# Telling Friend from Foe - Towards a Bayesian Approach to Sincerity and Deception

Sarit Adhikari
University of Illinois at Chicago
Chicago, IL 60607
sadhik6@uic.edu

Piotr J. Gmytrasiewicz
University of Illinois at Chicago
Chicago, IL 60607
piotr@uic.edu

## Abstract

In Communicative Interactive Partially Observable Markov Decision Processes (CIPOMDPs), agents use Bayes update to process their observations and messages without the usual assumption of cooperative discourse. We formalize the notion of sincerity and deception in terms of message in message space and belief of the agent sending the message. We then use a variant of the point-based value iteration method called IPBVI-Comm to study the optimal interactive behavior of agents in cooperative and competitive variants of the Tiger game. Modeling the belief and preferences of the opponent allows an agent to predict their optimal communicative behavior and physical action. Unsurprisingly, it may be optimal for agents to attempt to mislead others if their preferences are not aligned. But it turns out the higher depth of reasoning allows an agent to detect insincere communication and to guard against it. Specifically, in some scenarios, the agent can distinguish a truthful friend from a deceptive foe when the message received contradicts the agent's observations, even when the received message does not directly reveal the opponent type.

## 1 Introduction

The study of communication and interaction among self-interested agents in a partially observable and stochastic domain has application in several fields ranging from military [BAG$^+$05, GYPC20] to social robotics [BTK08]. The communicative behavior among the agents in multi-agent systems has been studied in cognitive science [AdR18], economics [EG19] and artificial intelligence [GL20, YFS$^+$20]. With the advancement of artificial intelligence, the topic of machine deception has become more important. In particular, since communication among agents is becoming ubiquitous, malicious agents trying to exploit the vulnerabilities in other AI systems and humans might be a common problem of the future. Thus it is important to lay the foundation for deception-resistant AI systems. Further, as more AI agents are becoming part of our social life, the study of emergent social behavior among communicating agents (both artificial and human) with varied preferences is vital.

Although vast literature exists on the topic of machine fooling humans through fake content [SSW$^+$17] and human fooling machines with adversarial attacks [KGB17], the study of deception in a sequential decision-making scenario, by modeling other agents have rarely been explored. As argued in [IB17], AI needs to guard itself against malevolent humans and sometimes be able to deceive as well. On the other hand, when the agents' preferences



In: R. Falcone, J. Zhang, and D. Wang (eds.): Proceedings of the 22nd International Workshop on Trust in Agent Societies, London, UK on May 3-7, 2021, published at http://ceur-ws.org

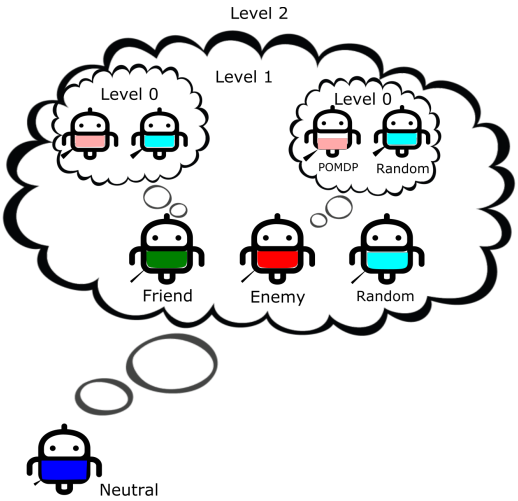



Figure 1: Theory of Mind (ToM) reasoning from the perspective of agent $i$ interacting with agent $j$ when there is uncertainty about the opponent type. Level is indicative of cognitive sophistication. The neutral agent $i$ thinks $j$ might be enemy, friend or random agent. Further, $i$ thinks $j$ thinks $i$ might be a sincere and gullible agent or a random agent. The behavior of $j$ is simulated by $i$ by putting itself in $j$'s shoes. Within that simulation, $i$ needs to reason how $j$ simulates $i$'s behavior.

align, then they benefit from sincere communication. Like physical actions, communicative actions are guided by the expected utility obtained in a particular state. Agents sometimes benefit from being sincere and sometimes it is in their best interest to deceive. To be able to cooperate, deceive or guard against deception, the agent needs to model the belief, intention, and preference of the other agent [BI11].

Humans are not very adept in detecting lies; we find it difficult to tell a skillful liar from a truthful, but possibly misinformed, friend. The main reason is that both may behave the same. We show examples of interactions during which agents, using Bayes update, can form informative beliefs about the identity of their opponents by keeping careful track of their possible states of knowledge, and by comparing the content of their communications to the agent's own observations. It is important to note that analyzing communicative behavior as separate from physical actions is bound to be insufficient, unless agents care directly about states of other agents' beliefs, as opposed to caring only about physical states of the world. In that latter, arguably more realistic, case, rational agents will attempt to modify others' beliefs only because this could induce them into advantageous physical actions - this could not be analyzed if one abstracts physical actions away.

Communicative Interactive Partially Observable Markov Decision Processes (CIPOMDPs) provide a principled framework for rational interaction and communication in a multi-agent environments [Gmy20]. CIPOMDP framework is an extension of interactive POMDP [GD05] to include the exchange of messages among the agents. IPOMDP, in turn, extends Partially Observable Markov Decision Process (POMDP) [Son78] to include other agents by incorporating their models into its state space. Thus, while POMDPs provide a theoretically sound framework to model uncertainty about the state of the world, the Theory of Mind (ToM) approach of IPOMDPs and CIPOMDPs allows modeling of uncertainty about other agents, including their beliefs about the world and other agents. Figure 1 shows the theory of mind (ToM) of the decision-making agent which is uncertain about the type of another agent and how it may model others including the original agent. At the bottom of the hierarchy of models could be a random agent, or a rational agent that does not model others, i.e., a classical POMDP agent. A great deal of research in psychology establishes a connection of deception to the recursive theory of mind reasoning, which starts at an early age in humans [STHP91, RO98, DWW$^+$15]. More recently, [OSV19] provides a comprehensive quantitative analysis of the role of rationality and theory of mind in deception and detecting deception.

We first formalize the definition of sincerity and deception in terms of the message sent by the agent and its belief about the state of the world. We then propose an offline solution method for communicative IPOMDPs, which adopts a point-based solution technique to solve for the optimal communicative behavior. The algorithm computes the policy to find the optimal action-message pair at each time-step of interaction with the other agent. The executed action may not be known to other agents but we assume that the message sent is received by the other agent with certainty. The subsequent policies are conditional on not only observation but also message

received from the other agent. Based on the preference of the agent and what is known about the preferences of the modeled agent, the agent may benefit from incorporating the message from another agent into its belief, but discounting it if it thinks the other agent has an incentive to lie. Since CIPOMDP agent needs message space to calculate the policy offline and messages express agent's beliefs, we construct the message space by discretizing the interactive belief space. The policy of POMDP agent on the bottom of a ToM hierarchy is augmented with a sincere messages it can send using a look-ahead reachability tree from the initial belief in the interactive state of the modeling agent. Similarly, we propose a way for an agent modeled as a POMDP to receive messages by augmenting its observation space. We apply CIPOMDPs to agents interacting in the multi-agent tiger game and we show that communication is valuable to agents and results in superior policies compared to its no communication counterpart. In cooperative scenarios, the agent can take advantage of messages from a sincere agent as additional observations, and can send sincere messages that inform the other agent. In competitive scenarios, the agent not only attempts to deceive the other agent but also ignores the message it knows to be deceitful. We then show how Bayesian update allows an agent higher in a cognitive hierarchy to tell a friend from foe based on the message received and its own observation.

## 2 Related work

The problem of agents communicating and interacting simultaneously has been addressed in several decision-theoretic as well as RL settings. [FAdFW16] uses DDRQN to learn communication protocol and solve a riddle by coordination. [FSH+19] combines multiagent RL with a bayesian update to compute communication protocols and policies in cooperative, partially observable multi-agent settings. [SsF16] uses a neural model that learns communication along with the policy. In the planning and control setting, [NMT13] uses communication for controllers to share part of their observation and control history at each step. More recently, [ULS20] used POMDP with communication for human-robot collaboration task. Other work in HRI include [CT16], [DCA17] and [WPH16] . [DA16] uses a theory of mind approach for the execution of a shared plan.

Communication has been studied extensively in other multi-agent decision theoretic frameworks [NRYT03], [NPY+04], [ZVlY04], [OSV07] In [RMG15], agents use extended belief state that contain approximation of other agents' beliefs. But these works assume fully cooperative interactions and mostly involve central planning. CIPOMDPs, on the other hand, provide subjective theory of mind reasoning during communication, and follows Bayesian approaches to pragmatics.

Deception has been widely studied across multiple disciplines including game theory [STX+18], [EJ10], psychology [GA14] and economics [Gne05]. When it comes to a sequential decision process to study attacker's and defender's approaches in cybersecurity research , decision theoretic framework of POMDP [AASN+20] and IPOMDP [SDS20] has been used. [SPB+19] combines ToM with components from deception theory and implements an epistemic agent using Agent-Oriented Programming Language.

## 3 Background

### 3.1 Communicative Interactive POMDPs

CIPOMDP [Gmy20] is the first general framework for an autonomous self-interested agent to communicate and interact with other agents in the environment based on Bayesian decision theory. A finitely nested communicative interactive POMDP of agent $i$ in an environment with agent $j$, is defined as:

$$CIPOMDP_i = \langle IS_{i,l}, A_i, \mathbb{M}, \Omega_i, T_i, O_i, R_i \rangle \tag{1}$$

where $IS_{i,l}$ is a set of interactive states, defined as $IS_{i,l} = S \times M_{j,k}, l \geq 1$, where $S$ is the set of physical states and $M_{j,k}$ is the set of possible models of agent $j$, $l$ is the strategy (nesting) level, and $k < l$. The set of possible models $M_{j,k}$ consists of intentional models, $\Theta_j$, or sub-intentional ones, $SM_j$. While the intentional models ascribe beliefs, preferences and rationality in action selection to the modeled agent, the sub-intentional models do not. We consider $k$th (less than $l$) level intentional models of agent $j$ defined as $\theta_{j,k} = \langle b_{j,k}, A_j, \Omega_j, T_j, O_j, R_j \rangle$, where $b_{j,k}$ is agent $j$'s belief nested to the level $k$, $b_{j,k} \in \Delta(IS_{j,k})$. The intentional model $\theta_{j,k}$, is sometimes called *type*, can be rewritten as $\theta_{j,k} = \langle b_{j,k}, \hat{\theta}_j \rangle$, where $\hat{\theta}_j$ includes all elements of the intentional model other than the belief and is called the agent $j$'s frame. Among the classes of sub-intentional models, we consider no-information model [GD00] which randomly selects action to execute and message to send in each time-step. The random models are possible in each level starting with level-0.

In contrast to classical POMDPs and similar to IPOMDPs, the transition, observation and reward functions in CIPOMDPs take actions of other agents into account. $A = A_i \times A_j$ is the set of joint actions of all agents, $\Omega_i$ is the set of agent i's possible observations, $T_i : S \times A \times S \to [0, 1]$ is the state transition function, $O_i : S \times A \times \Omega_i \to [0, 1]$ is the observation function, $R_i : S \times A \to R$ is the reward function.

The $IS_{i,l}$ can be defined inductively

$$
\begin{aligned}
IS_{i,0} &= S, && \Theta_{j,0} = \{\langle b_{j,0}, \hat{\theta}_j \rangle : b_{j,0} \in \Delta(S)\} \\
& && M_{j,0} = \Theta_{j,0} \cup SM_j \\
IS_{i,1} &= S \times M_{j,0}, && \Theta_{j,1} = \{\langle b_{j,1}, \hat{\theta}_j \rangle : b_{j,1} \in \Delta(IS_{j,1})\} \\
& && M_{j,1} = \Theta_{j,1} \cup SM_j \\
& \ldots\ldots \\
IS_{i,l} &= S \times \overset{l-1}{\underset{k=0}{\times}} M_{j,k}, && \Theta_{j,l} = \{\langle b_{j,l}, \hat{\theta}_j \rangle : b_{j,l} \in \Delta(IS_{j,l})\} \\
& && M_{j,l} = \Theta_{j,l} \cup SM_j
\end{aligned}
$$

The above defines the 0-level model, $\theta_{j,0}$ as having beliefs only over the physical state space, $S$. The level 1 agent model maintains beliefs over the physical states and 0-level models of the opponent. A level $l$ agent, $\theta_{j,l}$, maintains beliefs over $S$ and over models of the opponent nested up to $l-1$.

$\mathbb{M}$ is a set of messages the agents can send to and receive from each other , i.e., it is a communication language the agents share. Since agents' beliefs are probability distributions and communication is intended to share beliefs, it is natural to interpret a message in $\mathbb{M}$ as a marginal probability distribution over a subset of variables in the agents' interactive state spaces $IS_i$ and $IS_j$, which overlap. That way $\mathbb{M}$ is a set of probabilistic statements about the interactive state space. The message *nil*, i.e., silence, contains no variables. Note that we do not assume that messages reflect agents' actual beliefs. We will further discretize $\mathbb{M}$ below.

### 3.2 Belief update in CIPOMDPs

Belief update in CIPOMDPs is analogous to belief update in IPOMDPs when it comes to actions and observations. At any particular time step agents $i$ and $j$ can not only perform physical actions and observe but also send and receive messages. Call the message $i$ sent at time $t-1$ $m_{i,s}^{t-1}$, and one $i$ received at time $t$ $m_{i,r}^{t}$, and analogously for $j$. We assume all messages are in $\mathbb{M}$ and that message transmission is perfect. We provide precise definition of message space in our formulation in section 4.1. The belief update in CIPOMDPs has to update the probability of interactive state given the previous belief, action and observation, and given the message sent (at the previous time step) and received (at the current time): $P(is^t | b_i^{t-1}, a_i^{t-1}, m_{i,s}^{t-1}, o_i^t, m_{i,r}^t)$:

$$
\begin{aligned}
b_i^t(is^t) = P(is^t | b_i^{t-1}, a_i^{t-1}, m_{i,s}^{t-1}, o_i^t, m_{i,r}^t) = \eta \sum_{is^{t-1}} b_i^{t-1}(is^{t-1}) \sum_{a_j^{t-1}} P(m_{j,s}^{t-1}, a_j^{t-1} | \theta_j^{t-1}) \\
\times O_i(s^t, a^{t-1}, o_i^t) T_i(s^{t-1}, a^{t-1}, s^t) \sum_{o_j^t} \tau_{\theta_j^t}(b_j^{t-1}, a_j^{t-1}, m_{j,s}^{t-1}, o_j^t, m_{j,r}^t, b_j^t) O_j(s^t, a^{t-1}, o_j^t) \qquad (2)
\end{aligned}
$$

The term $P(m_{j,s}^{t-1}, a_j^{t-1} | \theta_j^{t-1})$ quantifies the relation between the message $i$ received from $j$ and the model, $\theta_j$, of agent $j$ that generated the message.[1] This term is the measure of $j's$ sincerity, i.e., whether the message $j$ sent reflects $j$'s beliefs which are part of the model $\theta_j$. $\eta$ is the normalizing constant.

### 3.3 Planning in CIPOMDPs

The utility of interactive belief of agent $i$, contained in $i$'s type $\theta_i$, is:

$$
\begin{aligned}
U_i(\theta_i) = \max_{(m_{i,s}, a_i)} \Big\{ \sum_{is \in IS} b_{is}(s) ER_i(is, m_{i,s}, a_i) \qquad (3) \\
+ \gamma \sum_{(m_{i,r}, o_i)} P(m_{i,r}, o_i | b_i, a_i) U_i(\langle SE_{\theta_i}(b_i, a_i, m_{i,s}, o_i, m_{i,r}), \hat{\theta}_i \rangle) \Big\}
\end{aligned}
$$

[1] Note that $m_{j,s}^{t-1} = m_{i,r}^t$ because message transmission is assumed to be perfect.

$ER_i(is, m_{i,s}, a_i)$ above is the immediate reward to $i$ for sending $m_{i,s}$ and executing action $a_i$ given the interactive state $is$ and is equal to $\Sigma_{a_j} R_i(is, a_i, a_j, m_{i,s}) P(a_j|\theta_j)$.

The planning in CIPOMDP makes use of equation 3 , which is based on the Bellman optimality principle. The policy computes optimal action, message pair which results in a maximum expected reward. Consequently, value iteration in CIPOMDP is analogous to that in IPOMDP and POMDP. The set of optimal message-action pairs, $(m^*_{i,s}, a^*_i)$ is obtained by replacing max in equation 3 with argmax. We call the resulting set of optimal message-action pairs $OPT(\theta_i)$. When agent $i$ models agent $j$ as a strict optimizer, $i$ predicts j would choose action-message pair in $OPT$ set with equal probability:

$$P(m_{j,s}, a_j|\theta_j) = \frac{1}{|OPT(\theta_j)|} \tag{4}$$

and that $P(m_{j,s}, a_j|\theta_j)$ is equal to zero if $(m_{j,s}, a_j)$ is not in $OPT$. The possibility that agents may be less than optimal is considered in [Gmy20]

Being able to compute the probabilities of messages given the preferences and beliefs of a speaker is of crucial importance when sincerity is not guaranteed. The belief update in CIPOMDPs provides the principled way to discount content of messages that may be insincere because it is in the interest of the speaker to transmit them. We give an example of this further below.

# 4 Approach

## 4.1 Message space

We define message space, $\mathbb{M}$, as a set of marginal probability distributions over a subset of variables in the interactive states of the agent. Since the belief space is continuous we make the computation more tractable by quantizing $\mathbb{M}$ into finite set of belief points. The message space is augmented with $nil$, which is analogous to no-op operation in the physical action set. Limiting message space to only $nil$ message reduces CIPOMDP to IPOMDP. Usually, a message contains information about only the subset of possible interactive states.[2] The variables that the message doesn't mention are interpreted as being marginalized. We use the principle of indifference and assume that probability is uniformly distributed among the remaining variables.

The message received $m_{i,r}$ can provide a mapping from physical state to belief marginalizing other variables of interactive state (belief of another agent, frame, etc). Then $m_{i,r}(s)$ denotes belief the message ascribes to physical state $s \in S$. Further, message may only contain probabilities of subset of values that the variables mentioned in message can take. Let $S'$ be the set of states not mentioned in the message. If a message $m$ does not mention state s' then

$$\forall_{s'} \in S'$$
$$m(s') = \frac{1 - \sum_{s \in S - S'} m(s)}{|S| - |S'|} \tag{5}$$

## 4.2 Sincerity and Deception

Let $X$ denote a set of variables describing the interactive state of the agent. $\wp(X)$ is a set of all non-empty subsets of $X$. The joint belief $b^t$ of an agent can be marginalized over the subset of variables in $X$. Let $b_{\overline{X}}$ represent belief marginalized over variables in $\overline{X}$. Accordingly message space $\mathbb{M}$ can be factored into the sets of messages marginalized over each of $\overline{X} \in \wp(X)$.

Sincere message can be defined as a message in message space $m \in \mathbb{M}_{\overline{X}}$ which is closest to the marginalized belief $b^t_{\overline{X}}$ consistent with the true joint belief $b^t$ of the agent at time $t$. The distance is defined in terms of the L1 norm. Thus the set of sincere messages is given by

$$\mathbb{M}_{sincere} = \bigcup_{\overline{X} \in \wp(X)} \arg\min_{m \in \mathbb{M}_{\overline{X}}} \|b^t_{\overline{X}} - m\| \tag{6}$$

Insincere(deceptive) message can be defined as any message in message space except the one closest to true belief $b_t$ of the agent at time $t$. Thus the set of insincere messages is given by

$$\mathbb{M}_{insincere} = \bigcup_{\overline{X} \in \wp(X)} \mathbb{M}_{\overline{X}} - \arg\min_{m \in \mathbb{M}_{\overline{X}}} \|b^t_{\overline{X}} - m\| \tag{7}$$

[2]For example, if the physical state space, $S$, is factored into variables $X$ and $Y$, a message, $m$, might be "$P(0 \leq X \leq 100) = 0.7$".

### 4.3 Communication for POMDP (level-0 CIPOMDP)

The recursion in CIPOMDP bottoms out as a flat POMDP which does not have a model of the other agent. We use the definition of literal speaker [3] from [Gmy20]. A literal listener can incorporate the incoming message as additional observation, which we describe in the following section. These assumptions enable an agent that does not model the other agent to participate in the exchange of messages.

#### 4.3.1 Augmented observation space and function for POMDP

We propose POMDP ($\theta_0$) can receive the message and include it in its belief update, by augmenting its observation space and consequently observation function. Observation space now becomes a Cartesian product of usual observation space and message space.

$$\Omega' = \Omega \times \mathbb{M} \tag{8}$$

The joint probability of observation and message received is obtained by combining the likelihood function for a physical observation with message distribution. The message distribution is represented by a triangular distribution with the idea that the likelihood of a message reflecting belief about the world state should increase monotonically as the belief. For e.g. if the belief is defined over two states $(s1, s2)$, and the true state of the world is $s1$, the likelihood of received message reflecting belief of (1,0) should be higher than the likelihood of received message reflecting belief of (0.5, 0.5). This formulation makes the agent gullible.

$$P(m_{i,r}|s) = \begin{cases} \frac{1}{|S|} & if m_{i,r} = nil \\ (m_{i,r}(s)) & otherwise \end{cases} \tag{9}$$

Given the state of the world, observation is conditionally independent of the message received, then the augmented observation function can be defined as

$$\forall_{m_{i,r} \in \mathbb{M}} \text{ and } \forall_{o \in \Omega} \text{ and } \forall_{s \in S}$$

$$\begin{aligned} O'(s, a, o, m_{i,r}) &= \frac{P(o, m_{i,r}|s, a)}{\sum_{o, m_{i,r}} P(o, m_{i,r}|s, a)} \\ &= \frac{P(o|s, a)P(m_{i,r}|s, a)}{\sum_o P(o|s, a) \sum_{m_{i,r}} P(m_{i,r}|s)} \\ &= \frac{O(s, a, o)P(m_{i,r}|s)}{\sum_{m_{i,r}} P(m_{i,r}|s)} \end{aligned} \tag{10}$$

### 4.4 Algorithm

#### 4.4.1 Value Iteration

As in POMDPs, the value function of CIPOMDPs is represented in terms of max over linear segments called alpha-vectors. Each alpha vector corresponds to a policy and each component of alpha vector ascribes value to an interactive state. Value iteration proceeds by backing up alpha-vectors to a higher time horizon starting from horizon 1.

#### 4.4.2 Complexity

POMDPs suffer from the curse of dimensionality and the curse of history. Naturally, These curses are carried over to IPOMDPs and CIPOMDPs, which require solving of nested POMDPs and CIPOMDPs. The curse of history is more prominent in CIPOMDPs because the policy is now conditional on both observation and message received. Since, computing optimal policies for POMDPs by exact solution methods are proven to be PSPACE-complete for finite time horizon and undecidable for an infinite time horizon [MHC03], a large amount of work has been done in computing approximate solution. [PGT03] introduced a point-based value iteration

[3] literal speaker generates a message reflecting its true belief about the physical states of the world $b_t$ with probability $1 - \alpha$ and all other messages including 'nil' with probability $\frac{\alpha}{|\mathbb{M}|-1}$. The messages are then broadcasted to "no one in particular" (NOIP), and do not take part in belief update for POMDP agent

(PBVI) algorithm to approximate exact value iteration by selecting a fixed set of representative belief points and maintaining alpha vectors that are optimal at those points only. Our work builds on the interactive point-based value iteration [DP08] which showed improvement in runtime over other IPOMDP solution methods like [DG05]. Exact value iteration quickly becomes intractable in PODMPs and IPOMDPs due to generation of large number of alpha-vectors which is exponential in observation space $|A||\nu^{t+1}|^{|\Omega|}$, where $\nu^{t+1}$ denote set of alpha vectors being backed-up from $t+1$ to $t$. In the case of CIPOMDP, the size is further exploded due to the inclusion of message space in the policy. The exact number of alpha-vectors generated at time $t+1$ will be $|A||\nu^{t+1}|^{|\Omega||\mathbb{M}|}$ . To keep the size of the alpha-set in each iteration tractable, we can use the point-based method, which only retains the vectors which are optimal at the fixed set of belief points. As in IPOMDP, we need to solve the lower-level model to compute the alpha vectors. Accordingly, we limit the initial model of other agents to a finite set.

#### 4.4.3 IPBVI-Comm

The point-based value iteration approach backs up the alpha-vectors optimal at a fixed set of belief points in each iteration starting from horizon 1. Each iteration consists of three steps, which we describe below, and along the way we highlight the difference with IPBVI [DP08]. The belief set for each horizon consists of randomly generated belief points across all interactive states.

Step 1

The first step involves calculating the intermediate set of alpha vectors $\Gamma^{a_i,m_{i,s},*}$ representing immediate reward for action $a_i$ and message $m_{i,s}$ (equation 11), and $\Gamma^{a_i,o_i,m_{i,s},m_{i,r}}$ representing future reward after receiving observation $o_i$ and message $m_{i,r}$ (equation 12). The step is performed for all actions, observations and messages. For horizon 1, computation of immediate reward is sufficient and will be used as initial alpha set for subsequent backups.

Different from point-based algorithm for IPOMDPs, we need to calculate $Pr(m_{i,r}|\theta_{j,l-1})$ and perform belief update for the other agent $j$ which now depends on message sent by $i$. Due to one-step delay in message exchange, the message sent in the current time step allows computing interactive states in next time step and backup the values from next time step. Assuming sincerity for $\theta_{j,0}$, only the message closest to belief will get probability 1- $\alpha$. All the other messages, including 'nil', will share the probability $\alpha$. When a message received is other than a sincere message, level-1 CIPOMDP ignores the message, and belief update proceeds as IPOMDP. For higher-level CIPOMDPs, the probability of message received is uniformly distributed among all the messages in $OPT(\theta_j)$ set, as defined in section 3.3.

$$\forall a_i \in A_i, \forall o_i \in \Omega_i, \forall m_{i,s} \in \mathbb{M}, \forall is \in IS$$

$$\Gamma^{a_i,m_{i,s},*} \leftarrow \alpha^{a_i,m_{i,s},*}(is) = \sum_{a_j \in A_j} R_i(s,a_i,a_j,m_{i,s}) Pr(a_j|\theta_{j,l-1}) \tag{11}$$

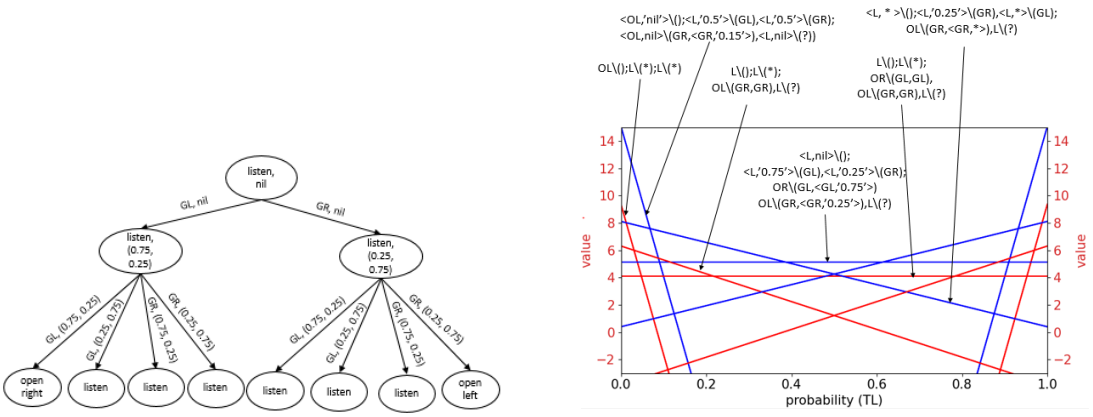
$$\Gamma^{a_i,o_i,m_{i,s},m_{i,r}} \leftarrow \alpha^{a_i,o_i,m_{i,s},m_{i,r}}(is) = \gamma \sum_{is' \in IS'} \sum_{a_j \in A_j} Pr(m_{i,r},a_j|\theta_{j,l-1}) T_i(s,a_i,a_j,s')$$
$$O_i(s',a_i,a_j,o_i) \sum_{o_j} O_j(s',a_i,a_j,o_j) \delta_D(SE_{\hat{\theta_j}}(b_{j,l-1},a_j,o_j,m_{j,s},m_{i,s}) - b'_{j,l-1}) \alpha^{'}(is') \tag{12}$$

Here, $\delta_D$ is the Dirac delta function taking the current belief and updated belief as an argument. The updated belief is returned by state estimator function $SE_{\hat{\theta_j}}$.

Step 2

The second step involves combining intermediate alpha vectors calculated in step 1 weighted by the observation and message likelihood using a cross sum operation. Due to the point-based approach, the cross sum operation in exact value iteration is simplified. The step proceeds by selecting only those intermediate alpha vectors which are optimal at any of the given set of belief points.

$$\Gamma^{a_i,m_{i,s}} \leftarrow \Gamma^{a_i,m_{i,s},*} \bigoplus_{o_i \in \Omega_i, m_{i,r} \in \mathbb{M}} \arg\max_{\Gamma^{a_i,o_i,m_{i,s},m_{i,r}}} (\alpha^{a_i,o_i,m_{i,s},m_{i,r}}.b_{i,l})$$
$$\forall b_{i,l} \in B_{i,l} \tag{13}$$




(a) (b)

Figure 2: (a)Policy tree for CIPOMDP agent with friend reward function modeling sinciere and gullible agent (b)Alpha vectors for friend reward function and sensor accuracy 0.85, for IPOMDP (without communication - in red) and for CIPOMDP (with communication - in blue). Each vector is labelled with a 3-step policy. For example $L\backslash(); L(*); OL\backslash(GR; GR), L(?)$ means that agent starts by listening (list of observations is empty), followed by listening no matter what it hears, followed in the third step by OL if previous two observations were GR and GR, but followed by another L if previous two observations were anything but (GR,GR). Communicative policies additionally include contents of messages received simultaneously with growls.

Step 3

In the final step, the belief points in set $B_{i,l}$ are used again to select the alpha vectors for the final set for the current iteration. Since different action, message pairs can be optimal for the modeled agent, we need to include alpha vectors corresponding to all optimal action, message pairs in the final alpha set.

$$\nu^t \leftarrow \underset{\alpha^t \in \cup_{a_j} \Gamma^{a_i,m_i}}{\arg\max} (\alpha^t.b_{i,l})$$
$$\forall b_{i,l} \in B_{i,l} \tag{14}$$

The recursion bottoms out as POMDP at level-0, which we assume to be a literal speaker. Since POMDP policy only computes physical action, we need to augment the policy with a sincere message.

## 5 Experiments and results

We report the results on the multi-agent tiger game, which elegantly models complex interaction with the environment and other agents in sequential decision-making scenarios under uncertainty. The game provides intuitive scenarios for the study of sincerity and deception because the agents can be incentivized to influence another agent to open right or wrong door while not ignoring uncertainty in the environment.

### 5.1 Multi-agent tiger game

In this version, two agents are facing two doors: "left" and "right". Behind one door lies a hungry tiger and behind the other is a pot of gold but the agents do not know the position of either. Thus, the set of states is $S = \{TL, TR\}$ indicating the tiger's presence behind the left, or right, door. Each agent can open either door. Agents can also independently listen for the presence of the tiger, so the actions are $A = \{OR, OL, L\}$ for opening the right door, opening the left door, and listening and is the same for both agents. The transition function $T$, specifies that every time either agent opens one of the doors, the state is reset to $TR$ or $TL$ with equal probability, regardless of the action of the other agent. However, if both agents listen, the state remains unchanged. After every action, each agent can hear the tiger's growl coming either from the left, $GL$, or from the right door, $GR$. The observation function $O$ (identical for both agents) specifies the accuracy of observations. We assume that tiger's growls are informative, with predefined sensor accuracy, only if the agents listen. If the agent opens the doors the growls have an equal chance to come from the left or right door and are thus completely uninformative.

Table 1: Neutral Reward

| $\langle a_i, a_j \rangle$ | TL | TR |
|---|---|---|
| OR, L | 10 | -100 |
| OL, L | -100 | 10 |
| L, L | -1 | -1 |
| OR, OL | 10 | -100 |
| OL, OL | -100 | 10 |
| L, OL | -1 | -1 |
| OR, OR | 10 | -100 |
| OL, OR | -100 | 10 |
| L, OR | -1 | -1 |

Table 2: Friend Reward

| $\langle a_i, a_j \rangle$ | TL | TR |
|---|---|---|
| OR, L | 9.5 | -100.5 |
| OL, L | -100.5 | 9.5 |
| L, L | -1.5 | -1.5 |
| OR, OL | -40 | -95 |
| OL, OL | -150 | 15 |
| L, OL | -51 | 4 |
| OR, OR | 15 | -150 |
| OL, OR | -95 | -40 |
| L, OR | 4 | -51 |

Table 3: Enemy Reward A

| $\langle a_i, a_j \rangle$ | TL | TR |
|---|---|---|
| OR, L | 10.5 | -99.5 |
| OL, L | -99.5 | 10.5 |
| L, L | -0.5 | -0.5 |
| OR, OL | 60 | -105 |
| OL, OL | -50 | 5 |
| L, OL | 49 | -6 |
| OR, OR | 5 | -50 |
| OL, OR | -105 | 60 |
| L, OR | -6 | 49 |

Table 4: Enemy Reward B

| $\langle a_i, a_j \rangle$ | TL | TR |
|---|---|---|
| OR, L | 10 | -100 |
| OL, L | -100 | 10 |
| L, L | -1 | -1 |
| OR, OL | 10 | -150 |
| OL, OL | -100 | -40 |
| L, OL | -1 | -51 |
| OR, OR | -40 | -100 |
| OL, OR | -150 | 10 |
| L, OR | -51 | -1 |

Table 5: Reward comparison for CIPOMDP agent against IPOMDP agent in different scenarios. For Enemy, reward function in Table 3 is used.

| Nesting Level | Agent | Opponent | Reward | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | h=3 | | h=4 | | h=5 | |
| | | | CIPOMDP | IPOMDP | CIPOMDP | IPOMDP | CIPOMDP | IPOMDP |
| 1 | Neutral | Sincere and Gullible | 3.4 ± 8.92 | 2.84 ± 16.02 | 3.9 ± 8.29 | 2.39 ± 7.95 | 3.5 ± 8.037 | 1.067 ± 20.275 |
| | Enemy | Sincere and Gullible | 46 ± 24.69 | 1.53 ± 18.07 | 66.48 ± 42.15 | 1.07 ± 8.887 | 86.00 ± 36.57 | -1.31 ± 24.79 |
| | Friend | Sincere and Gullible | 5.08 ± 10.91 | 4.15 ± 18.02 | 6.05 ± 9.42 | 3.56 ± 8.86 | 5.71 ± 17.10 | -0.81 ± 23.32 |
| 2 | Neutral | Enemy | 3.39 ± 8.08 | 2.81 ± 16.02 | 3.9 ± 11.40 | 2.39 ± 7.67 | 3.4942 ± 8.94 | 1.55 ± 17.14 |
| | Friend | Friend | 5.08 ± 10.5 | 4.14 ± 18.02 | 6.21 ± 8.56 | 3.67 ± 8.97 | 5.02 ± 10.099 | 3.65 ± 17.94 |
| | Enemy | Enemy | 5.44 ± 10.83 | 1.53 ± 18.07 | 8.99 ± 18.88 | 2.32 ± 15.72 | 10.78 ± 18.09 | 0.5 ± 19.81 |
| | Neutral | Uncertain (Enemy or Friend) | 3.43 ± 7.80 | 1.53 ± 18.07 | 4.19 ± 9.162 | 2.44 ± 7.87 | 3.45 ± 8.33 | 0.82 ± 13.71 |

#### 5.1.1 Reward functions

The reward functions are chosen to simulate cooperative and competitive scenarios. Table 1 represents the scenario when the reward of the agent is independent of the action of the other agent. Table 2 is the friend reward function where the agent gets half of the reward obtained by the other agent, in addition to its own reward. In table 3 the agent gets half of the negative of the reward obtained by the other agent, hence represents the competitive case. In other reward function Table 4, the agent gets -50 reward if another agent opens the correct door but there is no extra reward if the other agent opens the wrong door. Table 3 incentivizes the extreme lie while 4 incentivizes more believable lie.

#### 5.1.2 Experimental setup

Table 5 shows the total reward collected in multi-agent tiger game averaged across 10000 episodes for sensor accuracy of 0.85. The message space $\mathbb{M}$ is limited to distribution over physical states only and has been quantized into 5 equally spaced belief points (0.0,1.0),(0.25, 0.75),(0.5,0.5),(0.75, 0.25), (1.0, 0.0), and nil. The value of $\alpha$ is fixed to 0.01. The results show that the CIPOMDP agent outperforms the IPOMDP agent in terms of the average reward collected due to the sophistication of message exchange. The difference is more prominent when the agent is able to deceive the other agent. The behavior of the agent across multiple scenarios is discussed below.

### 5.2 Cooperative Scenarios

When level 2 friend $i$ models a friend $j$, agent $i$ sends a sincere message to $j$ reflecting its belief and further includes a message from another agent in its belief update. For e.g. after starting from uniform belief, if the agent $i$ hears GL, it will send the sincere message $m_{i,s} = (0.75, 0.25)$, which assigns probability 0.75 to TL and 0.25 to TR. Figure 2 (b) shows the alpha vectors plot corresponding to communicative policies for CIPOMDP

Table 6: Different scenarios of belief update of level 2 CIPOMDP agent $i$ modeling enemy, friend, and a random agent (initial belief is uniform over all the models and location of the tiger). For all scenarios, agent $i$ sends a nil message and executes listen action in each time-step. The belief update is shown in figure 3

| Scenario | time-step | observation | message received | Remarks |
|---|---|---|---|---|
| 1 | 1 | GL | nil | The received message contradicts agent's consecutive observations of GLs. The other agent is more likely to be enemy than friend |
| | 2 | GL | (0.25, 0.75) | |
| 2 | 1 | GL | nil | The received message jives in with the agent's consecutive observations of GLs. The other is more likely to be friend than enemy |
| | 2 | GL | (0.75, 0.25) | |
| 3 | 1 | GR | nil | Agent gets contradictory growls, there is an equal chance the other agent is an enemy (and tiger on left) or a friend (and tiger on right) |
| | 2 | GL | (0.25, 0.75) | |
| 4 | 1 | GR | nil | Agent gets contradictory growls, there is an equal chance the other agent is a friend (and tiger on left) or an enemy (and tiger on right) |
| | 2 | GL | (0.75, 0.25) | |
| 5 | 1 | GL | nil | Agent gets the message that is in OPT set of neither enemy nor friend. Hence, is certain that the other agent is a random agent |
| | 2 | GL | nil | |

agent with friend reward function modeling CIPOMDP (also with a friend reward). The resulting policy tree when the agents are uncertain about the initial location of the tiger is shown in figure 2 (a).

### 5.3 Non-cooperative Scenarios

When a level 1 CIPOMDP agent $i$ models a gullible agent $j$, it can collect a large reward by sending a deceitful message. For e.g. after starting from uniform belief and getting $GL$, the agent sends a message $m_{i,s} = (0, 1)$ which indicates agent $i$ is certain tiger is on the right, opposite to its own observation. When level 2 enemy $i$ models enemy $j$, the sophisticated agent $i$ can prevent itself from being deceived by $j$ and further take advantage of the deceitful message as an extra observation. Also, since level-2 agent knows level-1 CIPOMDP models the other as a sincere agent, the former tends to deceive the latter by sending a deceitful message. When level 2 neutral agent models an enemy, it has no incentive to deceive but can prevent itself from being deceived by ignoring the message from the other agent. When the nesting level is increased to level 3, the agent tends to send a sincere message with the intention to deceive the level 2 agent which is expecting a deceitful message.When the agents are indifferent to each other's actions i.e. both have a neutral reward function, then communication doesn't provide any added value and the reward collected is the same as the IPOMDP agent.

### 5.4 Uncertainty about the opponent

#### 5.4.1 Message non-revealing of the agent's type

Let's consider a two-agent interaction in a multi-agent tiger game where both the agents $i$ and $j$ start with no information about the location of the tiger. The level-2 CIPOMDP agent $i$ is uncertain about the level-1 opponent $j$ 's type and thus assigns uniform probability over the other agent being friend, enemy or random [4]. The friend is characterized by the friend reward function ($R_f$) which incentivizes sincere communication while the enemy is characterized by the enemy reward function ($R_e$) incentivizing the deceptive behavior. Also, all other elements in the frame of the CIPOMDP agent at all levels are assumed to be the same.

Behavior of level-0 agent

The POMDP agent at level-0 which starts with a uniform distribution over the physical states of the world $TL$ and $TR$, would open the door if it received two consecutive growls indicating the same location of tiger and the message non-indicative of opposing growl. Further, the agent would open the door at any time-step, regardless

[4]It can be the case that level-2 CIPOMDP is uncertain about the strategic level of the opponent and might want to model POMDP as well, but to simplify the illustration, we stick to 3 types of models

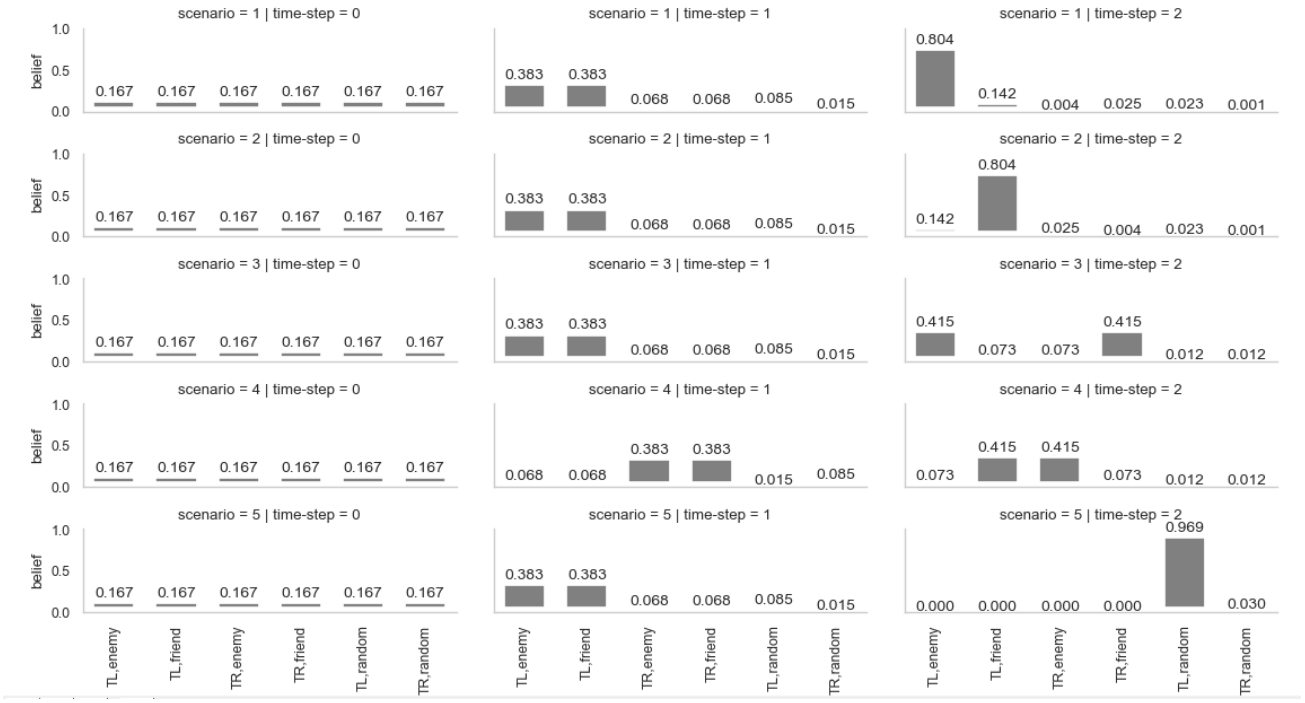



Figure 3: Belief Update for specific scenarios for level-2 CIPOMDP agent $i$ modeling enemy, friend, and random agent. Each row represents a scenario from table 6 and column represent time-step. In each scenario, the agent $i$ has a uniform initial belief over the state and the frame of another agent $j$. Each belief update is due to physical action, message sent, observation and message received which can be referenced from table 6

of its observation, if it got the extreme message indicating the location of a tiger with certainty. Also, being a literal speaker, the agent sends out its true belief at every time-step with probability $\alpha$.

Behavior of level-1 agent

The less sophisticated level-1 agent $\theta_{j,1}$ models sincere and gullible POMDP agent $\theta_{i,0}$, and the random agent $SM_i$. Let's consider the behavior of level-1 agent $j$ which has predicted the behavior of $\theta_{i,0}$ for every possible message it can send. The agent $\theta^e_{j,1}$ with enemy reward function (Table 4), has incentive to deceive the opponent, because every time $i$ opens the right door, $j$ gets -50 reward. Let's suppose the agent $\theta^e_{j,1}$ listens at time step $t = 0$ , to be certain of the tiger's location and then sends the message at time step $t = 1$ indicating growl from opposite location than what it heard. Let's suppose $\theta^e_{j,1}$ got $GL$, then it would send the message (0.25, 0.75). On other hand, if the level-1 agent had a friend reward function (Table 2), it would get a share of $i$'s reward. Let's call the agent $\theta^f_{j,1}$. $\theta^f_{j,1}$ would want the other agent $i$ to open the right door, and hence send the sincere message, indicative of its belief updated based on the observation. For an observation of $GR$, it would send a message (0.25, 0.75) at $t = 1$.

Belief update for level-2 agent

Level-2 CIPOMDP $i$ uses IPBVI-Comm to solve for the anticipated behavior of modeled agents $\theta^f_{j,1}$ and $\theta^e_{j,1}$. Now let's see how the belief update proceeds for $i$. Note, that optimal messages from level-1 agents are not indicative of the agent type. For e.g. if the enemy received GL, it would send the message (0.25, 0.75) and if the friend received GR, it would again send the message (0.25, 0.75). We study the scenario when the message itself is indicative of agent type in the next section. It turns out the agent $i$ is still able to assign a higher probability to one agent type than the other based on the observation it received. For e.g. if the observation and message received for $i$ in two time-steps are $< GL, 'nil' >$ and $< GL, (0.25, 0.75) >$, as shown in table 6, scenario 1, the agent is more likely to be an enemy than the friend because its observations contradict the received message. The belief update of other scenarios are shown in figure 3. The description of each scenario is provided in table 6.

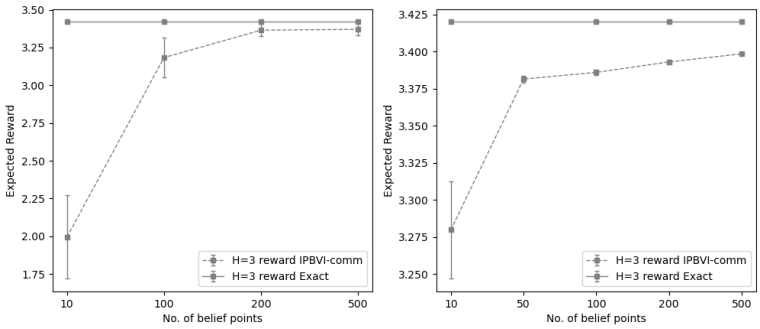



Figure 4: The performance profile of IPBVI-Comm for level 2 (left) and level 1 (right) CIPOMDP agent. The increase in expected reward is due to improvement in policy after increasing number of belief points. It also shows the comparison with exact value iteration.

#### 5.4.2 Message revealing the agent's type

Again we consider the scenario when there is uncertainty about the opponent type, i.e. when level 2 neutral agent $i$ models both enemy and friend. The only difference now is that the agent modeled at level-1 $\theta_{j,1}$ has the reward function that incentivizes the extreme lie (Table4). In this case, the higher level agent $\theta_{i,2}$ can figure out if the other agent $j$ is friend or enemy based on the message content only. The CIPOMDP agent $\theta_{i,2}$ incorporates the message from $j$ as a sincere message if it is closer to the belief of the modeled agent $\theta_{j,1}$ and discards the message as an insincere message if the incoming message is way off the belief of the modeled agent. For e.g. if the belief of the modeled agent is $(0.85, 0.15)$, $(0.75, 0.25)$ is considered a sincere message while $(1, 0)$ is considered insincere message. Let's suppose $\theta_{i,2}$ starts by listening at $t = 0$. After receiving $GL$ and message $nil$, the belief shifts towards physical state $TL$ but belief is equally distributed among both frames. At $t = 1$, after receiving message $(0.75, 0.25)$ and observation $GL$, the belief concentrates on tiger being on left and other agent being a friend. The agent is able to detect friend from enemy by calculation of sincerity term $P(m_{i,r}|\theta_j)$. When $\theta_j = (b_j, R_f)$, $P(m_{i,r}|\theta_j) = 1$ and when $\theta_j = (b_j, R_e)$, $P(m_{i,r}|\theta_j) = 0$. This happens because for level-1 CIPOMDP with enemy reward function, the optimal action would be to lie to the extreme. This would make $\theta_{i,0}$ open the door intended by the deceiver $\theta_{j,1}$, disregarding its own observation. This message reveals $j$ as an enemy to the higher level agent $i$.

### 5.5 Algorithm Performance

Since the point-based algorithm only backs up alpha-vectors optimal at fixed set of belief points, the performance would depend on the number of belief points chosen. Figure 4 shows the comparison of expected reward for the different number of belief points.

## 6 Conclusion

We started by devising a technique to incorporate the message into POMDP belief update in order to allow an agent that does not model other agents to take part in exchange of (literal) messages. We then formalized the notion of sincerity and deception in terms of belief of the agent and messages in message space. We adopted a point based solution method to CIPOMDPs to alleviate the complexities of considering communication as well as observations and physical actions. The analysis of computed policies shows the added sophistication of communication results in policies of superior quality, which is further supported by the empirical results on several experiments conducted on multi-agent tiger game. More importantly, we showed how higher levels of theory of mind may allow an agent to disambiguate a sincere friend from deceitful foe based on received message and observation from the environment.

In future work, we want to explore the higher depth of nesting, and more relaxed soft maximization criterion for action selection which can give rise to richer rational communicative behavior agents can engage in. We are considering online planning method like Monte Carlo tree search for computing policy which provides scalability and with some variation, could accomodate continuous message space.

## References


[AASN+20] Md Ali Reza Al Amin, Sachin Shetty, Laurent L. Njilla, Deepak K. Tosh, and Charles A. Kamhoua. *Dynamic Cyber Deception Using Partially Observable Monte-Carlo Planning Framework*, chapter 14, pages 331–355. John Wiley Sons, Ltd, 2020.

[AdR18] Saul Albert and J. P. de Ruiter. Repair: The interface between interaction and cognition. *Topics in Cognitive Science*, 10(2):279–313, 2018.

[BAG+05] P. Beautement, David H. Allsopp, M. Greaves, Steve Goldsmith, S. Spires, S. Thompson, and H. Janicke. Autonomous agents and multi -agent systems (aamas) for the military - issues and challenges. In *DAMAS*, 2005.

[BI11] Will Bridewell and Alistair Isaac. Recognizing deception: A model of dynamic belief attribution. In *AAAI Fall Symposium: Advances in Cognitive Systems*, 2011.

[BTK08] Cynthia Breazeal, Atsuo Takanishi, and Tetsunori Kobayashi. *Social Robots that Interact with People*, pages 1349–1369. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[CT16] Crystal Chao and Andrea Thomaz. Timed petri nets for fluent turn-taking over multimodal interaction resources in human-robot collaboration. *The International Journal of Robotics Research*, 35(11):1330–1353, 2016.

[DA16] S. Devin and R. Alami. An implemented theory of mind to improve human-robot shared plans execution. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 319–326, 2016.

[DCA17] Sandra Devin, Aurélie Clodic, and Rachid Alami. About decisions during human-robot shared plan achievement: Who should act and how? In *ICSR*, 2017.

[DG05] Prashant Doshi and Piotr Gmytrasiewicz. Approximating state estimation in multiagent settings using particle filters. In *Proceeding of AAMAS 2005*, 2005.

[DP08] Prashant Doshi and Dennis Perez. Generalized point based value iteration for interactive pomdps. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 1*, AAAI'08, page 63–68. AAAI Press, 2008.

[DWW+15] Xiao Pan Ding, Henry M. Wellman, Yu Wang, Genyue Fu, and Kang Lee. Theory-of-mind training causes honest young children to lie. *Psychological Science*, 26(11):1812–1821, 2015. PMID: 26431737.

[EG19] Piotr Evdokimov and Umberto Garfagnini. Communication and behavior in organizations: An experiment. *Quantitative Economics*, 10(2):775–801, 2019.

[EJ10] David Ettinger and Philippe Jehiel. A theory of deception. *American Economic Journal: Microeconomics*, 2(1):1–20, February 2010.

[FAdFW16] Jakob N. Foerster, Yannis M. Assael, Nando de Freitas, and Shimon Whiteson. Learning to communicate to solve riddles with deep distributed recurrent q-networks, 2016.

[FSH+19] Jakob N. Foerster, H. Francis Song, Edward Hughes, Neil Burch, Iain Dunning, Shimon Whiteson, Matthew M Botvinick, and Michael H. Bowling. Bayesian action decoder for deep multi-agent reinforcement learning. *ArXiv*, abs/1811.01458, 2019.

[GA14] Matthias Gamer and Wolfgang Ambach. Deception research today. *Frontiers in Psychology*, 5:256, 2014.

[GD00] Piotr J. Gmytrasiewicz and Edmund H. Durfee. Rational coordination in multi-agent environments. *Autonomous Agents and Multiagent Systems Journal*, 3(4):319–350, 2000.

[GD05] Piotr Gmytrasiewicz and Prashant Doshi. A framework for sequential planning in multiagent settings. *Journal of Artificial Intelligence Research*, 24:49–79, 2005. http://jair.org/contents/v24.html.

[GL20] Andrea L Guzman and Seth C Lewis. Artificial intelligence and communication: A human–machine communication research agenda. *New Media & Society*, 22(1):70–86, 2020.

[Gmy20] Piotr J. Gmytrasiewicz. How to do things with words: A bayesian approach. *J. Artif. Intell. Res.*, 68:753–776, 2020.

[Gne05] U. Gneezy. Deception: The role of consequences. *The American Economic Review*, 95:384–394, 2005.

[GYPC20] J. George, C. T. Yilmaz, A. Parayil, and A. Chakrabortty. A model-free approach to distributed transmit beamforming. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5170–5174, 2020.

[IB17] Alistair Isaac and Will Bridewell. White lies on silver tongues: Why robots need to deceive (and how). pages 157–172, 10 2017.

[KGB17] A. Kurakin, Ian J. Goodfellow, and S. Bengio. Adversarial examples in the physical world. *ArXiv*, abs/1607.02533, 2017.

[MHC03] Omid Madani, Steve Hanks, and Anne Condon. On the undecidability of probabilistic planning and related stochastic optimization problems. *Artificial Intelligence*, 147(1):5 – 34, 2003. Planning with Uncertainty and Incomplete Information.

[NMT13] A. Nayyar, A. Mahajan, and D. Teneketzis. Decentralized stochastic control with partial history sharing: A common information approach. *IEEE Transactions on Automatic Control*, 58(7):1644–1658, 2013.

[NPY+04] Ranjit Nair, David Pynadath, Makoto Yokoo, Milind Tambe, and Stacy Marsella. Communication for improving policy computation in distributed pomdps. In *Proceedings of the Agents and Autonomous Multiagent Systems (AAMAS)*, 2004.

[NRYT03] Ranjit Nair, Maayan Roth, Makoto Yokoo, and Milind Tambe. Taming decentralized pomdps: Towards efficient policy computation for multiagent settings. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, 2003.

[OSV07] Frans Olihoek, Mathijs Spaan, and Nikos Vlassis. Dec-pomdps with delayed communication. In *Proceedings of MSDM 2007 May 15, 2007, Honolulu, Hawai'i, USA*, 2007.

[OSV19] Lauren A Oey, Adena Schachner, and Edward Vul. Designing good deception: Recursive theory of mind in lying and lie detection, May 2019.

[PGT03] Joelle Pineau, Geoff Gordon, and Sebastian Thrun. Point-based value iteration: An anytime algorithm for pomdps. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, IJCAI'03, page 1025–1030, San Francisco, CA, USA, 2003. Morgan Kaufmann Publishers Inc.

[RMG15] J. Renoux, A. Mouaddib, and S. L. Gloannec. A decision-theoretic planning approach for multi-robot exploration and event search. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5287–5293, 2015.

[RO98] Nancy K. Ratner and Rose R. Olver. Reading a tale of deception, learning a theory of mind? *Early Childhood Research Quarterly*, 13(2):219 – 239, 1998.

[SDS20] Aditya Shinde, Prashant Doshi, and Omid Setayeshfar. Active deception using factored interactive pomdps to recognize cyber attacker's intent, 2020.

[Son78] Edward J. Sondik. The optimal control of partially observable markov processes over the infinite horizon: Discounted costs. *Oper. Res.*, 26(2):282–304, April 1978.

[SPB+19] S. Sarkadi, Alison R. Panisson, Rafael Heitor Bordini, P. McBurney, S. Parsons, and M. Chapman. Modelling deception using theory of mind in multi-agent systems. *AI Commun.*, 32:287–302, 2019.

[SsF16] Sainbayar Sukhbaatar, arthur szlam, and Rob Fergus. Learning multiagent communication with backpropagation. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2244–2252. Curran Associates, Inc., 2016.

[SSW+17] K. Shu, A. Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *SIGKDD Explor.*, 19:22–36, 2017.

[STHP91] Beate Sodian, Catherine Taylor, Paul L. Harris, and Josef Perner. Early deception and the child's theory of mind: False trails and genuine markers. *Child Development*, 62(3):468–483, 1991.

[STX+18] Aaron Schlenker, Omkar Thakoor, Haifeng Xu, Fei Fang, Milind Tambe, Long Tran-Thanh, Phebe Vayanos, and Yevgeniy Vorobeychik. Deceiving cyber adversaries: A game theoretic approach. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '18, page 892–900, Richland, SC, 2018. International Foundation for Autonomous Agents and Multiagent Systems.

[ULS20] Vaibhav V. Unhelkar, Shen Li, and Julie A. Shah. Decision-making for bidirectional communication in sequential human-robot collaborative tasks. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '20, page 329–341, New York, NY, USA, 2020. Association for Computing Machinery.

[WPH16] Ning Wang, David V. Pynadath, and Susan G. Hill. Trust calibration within a human-robot team: Comparing automatically generated explanations. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*, HRI '16, page 109–116. IEEE Press, 2016.

[YFS+20] Luyao Yuan, Zipeng Fu, Jingyue Shen, Lu Xu, Junhong Shen, and Song-Chun Zhu. Emergence of Pragmatics from Referential Game between Theory of Mind Agents. *arXiv e-prints*, page arXiv:2001.07752, Jan 2020.

[ZVlY04] Yu Zhang, Richard A. Volz, Thomas R. loerger, and John Yen. A decision-theoretic approach for designing proactive communication in multi-agent teamwork. In *Proceedings of the 2004 ACM Symposium on Applied Computing*, SAC '04, page 64–71, New York, NY, USA, 2004. Association for Computing Machinery.