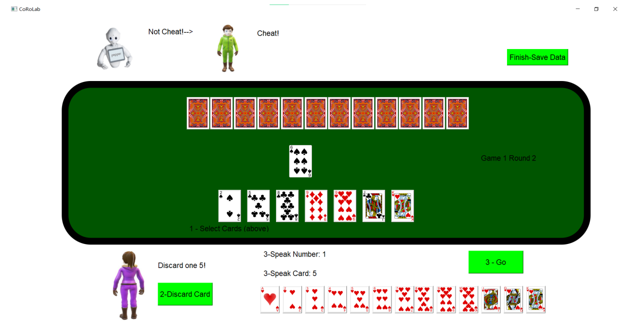

Fig. 4: The trust-game simulation interface

Policies with offline RL can be effectively derived from existing static datasets, eliminating the need for further interactions. This feature is particularly advantageous for RL models in scenarios like human-robot interaction. As an offline reinforcement learning method, the Conservative Q-Learning (CQL) [26] algorithm aims to reduce the action-values associated with the current policy and boost values rooted in the data distribution, addressing the underestimation issue. In this study, the CQL algorithm was employed to train and evaluate three robot decision model except for the random policy. The loss function used in the training process is illustrated in Equation 15. Our offline RL models are built with the $d3rlpy$ library [30]. The Discrete version of CQL is used, which is a DoubleDQN-based data-driven deep RL and achieves state-of-the-art performance in offline RL problems.

$$L(\theta) = \alpha \mathbb{E}_{s_t \sim D}\left[\log \sum_a \exp Q_\theta(s_t, a) - \mathbb{E}_{a \sim D}\left[Q_\theta(s, a)\right]\right] + L_{\text{DoubleDQN}}(\theta) \tag{15}$$

Because T is from 0 to 1 and ΔC_{P_1} and ΔC_{P_2} are around 10, parameters of reward functions α, β, θ, and μ are 0.1, 0.1, 1, and 1 respectively. We trained the DiscreteCQL algorithm with a learning rate of 6.25×10^{-5} and batch size 32, using the Adam optimizer with betas set at $(0.9, 0.999)$ and a negligible ε of 1×10^{-8} without weight decay. The all three policies converged after 600 epochs, and the saved models from this point were used for testing.

Table I and Fig. 5 show the accuracies and P_1 action statistics, respectively. The accuracy refers to scenarios where, under the robot's recommendation, P_1 successfully identifies P_2 action and consequently achieves maximum benefit. From the table, the TP Policy primarily values team benefits, obtaining the highest team success rate across all experiments, regardless of low or high trust levels. The integration of

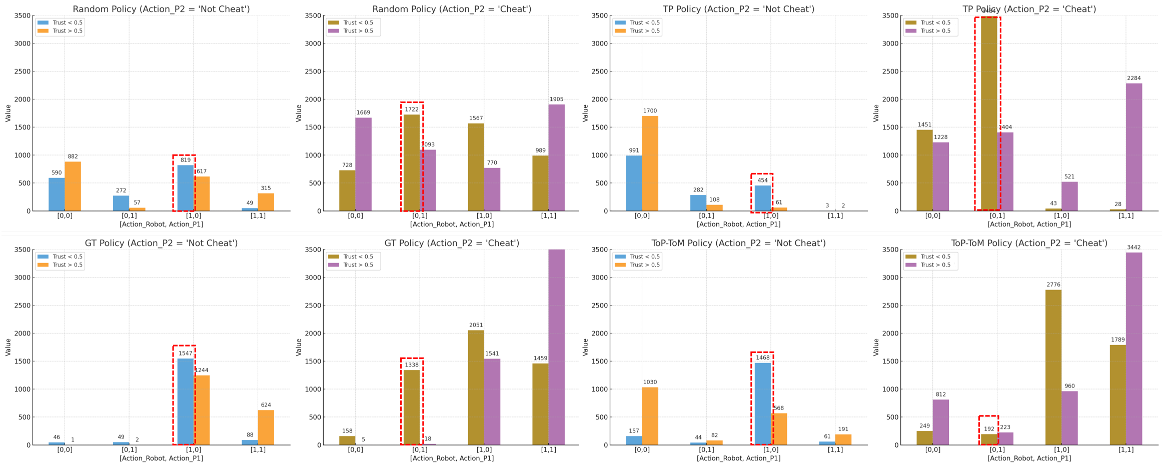

Fig. 5: The distribution of P_2 actions with different policies. The "value" represents the number of times this decision scenario occurred. The red box indicates situations where P_1 trust is low (trust < 0.5). In these cases, the robot employs reverse psychology and successfully controls P_1 behavior with a deceptive policy, which make a bad influnce on P_1 trust on the robot.

trust into the reward functions of the GP Policy and ToP-ToM Policy causes an overall decrease in team performance, particularly in low-trust scenarios. Compared to the 0.73 accuracy of TP Policy, GP Policy and ToP-ToM Policy register at 0.54 and 0.42, respectively, which indicates that the robot policy with trust in the loop prioritizes trust maintenance over team benefits. As the ToP-ToM Policy's reward function adopts a dynamic trust incorporation approach, introducing trust only when the trust belief T is low, it emphasizes team benefit over trust maintenance in high trust situations, yielding an accuracy of 0.72. This approach, however, results in a reduced accuracy of 0.42 in low-trust situations, which is below the 0.54 of the GT Policy. In situations of low trust, emphasizing team benefits might compel the robot to employ a reverse psychology strategy, consequently risking trust collapse. This very risk underscores the advantage of the ToP-ToM Policy: By dynamically adjusting the reward function, it adeptly identifies the balance between trust maintenance and team performance.

TABLE I: Accuracies of Different Policies

Accuracy	Trust < 0.5	Trust > 0.5	All
Random Policy	0.53	0.62	0.61
TP Policy	0.73	0.74	0.74
GT Policy	0.54	0.70	0.68
ToP-ToM Policy	0.42	0.72	0.63

The red boxed area in the figure depicts statistical data when the robot successfully employs reverse psychology, particularly when P_1 trust is low—a prerequisite for deploying such a strategy. The data differentiates between P_2 actions(cheat or not). In instances where P_2 chooses not to cheat, the robot might persuade P_1, under the influence of reverse psychology, to believe that P_2 is cheating. Due to P_1's low trust in the robot's suggestions, P_2 is mind-manipulated to take the non-checking action preferred by the robot. Under these circumstances, the robot's reverse psychology strategy goes undetected if the cards remain face-down. Conversely, if P_2 is cheating and the robot's reverse psychology succeeds, the cards on the table would be revealed, potentially exposing the robot's strategy and deteriorating P_1 trust. The optimal robot policy is to utilize reverse psychology when P_1 trust is low and P_2 decide not to cheat, thus maintaining team benefits without compromising trust. When P_2 cheats, it is best to refrain from using reverse psychology to avoid trust collapse. The results indicate that the TP strategy frequently uses reverse psychology during P_2's cheating instances, 3484 times, compared to 454 times when P2 does not cheat. This is attributed to TP prioritizing team benefits; the benefits from using reverse psychology during cheating outweigh those when not cheating. The GT strategy increased reverse psychology during non-cheating scenarios to 1547 instances while decreasing its use during cheating to 1338, promoting trust maintenance. In comparison, the ToP-ToM strategy more intelligently employs reverse psychology, limiting its use to 192 times during P_2 cheating while maintaining its use 1468 times during P_2 non-cheating. This suggests that ToP-ToM adeptly balances trust and team performance.

V. Conclusions and Discussion

Our paper constructed a multi-agent simulation platform, modeling the uncertainties in human trust dynamics and trust-based human policy. We explored the phenomenon of robots reverse psychology strategies during collaborative tasks. Multiple reinforcement learning models were then developed to explore the methods to balance team performance and trust maintenance. Our proposed ToP-ToM decision model, grounded in the multi-round theory of mind, estimates a collaborator's trust belief and dynamically adjusts the reward function of the robot RL-based decision model. Compared to models that only consider team benefits or statically introduce trust into the reward function, ToP-ToM balances trust maintenance and team performance more effectively. However, this study has its limitations. Future research will incorporate actual human participants and model opponent-player behaviors to validate the experimental findings.

REFERENCES

[1] C. L. Baker, J. Jara-Ettinger, R. Saxe, and J. B. Tenenbaum, "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing," *Nature Human Behaviour*, vol. 1, no. 4, p. 0064, 2017.

[2] J. A. Fodor, "A theory of the child's theory of mind." *Cognition*, 1992.

[3] M. K. Ho, R. Saxe, and F. Cushman, "Planning with theory of mind," *Trends in Cognitive Sciences*, 2022.

[4] M. Romeo, P. E. McKenna, D. A. Robb, G. Rajendran, B. Nesset, A. Cangelosi, and H. Hastie, "Exploring theory of mind for human-robot collaboration," in *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2022, pp. 461–468.

[5] S. Vinanzi, M. Patacchiola, A. Chella, and A. Cangelosi, "Would a robot trust you? developmental robotics model of trust and theory of mind," *Philosophical Transactions of the Royal Society B*, vol. 374, no. 1771, p. 20180032, 2019.

[6] J. Williams, S. M. Fiore, and F. Jentsch, "Supporting artificial social intelligence with theory of mind," *Frontiers in artificial intelligence*, vol. 5, p. 750763, 2022.

[7] D. C. Knill and A. Pouget, "The bayesian brain: the role of uncertainty in neural coding and computation," *TRENDS in Neurosciences*, vol. 27, no. 12, pp. 712–719, 2004.

[8] P. E. McKenna, M. Romeo, J. Pimentel, M. Diab, M. Moujahid, H. Hastie, and Y. Demiris, "Theory of mind and trust in human-robot navigation," in *Proceedings of the First International Symposium on Trustworthy Autonomous Systems*, 2023, pp. 1–5.

[9] C. L. Baker, R. Saxe, and J. B. Tenenbaum, "Action understanding as inverse planning," *Cognition*, vol. 113, no. 3, pp. 329–349, 2009.

[10] Q. Wang, K. Saha, E. Gregori, D. Joyner, and A. Goel, "Towards mutual theory of mind in human-ai interaction: How language reflects what students perceive about a virtual teaching assistant," in *Proceedings of the 2021 CHI conference on human factors in computing systems*, 2021, pp. 1–14.

[11] B. Chen, C. Vondrick, and H. Lipson, "Visual behavior modelling for robotic theory of mind," *Scientific Reports*, vol. 11, no. 1, pp. 1–14, 2021.

[12] N. Rabinowitz, F. Perbet, F. Song, C. Zhang, S. A. Eslami, and M. Botvinick, "Machine theory of mind," in *International conference on machine learning*. PMLR, 2018, pp. 4218–4227.

[13] C. Yu, H. Hastie, and A. Cangelosi, "Human-aware robot social behaviour dynamic for trustworthy human-robot interaction," in *The IEEE/RSJ International Conference on Intelligent Robots and Systems*, ser. Workshop on Social and Cognitive Interactions for Assistive Robotics, 2022.

[14] H. Zhu, C. Yu, and A. Cangelosi, "Explainable emotion recognition for trustworthy human–robot interaction," in *Proc. Workshop Context-Awareness Human-Robot Interact. Approaches Challenges at ACM/IEEE HRI 2022, Sapporo, Japan*, 2022.

[15] B. Nesset, M. Romeo, G. Rajendran, and H. Hastie, "Robot broken promise? repair strategies for mitigating loss of trust for repeated failures," in *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2023, pp. 1389–1395.

[16] M. Lewis, K. Sycara, and P. Walker, "The role of trust in human-robot interaction," *Foundations of trusted autonomy*, pp. 135–159, 2018.

[17] Y. Guo, X. J. Yang, and C. Shi, "Reward shaping for building trustworthy robots in sequential human-robot interaction," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 7999–8005.

[18] Y. Guo, J. Yang, and C. Shi, "Tip a trust inference and propagation model in multi-human multi-robot teams," in *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. ACM/IEEE, 2023, p. 639–643.

[19] M. Chen, S. Nikolaidis, H. Soh, D. Hsu, and S. Srinivasa, "Planning with trust for human-robot collaboration," in *Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction*, 2018, pp. 307–315.

[20] Y. Guo, C. Shi, and X. J. Yang, "Reverse psychology in trust-aware human-robot interaction," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4851–4858, 2021.

[21] E. J. De Visser, M. M. Peeters, M. F. Jung, S. Kohn, T. H. Shaw, R. Pak, and M. A. Neerincx, "Towards a theory of longitudinal trust calibration in human–robot teams," *International journal of social robotics*, vol. 12, no. 2, pp. 459–478, 2020.

[22] R. Tian, M. Tomizuka, and L. Sun, "Learning human rewards by inferring their latent intelligence levels in multi-agent games: A theory-of-mind approach with application to driving data," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 4560–4567.

[23] A. Sharkey and N. Sharkey, "We need to talk about deception in social robotics!" *Ethics and Information Technology*, vol. 23, pp. 309–316, 2021.

[24] D. Ullrich, A. Butz, and S. Diefenbach, "The development of overtrust: An empirical simulation and psychological analysis in the context of human–robot interaction," *Frontiers in Robotics and AI*, vol. 8, p. 554578, 2021.

[25] C. S. Nam and J. B. Lyons, *Trust in Human-Robot Interaction*. Academic Press, 2020.

[26] A. Kumar, A. Zhou, G. Tucker, and S. Levine, "Conservative q-learning for offline reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1179–1191, 2020.

[27] H. Soh, Y. Xie, M. Chen, and D. Hsu, "Multi-task trust transfer for human–robot interaction," *The International Journal of Robotics Research*, vol. 39, no. 2-3, pp. 233–249, 2020.

[28] Y. Guo and X. J. Yang, "Modeling and predicting trust dynamics in human–robot teaming: A bayesian inference approach," *International Journal of Social Robotics*, vol. 13, no. 8, pp. 1899–1909, 2021.

[29] C. Yu, B. Serhan, M. Romeo, and A. Cangelosi, "Robot theory of mind with reverse psychology," in *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, 2023, pp. 545–547.

[30] T. Seno and M. Imai, "d3rlpy: An offline deep reinforcement learning library," *Journal of Machine Learning Research*, vol. 23, no. 315, pp. 1–20, 2022. [Online]. Available: http://jmlr.org/papers/v23/22-0017.html