Technical Report
How to Do Things With Words: Towards a Bayesian Approach

UIC CIPOMDPs Team

May 31, 2018

1 Abstract

Our premise is that communication is a type of action, and that decisions regarding communicative acts should be based on decision-theoretic planning using Bellman optimality principle already utilized in many AI systems. As in any form of planning, the results of actions need to be precisely specified. The main result contained in this paper is the principled derivation of how the sender and the receiver of message(s) should update their beliefs as the result of the communication using Bayes rule. We build on the interactive POMDP (IPOMDP) framework, which in turn extends POMDPs to multi-agent settings. IPOMDPs endow the agents with models of their environments and models of other agents. We supplement IPOMDPs with communicative acts available to the agents to formulate communicative IPOMDPs (CIPOMDPs).

2 Introduction

The idea of communicating with machines like we do with other people is a very appealing one for AI and important for its applications. The current state-of-the-art systems like Alexa and Siri seem mostly hard-coded to respond to queries, but lack the ability to initiate conversations and are not directed by any general notion of purpose of the interaction. Our work is predicated on the idea that interaction in general, and communication in particular, should be approached using already existing techniques allowing, say autonomous vehicles navigate a partially observable and non-deterministic environments. The reason

these techniques are applicable is because interactive settings are also non-deterministic and partially observable.

A key notion allowing planning is the specification of the consequences of an agent's action. In Markov decision processes (MDPs) and their partially observable variant (POMDPs) this specification is contained in the transition function T [17, 28]. This paper reports on the Bayesian approach to specifying the consequences of communicative acts that agents can engage in, since the main function of communication is to change the agents' beliefs [12, 21, 30]. On the one hand communication is action [2] executed by a speaker, but on the other hand it is perception for the hearer. Let's consider what happens when your friend Joe tells you that it's 40 degrees C outside in Melbourne this morning. Do you trust/believe him? What, precisely, happens to the state of the world, your state of knowledge about the state of the world, what you know about what Joe knows, etc.? How does your prior knowledge about the temperature (say that it's 15C plus minus 5C) factor into this? What if you think that Joe may be not well informed? And what if you suspect that Joe is a prankster, or maybe has been hacked by a foreign power? Finally, why should you even care about some of these?

Our ability to answer questions above is crucial to creating robust AI communicative systems and a vast amount of previous work is relevant. One can start with Austin's "How to do things with words" [2], approaches that make a cooperative assumption (Grice's Cooperative Principle [13]) that everything that has been said is true and will be believed, or pursue the game-theoretic and Bayesian approaches to pragmatics [12, 26, 31, 33] which analyze the interests of the sender of the information to interpret it. Communication among fully cooperative agents as in DEC-POMDPs was addressed in [25, 27, 32] but agents were assumed to communicate their most recent observations so the issue of pragmatics does not arise, and we consider a more general case of agents' being uncertain about the reward functions of other agents. As we show below, agents' having explicit models of others is necessary for Bayesian approach to pragmatics.

Our contribution is to propose a principled approach based on Bayesian decision theory and decision-theoretic planning. First, we build on interactive POMDPs [11] which allow agents to represent their state of knowledge about the physical states and about possible models of other agents. The ability of agents to model other agents has also been called the theory of mind. Its usefulness while interacting with others has long been established in psychology, linguistics, economics and AI [9, 10, 11, 13, 20, 29, 31, 26]. We limit our attention to intentional models, which also are IPOMDPs, and in which agents' beliefs represent what they know about the state of the world and about other agents, including their preferences and beliefs about other agents' beliefs, others' beliefs about others and so on. Such beliefs are called interactive beliefs [1, 11]. We augment IPOMDPs by allowing agents to send and receive messages, and call the resulting framework communicative

IPOMDPs (CIPOMDPs). In finitely nested CIPOMDPs the models agents have of each other terminate at a finite level, called strategy level l, with "flat" POMDP models, like in IPOMDPs. There is no a priori bound on the value of the strategy level; agents may be free to choose one as needed. For example in the "Cheryl's Birthday" problem[1] the audience needs to create models of Albert and Bernard, and their models of each other (and of Cheryl) to interpret the information shared about who does and does not know what and to figure out Cheryl's birthday. But, if only Cheryl were less opaque and simply informed us (the audience) of her birthday directly the nested models of Albert and Bernard would not be needed.

Second, we use Bayes update to describe the results of agents' communicative actions on agents' beliefs. This defines the analogue to the transition function for POMDPs, formalizes the game-theoretic pragmatics (see [8, 31, 26] and references therein) and implements Bayesian approach to language pragmatics [33] (but see [12] for an alternative approach). As we show, the update of agents' beliefs due to communication recursively descends l levels down to "flat" POMDP models where the communicative acts are processed according to their literal meaning. The Bayesian update combines the hearer's prior information with what the hearer knows about the speaker's sincerity (i.e., whether the content of the message reflects the speaker's beliefs), and with what the hearer knows about how informed the speaker is (i.e., whether the speaker's beliefs reflect reality.) Using the example of the tiger game we show that if the hearer knows the speaker is sincere and knows the speakers' observation capabilities then agents' communicating their beliefs may be equivalent to them sharing their observations (see [23] for comparison).

Third, we describe an approach to decision-theoretic planning, i.e., forward search through the space of interactive beliefs, the agents can use to plan the most advantageous communicative action sequences. It is based on Bellman optimality principle and mirrors value iteration in POMDPs [17, 28] and IPOMDPs [11]. Decision-theoretic planning backs up values of future reachable states while the agents optimize by making their utility maximizing choices. The agents' capability to examine possible futures during planning can shield them from being taken advantage of by insincere speakers. Intuitively, consider that Joe pranks us by lying about the temperature this morning and gets us to come out of the hotel in shorts and T shirt to 10 degrees C. Joe may find it hilarious and get a short term reward, but a more distant possible future may include our getting back at Joe and never believing him again, discounting his initial reward substantially. Our knowing he knows that we can make him pay may allow us to trust him (this is analogous to agents who play a repeated game of Prisoner's Dilemma being able to trust each other to cooperate due to them being able to punish the opponent for defection [3].) On the other hand, if we know that we will not have a chance to get Joe back for lying to us we may want to be

[1] See `https://en.wikipedia.org/wiki/Cheryl's_Birthday`

less trusting and check the temperature for ourselves before getting dressed. In general, out-thinking an opponent in terms of either the strategy level or planning time horizon are crucial in preventing an agent from being exploited (see [19] for some related reading.)

We should mention at the outset that we are barely scratching the surface and this paper leaves much of the important remaining issues to future work. For example, we will not have much to say about the communication language the agents can use, about the vast variety of communicative acts explored in the linguistics literature and how they relate to our approach, and about the intricacies of possible deceptive and exploitative communicative behaviors.

3 The Formalism

We briefly describe interactive POMDPs first since our approach to communication builds on that approach. IPOMDPs extend POMDPs to interactive settings, [11]. We consider only two agents, i and j, for simplicity. A finitely nested interactive POMDP of agent i is defined as:

$$IPOMDP_i = \langle IS_{i,l}, A, \Omega_i, T_i, O_i, R_i \rangle \tag{1}$$

where $IS_{i,l}$ is a set of interactive states, defined as $IS_{i,l} = S \times M_{j,l-1}, l \geq 1$, where S is the set of physical states and $M_{j,l-1}$ is the set of possible models of agent j, and l is the strategy (nesting) level. We consider only one class of models which are the $(l-1)$th level *intentional* models of agent j: $\theta_{j,l-1} = \langle b_{j,l-1}, A_j, \Omega_j, T_j, O_j, R_j \rangle$[2], $b_{j,l-1}$ is agent j's belief nested to the level $(l-1)$, $b_{j,l-1} \in \Delta(IS_{j,l-1})$. The intentional model $\theta_{j,l-1}$, is sometimes called *type*, can be rewritten as $\theta_{j,l-1} = \langle b_{j,l-1}, \hat{\theta}_j \rangle$, where $\hat{\theta}_j$ includes all elements of the intentional model other than the belief and is called the agent j's frame.

The $IS_{i,l}$ can be defined in an inductive manner:

$$\begin{aligned}
&IS_{i,0} = S, && \theta_{j,0} = \{\langle b_{j,0}, \hat{\theta}_j \rangle : b_{j,0} \in \Delta(S)\} \\
&IS_{i,1} = S \times \theta_{j,0}, && \theta_{j,1} = \{\langle b_{j,1}, \hat{\theta}_j \rangle : b_{j,1} \in \Delta(IS_{j,1})\} \\
&..... && \\
&IS_{i,l} = S \times \theta_{j,l-1}, && \theta_{j,l} = \{\langle b_{j,l}, \hat{\theta}_j \rangle : b_{j,l} \in \Delta(IS_{j,l})\}
\end{aligned} \tag{2}$$

All remaining components in an IPOMDP are analogues to those in a POMDP; $A = A_i \times A_j$ is the set of joint actions of all agents, Ω_i is the set of agent i's possible observations, $T_i : S \times A \times S \rightarrow [0, 1]$ is the state transition function, $O_i : S \times A \times \Omega_i \rightarrow [0, 1]$ is the observation function, $R_i : S \times A \rightarrow R$ is the reward function.

[2]We neglect the optimality criterion introduced in [11] for simplicity.

Given all the definitions above, the IPOMDP belief update, derived in [11], is:

$$b_i^t(is^t) = P(is^t|b_i^{t-1}, a_i^{t-1}, o_i^t) = \alpha \sum_{is^{t-1}} b_i^{t-1}(is^{t-1}) \times$$
$$\times \sum_{a_j^{t-1}} P(a_j^{t-1}|\theta_j^{t-1}) T_i(s^{t-1}, a^{t-1}, s^t) \tag{3}$$
$$\times O_i(s^t, a^{t-1}, o_i^t) \times \sum_{o_j^t} O_j(s^t, a^{t-1}, o_j^t) \times$$
$$\times \tau_{\theta_j}(b_j^{t-1}, a_j^{t-1}, o_j^t, b_j^t)$$

Unlike plain belief update in POMDP, the interactive belief update in an IPOMDP takes two additional elements into account. First, the probability of others' actions given their models, defined below, in the second summation is included since the state of physical environment depends on both agents' actions. Second, the modeling agent needs to update its beliefs about the models of others based on the anticipation of what observations the other agent might get and how it updates, represented in τ_{θ_j}.

Customarily the belief update is represented as a function SE so that the new belief is: $b_i^t = SE(b_i^{t-1}, a_i^{t-1}, o_i^t)$. $\tau_{\theta_i}(b_i^{t-1}, a_i^{t-1}, o_i^t, b_i^t)$ is defined as equal 1 when b_i^t is equal to $SE(b_i^{t-1}, a_i^{t-1}, o_i^t)$ and zero otherwise, and similarly for j.

An optimal action, a_i^*, for an infinite horizon criterion with discounting, an element of the set of optimal actions, $OPT(\theta_i)$, defined as:

$$OPT(\theta_i) = \arg\max_{a_i \in A_i} \Big\{ \sum_{is \in IS} b_i(is) ER_i(is, a_i) + \tag{4}$$
$$+\gamma \sum_{o_i \in \Omega_i} P(o_i|a_i, b_i) \times U(\langle SE_{\theta_i}(b_i, a_i, o_i), \hat{\theta}_i \rangle) \Big\}$$

Where the $ER_i(is, a_i)$ is the expected value of the immediate reward to i for executing action a_i given interactive state is and is equal to $\Sigma_{a_j} R_i(is, a_i, a_j) P(a_j|\theta_j)$. The utility (value), $U(b_i)$, of an interactive belief state, which is the first element of θ_i, is defined by the Bellman equation for IPOMDPs, and is analogous to the one above with argmax replaced by max. The agent's ability to compute optimal utility maximizing actions is the basis for rationality. The Bellman equation for IPOMDPs describes the back-up operation during value-driven decision-theoretic search through an agent's interactive beliefs[3], and

[3] Analogously to search through physical state space for MDPs and through single agent belief space in POMDPs.

is a decision-theoretic version of epistemic planning [4, 5, 14, 18]. Same value-driven search through interactive beliefs states applies to planning for communicative behavior, as we show below.

The right-hand side of Equation (4) for any particular action a_i is the expected utility of this action to agent i, $U_i(a_i)$. Equation (4) allows agents to predict which actions other agents are likely to execute. According to hard maximization criterion i would model j as a strict optimizer and predict that j could execute only actions in $OPT(\theta_j)$. A more relaxed criterion is soft maximization [22] according to which the probability of j executing a_j is:

$$P(a_j|\theta_j) = \frac{exp[\lambda U_i(a_j)]}{\sum_{a_j} exp[\lambda U_i(a_j)]} \tag{5}$$

λ above is the rationality parameter; when λ is 0 the choice is random, when λ is infinity the soft max criterion becomes hard max. The same principle of modeling other agents as rational optimizers is central to Bayesian pragmatics of communicative behavior based on Rational Speech Acts model [12]. We use it as well in our formalism below.

3.1 Communicative IPOMDPs

Communicative IPOMDPs (CIPOMDPs) build on IPOMDPs but include additional action of sending a message, m_s, and an additional observation - a message that could be received, m_r. Either message can be nil.

$$CIPOMDP_i = \langle IS_{i,l}, A, M, \Omega_i, T_i, O_i, R_i \rangle \tag{6}$$

where M is a set of messages the agents can send to and receive from each other (so that both m_s and m_r above are in M). All of the other elements are as defined previously for IPOMDPs except that the reward function has an additional argument; $R_i : S \times A \times M \rightarrow R$ so it can also depend on the messages i sends (messages can be costly.) The set of messages constitute the language of communication the agents share. We leave the exact specification of M for future work but we make an assumption that each message in M can be interpreted as a marginal probability distribution spanned on the agents' interactive state spaces IS_i (and IS_j). This allows agent i to send a message containing information about any variable(s) in i's belief space, and similarly for j. Messages with value nil (silence) contain no variables. The fact that the agents' beliefs and messages exchanged are probability distributions facilitates incorporation of information received into the agent's beliefs, subject to it's veracity. Note that while M is over the same state space as agent's beliefs we do not demand that it be in any way tied to the actual beliefs - agents are free to lie or be truthful in any way they find advantageous.

Since interactive states encompass states of the world and other agents' nested models the message spaces defined above can contain information about the world, information about other agents' anticipated actions, beliefs of other agents about the world and what they think others may do, and so on.

3.1.1 Belief Update

We now move on to Bayesian belief update during interaction. For simplicity we consider two agents i and j. At any particular time step agents perform physical actions and observe and 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 $m_{i,r}^{t}$. All messages are in M.

The update below is analogous to belief update in IPOMDPs which updates the probability of interactive states given previous action and current observation: $P(is^t|b_i^{t-1}, a_i^{t-1}, o_i^t)$. 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): $P(is^t|b_i^{t-1}, a_i^{t-1}, m_{i,s}^{t-1}, o_i^t, m_{i,r}^t)$:

Proposition 1:

$$
\begin{aligned}
&b_i^t(is^t) = P(is^t|b_i^{t-1}, a_i^{t-1}, m_{i,s}^{t-1}, o_i^t, m_{i,r}^t) = \\
&= \alpha \sum_{is^{t-1}} b_i^{t-1}(is^{t-1}) \sum_{(m_{j,s}^{t-1}, a_j^{t-1})} Pr(m_{j,s}^{t-1}, a_j^{t-1}|\theta_j^{t-1}) \times \\
&\times O_i(s^t, a^{t-1}, o_i^t) T_i(s^{t-1}, a^{t-1}, s^t) \times \\
&\times \sum_{o_j^t} \tau_{\theta_j^t}(b_j^{t-1}, a_j^{t-1}, m_{j,s}^{t-1}, o_j^t, m_{j,r}^t, b_j^t) O_j(s^t, a^{t-1}, o_j^t)
\end{aligned}
\tag{7}
$$

Proof of Proposition 1 appears in the Appendix. Above update treats messages sent as actions and messages received as additional observations. For simplicity the update above assumes perfect message transmission, i.e., that the message sent by i at $t-1$ is the one received by j at t and that the message received by i at t is the one sent by j at $t-1$; they appear in j's belief update τ_{θ_j}. If transmission could be imperfect i has to maintain probabilities describing accuracy of transmission: $P(m_{j,s}^{t-1}|m_{i,r}^t)$ (probability of what j sent given what i received) and $P_i(m_{j,r}^t|m_{i,s}^{t-1})$ (probability of what j received given what i sent) for all possible $m_{j,s}^{t-1}$ and $m_{j,r}^{,t}$.

Update in Proposition 1 is analogous to belief update in IPOMDPs when it comes to actions and observations. With respect to communication Eq. (7) combines three important elements. First, the updated belief depends on the agent's prior information $b_i^{t-1}(is^{t-1})$, as should be expected. Second, the term $P(m_{i,r}^t|\theta_j^{t-1}, a_j^{t-1})$ (or $P(m_{j,s}^{t-1}|\theta_j^{t-1}, a_j^{t-1})$ in case

of possibly imperfect transmission) is analogous to the accuracy of an observation given a state, which quantifies the relation between the observed signal and the reality that generates that signal. In Proposition 1 the $P(m^t_{i,r}|\theta^{t-1}_j), a^{t-1}_j$ quantifies the relation between the message i received from j and the model, θ_j, of agent j that generated the message. 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 θ_j[4]. We assume that agents are sincere to the extend that it pays off for them; we define this further below. Third, Proposition 1 includes the dependence of agent j's belief and the state of the world included in the interactive state $is =< s, \theta_j >$, both at time t and $t-1$. More explicitly, the joint probability of a particular state s and θ_j is $P(s|\theta_j)P(\theta_j)$ and represents the degree to which i estimates that j is informed about the true state of the world. i maintains its estimate of j's beliefs through the τ_{θ_j} term; if j's observation function contained in θ_j is accurate i would expect that j's beliefs reflect the state of the world s accurately.

In summary, Proposition 1 generalizes IPOMDP belief update to include communication and combines agent i's prior information, i's estimate of the relation between the message it received and agent j's belief about the world, and the relation between j's belief and the state of the world. In the cooperative case, if i thinks j is sincere and well informed, i combines its own prior beliefs with the content of the message to update its beliefs about the world.

Finally, Proposition 1 assumes that the agents' frames (i.e., all of their characteristics other than beliefs - reward functions, action spaces, and transition and observation functions) remain constant.[5] We provide an example of Bayesian pragmatic update in a multi-agent version of the well-known Tiger game environment [17] below.

3.1.2 Decision-Theoretic Planning for Communication and Interaction

Given that belief update in CIPOMDPs[6] is analogous to Equation (3), we similarly proceed to define the Bellman equation which includes communicative actions and hard and soft criteria quantifying speaker's sincerity.

The belief update, defined in Proposition 1, over the whole space IS_i due to communication is again represented as a function SE so that the new belief is: $b^t_i = SE(b^{t-1}_i, a^{t-1}_i, m^{t-1}_{i,s}, o^t_i, m^t_{i,r})$. $\tau_{\theta_i}(b^{t-1}_i, a^{t-1}_i, m^{t-1}_{i,s}, o^t_i, m^t_{i,r}, b^t_i)$ is defined as equal 1 when b^t_i is equal to $SE(b^{t-1}_i, a^{t-1}_i, m^{t-1}_{i,s}, i^t_i, m^t_{i,r})$ and zero otherwise, and analogously for j (which is a factor in Proposition 1 above).

The utility of interactive belief of agent i, contained in i's type θ_i, is:

[4]The message j sent may also depend on j's previous action a^{t-1}_j.

[5]In [11] this assumption is called model non-manipulability.

[6]Recall that either message may be nil – silence can be informative and takes part in the belief update.

$$U_i(\theta_i) = \max_{(m_{i,s},a_i)} \Big\{ \sum_{is \in IS} b_{is}(s) ER_i(is, m_{i,s}, a_i) + \tag{8}$$
$$+\gamma \sum_{(m_{i,r},a_i)} P(m_{i,r}, o_i | b_i, a_i) \times$$
$$\times U_i(\langle SE_{\theta_i}(b_i, a_i, m_{i,s}, o_i, m_{i,r}), \hat{\theta}_i \rangle) \Big\}$$

$ER_i(is, m_{i,s}, a_i)$ 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 $\Sigma_{a_j} R_i(is, a_i, a_j, m_{i,s}) P(a_j|\theta_j)$, ($i$'s reward can depend on the cost of sending $m_{i,s}$ as we mentioned before.)

Equation (13) defines the utility of an interactive belief in θ_i and is the Bellman equation for interactive (physical and communicative) behavior. The agent's ability to compute optimal utility maximizing interactive behavior is the basis for rational interaction. The Bellman equation above describes the back-up operation during value-driven decision-theoretic search through an agent's interactive beliefs reachable by both agents' executing communicative and physical actions during interaction. It is a decision-theoretic version of multi-agent epistemic planning [4, 5, 14, 18]. As we mentioned, the agents' ability to consider possible future interactions could allow them to prevent themselves from being exploited by insincere speakers.

An optimal message[7] action pair, $(m^*_{i,s}, a^*_i)$, agent i should execute (assuming infinite time horizon criterion with discounting) is an element of the set of optimal messages, $OPT(\theta_i)$, defined as:

$$OPT(\theta_i) = \arg\max_{(m_{i,s},a_i)} \Big\{ \sum_{is \in IS} b_{is}(s) ER_i(is, m_{i,s}, a_i) + \tag{9}$$
$$+\gamma \sum_{(m_{i,r},a_i)} P(m_{i,r}, o_i | b_i, a_i) \times$$
$$\times U_i(\langle SE_{\theta_i}(b_i, a_i, m_{i,s}, o_i, m_{i,r}), \hat{\theta}_i \rangle) \Big\}$$

The right-hand side of Equation (14) for any particular message action pair $(m_{i,s}, a_i)$ is the expected utility, $U_i(m_{i,s}, a_i)$, of sending $m_{i,s}$ and executing action a_i, and similarly for agent j.

Equation (14) allows agents to predict which messages and actions are rational for other agents. As for IPOMDPs, there is a hard maximization criterion according to which

[7] There could be more than one optimal message.

i would model j as a strict optimizer and predict that j could only perform interactive actions in $OPT(\theta_j)$. Soft maximization defines the the probability of j sending $m_{j,s}$ and performing a_j as:

$$Pr(m_{j,s}, a_j|\theta_j) = \frac{exp[\lambda U_j(m_{j,s}, a_j)]}{\sum_{(m_{j,s},a_j)} exp[\lambda U_j((m_{j,s}, a_j)]} \tag{10}$$

As before, when λ is 0 the choice is random and when λ is infinity the soft max criterion becomes hard max. Equation (15) treats agents as rational and, when it comes to communication, is central to Rational Speech Acts model [12]. Equation (15) quantifies an agent's sincerity by tying the message it decides to send to it's beliefs (contained in θ_j) by modeling it as a self-interested rational speaker. In other words, a speaker may be insincere and intend to deceive to further its interests, or the communication network (or speaker itself) may be compromised. As we mentioned agents can protect themselves from being lied to by out-thinking the other agent in terms of depth of the nested theories of mind and in terms of the time horizon. Further investigation of this topic is left for future work.

3.1.3 Literal Meaning

Given that the interactive state space IS contains nested agent models, the Bayesian pragmatics update in Proposition 1 descends down all on the levels of nesting of the theories of mind agent has in alternating invocations of the agents' belief update, τ. These simulations are characteristic of game-theoretic pragmatics [15]. The recursion bottoms out at level l with "flat" POMDP models of agents who do not have any explicit models of other agents. At that point messages have only their **literal meaning** (see [7, 8] and references therein.) While our approach builds on this previous work it leaves a number of unanswered questions, we discuss these briefly below.

We assume that an agent that does not model the other agent can still participate in exchange of messages. A **literal speaker** can generate a message (including nil) by choosing a subset of variables it uses to describe the physical state S and forming a message containing information about these variables. The message can then be broadcasted to "no one in particular" (NOIP). Similarly, a **literal listener**, upon receiving a message from NOIP can incorporate the content of the message into its beliefs about the physical state. To model why the speaker and the listener that do not model any other agent(s) would participate in information exchange we can turn to reinforcement learning. Intuitively, we want to model the possibility that sending messages to, and accepting them from, NOIP may have resulted in enhanced rewards in the past.

We propose two functions. G_α is the generation function, with a parameter α, used

by a literal speaker to convert its beliefs about the world into a message it can then send: $G_\alpha(b^t) = m_s^t$. For example G could be a non-deterministic function which returns a nil message with probability $(1-\alpha)$, and a message equal to a marginal over space b^t for a subset of most important variables used to describe IS, with probability α. In this case α is a single number in $[0,1]$ and the message generated is be a true reflection of the agent's beliefs. The value of α can be increased if an agent finds that sending sincere messages results in increased rewards. A richer parametrisation is needed to model generating insincere messages (see [6] for an example of insincere drongo bird's warning calls scaring other animals away from food.)

A combination function, C_β, is one that a literal listener can use to incorporate the incoming message m_r into its beliefs: $b^{t+1} = C_\beta(b^t, m_r^t)$. Given that m_r can be interpreted as a probability distribution a simple combination function is a weighted sum of prior belief and newly received message: $C_\beta = (1-\beta)b^t + \beta m_r$. Again, the value of β can be increased as it turns out that relying on previously received messages leads to positive rewards.

Specification of more precise models of literal listener and speaker, their parametrisations, and parameter update based on reinforcement learning is left for future work.

4 Example

We consider a simple cooperative interaction between two agents engaged in a multi-agent Tiger game [17, 24] 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 (gold is always opposite the tiger.) 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.[8] The observation function O (identical for both agents) specifies the accuracy of observations. We assume that tiger's growls are informative, with 60% accuracy, only if the agents listen (i.e., if tiger is to the left of a listening agent will get a growl from the left, GL, with probability 0.6 and growl from the right, GR, with probability 0.4). If either of them opens the doors the growls

[8] In [11] agents can also hear the creek of an opened door but we neglect this here.

$\langle a_i, a_j \rangle$	TL	TR
OR,OR	20	-50
OL,OL	-50	20
OR,OL	-100	-100
OL,OR	-100	-100
L,L	-1	-1

Table 1: Reward for Multi-agent Tiger Game

have equal chance to come from left or right door and are thus completely uninformative.

What makes the interaction cooperative is the reward function R. We assume reward values as defined in Table 1; a joint action of both agents opening the door with the gold behind it results in the payoff of 20 to each agent, a joint action of both agents opening the door with the tiger behind it results in payoff of -50 each, but when they miscoordinate and each open different door they each get -100. Also, listening costs -1. We will call this reward function a "friend" reward function, R_F. In general, of course, agents cannot directly observe another agent's reward function directly. Here we will assume that it is known the agents' reward function is R_F for simplicity, and that deception is ill advised.

To illustrate the CIPOMDP belief update in this setting we will assume the following scenario: Both agents i and j start off with no information about the position of the tiger and the initial beliefs of agents consist of probability of 0.5 assigned to state TL and 0.5 to TR. Also, both agents listen in time step 1, and then listen again and exchange messages containing their beliefs about tiger's position in time step 2. We will analyze the scenario form the point of view of agent i who has a strategy level of $l = 2$, i.e., i models j as a I-POMDP with strategy level of $l = 1$. This means i models j as modeling i as flat POMDP. As i listens, it will update its belief over its interactive state $IS_{i,2} = \{(TL, IS_{j,1}), (TR, IS_{j,1})\}$ within which $j's$ model is an IPOMDP with interactive states $IS_{j,1} = \{(0.5, \theta_{i,0}), (0.5, \theta_{i,0}\}$. Finally, j's model of i is a POMDP such that $\theta_{i,0} = \langle (0.5, 0.5), A, \Omega, T, O, R_F \rangle$. Here $(0.5, 0.5)$ is $j's$ belief about i's belief of the tiger's position and the other elements are as defined before. i's initial belief, b_i^0, over the two elements in $IS_{i,2}$ is $(0.5, 0.5)$ since i does not know the tiger's position and knows the model of agent j.

Note that if agent j was not present and i listened and received, say, growl form the left, GL', its "flat" belief over $S = (TL, TR)$ would become $(0.6, 0.4)$.[9] Also, if a lone agent

[9]The updated belief in TL is $p(TL|GL') = \frac{p(GL'|TL)p(TL)}{p(GL')} = \frac{0.6 \times 0.5}{0.5} = 0.6$.

i would listen again, and receive second growl GL", it's updated belief over $\{TL, TR\}$ would be $(0.6923, 0.3077)$.[10]

Coming back to our scenario with both agents, initially both listen since opening the doors given agents do not know anything about the tiger's location is ill advised. Again assume that agent i received growl from the left, GL. i now uses CIPOMDP update, Eq. (7), and there now four interactive states that, for i, have positive probabilities: $\{(TL, \langle(0.6, 0.4), \hat{\theta_{j,1}}\rangle), (TL, \langle(0.4, 0.6), \hat{\theta_{j,1}}\rangle),$
$(TR, \langle(0.6, 0.4), \hat{\theta_{j,1}}\rangle), (TR, \langle(0.4, 0.6), \hat{\theta_{j,1}}\rangle)\}$, reflecting the fact that there are two physical states and two possible beliefs of agent j; one, $(0.6, 0.4)$ corresponding to j having received GL and the other corresponding to j getting GR. The probabilities i assigns to these four possible interactive states turn out to be: $(0.36, 0.24, 0.16, 0.24)$; this is $i's$ updated belief at time $t = 1$, b_i^1. Note that $i's$ marginalized belief over physical states $\{TL, TR\}$ at $t = 1$ is $(0.6, 0.4)$ as should be expected. We marginalized j's belief over i's possible beliefs for simplicity. This update gives same result for both IPOMDP as well as CIPOMDP belief updates because agents do not exchange messages.

Now we can consider two further scenarios. First scenario involves agents performing listen action only. It is similar to IPOMDP belief update. Let's suppose i again gets GL. Now, there are six interactive states for i: $\{(TL, \langle(0.692, 0.308), \hat{\theta_{j,1}}\rangle), (TL, \langle(0.5, 0.5), \hat{\theta_{j,1}}\rangle),$
$(TL, \langle(0.308, 0.692), \hat{\theta_{j,1}}\rangle), (TR, \langle(0.692, 0.308), \hat{\theta_{j,1}}\rangle),$
$(TR, \langle(0.5, 0.5), \hat{\theta_{j,1}}\rangle), (TR, \langle(0.308, 0.692), \hat{\theta_{j,1}}\rangle)\}$. The probabilities i assigns to these six interactive states turn out to be: $(0.2492, 0.3323, 0.1107, 0.0492, 0.14769, 0.110769)$. and i's marginalized belief over physical states is $(0.692, 0.308)$. At this point ($t = 2$), i's optimal action would be to listen because expected cost of opening the door is still higher than listening.[11]. i also knows j won't open the door. Moreover i knows j knows i won't open the door because of two levels of nested modeling in this case.

In the second scenario, the agents not only perform listening actions at $t = 1$, but exchange messages containing their true beliefs about tiger's location. Assume that at $t = 1$ i sends $m_{i,s}^1$ with the content "Probability of tiger being on the left is 0.6", and that at time $t = 2$, i receives $m_{i,r}^2$ with the identical content j sent at $t = 1$. Assuming i again gets growl left as observation at time $t = 2$, we use CIPOMDP belief update equation to get updated belief of i. The content of the message sent by sincere speaker j[12] means that the term $P(m_{i,r}^2|\theta_j^1)$ is 1 only for models of j in which $j's$ belief is $(0.6, 0.4)$ and zero for others. After receiving the message and observing GL, the probabilities assigned to four interactive states $\{(TL, \langle(0.77, 0.23), \hat{\theta_{j,1}}\rangle), (TL, \langle(0.6, 0.4), \hat{\theta_{j,1}}\rangle),$

[10] $p(TL|GL', GL") = \frac{p(GL"|TL)p(TL|GL')}{p(GL"|GL')} = \frac{0.6 \times 0.6}{0.6^2 + 0.4^2}$

[11] 0.692*20+0.308*-50 = -1.56 < -1

[12] Since the interaction is cooperative.

$(TR, \langle (0.77, 0.23), \hat{\theta_{j,1}} \rangle), (TR, \langle (0.6, 0.4), \hat{\theta_{j,1}} \rangle)\}$ will be $(0.4628, 0.3085, 0.091, 0.1371)$. As expected i's marginalized belief over physical states is $(0.77, 0.23)$. This is the same belief update as if i would have listened at time $t = 2$ and got GL at time $t = 3$. Thus, exchanging messages allowed the agents to update their beliefs about the state of the physical world beyond what was contained in the observations they made – their sharing information, in this simple case, is equivalent to an additional observation.

5 Conclusion and Future Work

We presented an approach to decision-theoretic planning for communication and interaction by building on IPOMDPs. We added to IPOMDPs the capability for agents to exchange messages, derived Bayesian update of agents' beliefs due to message exchange, which can be simultaneous with physical actions and observations, and formulated Bellman optimality for interactive behavior.

The approach we discussed leaves a number of important avenues open for future work. First, more detailed specification of the agent communication language, M is needed. For example, how does an agent formulate messages not only about its current beliefs about interactive states IS but also about the observations it has received, predicted behaviors of other agents, etc? Second, the generation of all possible insincere messages needs to be further described and parametrised. Third, the search techniques used to plan interactive and communicative behaviors have to be defined, Since possible messages span a continuous space of probability distributions. The infinite POMDPs [**Doshi-Velez2009:Infinite**] techniques seem promising in this regard. Finally, a principled investigation to the deceptive communicative behavior, paired with physical actions, based on the Bayesian pragmatics update we derive in Proposition 1 is fascinating avenue for future work, which should compare our approach to game-theoretic techniques [**Kamenica2017:Bayesian**, 16].

********** The section below is the earlier one, without physical action and observation.

5.1 Communicative IPOMDPs

Communicative IPOMDPs (CIPOMDPs) build on IPOMDPs but include additional action of sending a message, m_s, and an additional observation - a message that could be received, m_r. Either message can be nil.

$$CIPOMDP_i = \langle IS_{i,l}, A, M, \Omega_i, T_i, O_i, R_i \rangle \tag{11}$$

where M is a set of messages the agents can send to and receive from each other (so that both m_s and m_r above are in M). All of the other elements are as defined previously for IPOMDPs except that the reward function has an additional argument; $R_i : S \times A \times M \rightarrow R$ so it can also depend on the messages i sends (messages can be costly.) The set of messages constitute the language of communication the agents share. We leave the exact specification of M for future work but we make an assumption that each message in M can be interpreted as a marginal probability distribution spanned on the agents' interactive state spaces IS_i (and IS_j). This allows agent i to send a message containing information about any variable(s) in i's belief space, and similarly for j. Messages with value nil (silence) contain no variables. The fact that the agents' beliefs and messages exchanged are probability distributions facilitates incorporation of information received into the agent's beliefs, subject to it's veracity. Note that while M is over the same state space as agent's beliefs we do not demand that it be in any way tied to the actual beliefs - agents are free to lie or be truthful in any way they find advantageous.

Since interactive states encompass states of the world and other agents' nested models the message spaces defined above can contain information about the world, information about other agents' anticipated actions, beliefs of other agents about the world and what they think others may do, and so on.

5.1.1 Bayesian Pragmatics

We now move on to Bayesian belief update due to communication. For simplicity we consider only communication between two agents i and j, and assume that at the particular time step agent's physical actions are nil (joint action of both agents (nil, nil) is denoted as Nil), and that both agents' observations are also nil. Consequently the update below is not full belief update for CIPOMDPs, which should also allow for agents' actions and observations as well as communication. The full belief update for CIPOMDPs is left for future work. At each time agents can 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 $m_{i,r}^{t}$. All messages are in M.

The update below is analogous to belief update in IPOMDPs which updates the probability of interactive states given previous action and current observation: $P(is^t|b_i^{t-1}, a_i^{t-1}, o_i^t)$. The full 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): $P(is^t|b_i^{t-1}, a_i^{t-1}, m_{i,s}^{t-1}, o_i^t, m_{i,r}^t)$, but as we mentioned we consider only update due to communication and we compute $P(is^t|b_i^{t-1}, nil, m_{i,s}^{t-1}, nil, m_{i,r}^t)$:

Proposition 1:

$$b_i^t(is^t) = P(is^t|b_i^{t-1}, nil, m_{i,s}^{t-1}, nil, m_{i,r}^t) = \tag{12}$$
$$= \alpha \sum_{is^{t-1}} b_i^{t-1}(is^{t-1}) P(m_{i,r}^t|\theta_j^{t-1}) T_i(s^{t-1}, Nil, s^t)$$
$$\times \tau_{\theta_j}(b_j^{t-1}, nil, m_{i,r}^t, nil, m_{i,s}^{t-1}, b_j^t)$$

Proof of Proposition 1 appears in the Appendix. Above update treats messages sent as actions and messages received as additional observations. For simplicity the update above assumes perfect message transmission, i.e., that the message sent by i at $t-1$ is the one received by j at t and that the message received by i at t is the one sent by j at $t-1$; they appear in j's belief update τ_{θ_j}. If transmission could be imperfect i has to maintain probabilities describing accuracy of transmission: $P(m_{j,s}^{t-1}|m_{i,r}^t)$ (probability of what j sent given what i received) and $P_i(m_{j,r}^t|m_{i,s}^{t-1})$ (probability of what j received given what i sent) for all possible $m_{j,s}^{t-1}$ and $m_{j,r}^t$. These would enter into Equation (12) and $j's$ update would then be $\tau_{\theta_j}(b_j^{t-1}, nil, m_{j,s}^{t-1}, nil, m_{j,r}^t, b_j^t)$.

Bayesian pragmatics formalized in Proposition 1, combines three important elements. First, the updated belief depends on the agent's prior information $b_i^{t-1}(is^{t-1})$, as should be expected. Second, the term $P(m_{i,r}^t|\theta_j^{t-1})$ (or $P(m_{j,s}^{t-1}|\theta_j^{t-1})$ in case of possibly imperfect transmission) is analogous to the accuracy of an observation given a state, which quantifies the relation between the observed signal and the reality that generates that signal. In Proposition 1 the $P(m_{i,r}^t|\theta_j^{t-1})$ quantifies the relation between the message i received from j and the model, θ_j, of agent j that generated the message. 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 θ_j. We assume that agents are sincere to the extend that it pays off for them; we define this further below. Third, Proposition 1 includes the dependence of agent j's belief and the state of the world included in the interactive state $is =< s, \theta_j >$, both at time t and $t-1$. More explicitly, the joint probability of s and θ_j is $P(s|\theta_j)P(\theta_j)$ and represents the degree to which i estimates that j is informed about the true state of the world. i maintains its estimate of j's beliefs through the τ_{θ_j} term; if j's observation function contained in θ_j is accurate i would expect that j's beliefs reflect the state of the world s.

In summary, Proposition 1 combines agent i's prior information, i's estimate of the relation between the message it received and agent j's belief about the world, and the relation between j's belief and the state of the world. In the intuitive cooperative case, i.e., if i thinks j is sincere and well informed, i combines its own prior beliefs with the content of the message to update its beliefs about the world.

The $T_i(s^{t-1}, Nil, s^t)$ term in Proposition 1 accounts for a possibly dynamic environment which can transition from s^{t-1} to s^t even if neither of the agents perform any physical

actions. Finally, Proposition 1 assumes that the agents' frames (i.e., all of their characteristics other than beliefs - reward functions, action spaces, and transition and observation functions) remain constant.[13] We provide an example of Bayesian pragmatic update in a multi-agent version of the well-known Tiger game environment [17] below.

5.1.2 Decision-Theoretic Planning for Communicative Acts

Given that belief update due to message exchange[14] is analogous to Equation (3), we similarly proceed to define the Bellman equation for communicative actions and hard and soft criteria quantifying the speaker's sincerity.

The belief update, defined in Proposition 1, over the whole space IS_i due to communication is again represented as a function SE so that the new belief is: $b_i^t = SE(b_i^{t-1}, nil, m_{i,s}^{t-1}, nil, m_{i,r}^t)$. $\tau_{\theta_i}(b_i^{t-1}, nil, m_{i,s}^{t-1}, nil, m_{i,r}^t, b_i^t)$ is defined as equal 1 when b_i^t is equal to $SE(b_i^{t-1}, nil, m_{i,s}^{t-1}, nil, m_{i,r}^t)$ and zero otherwise, and analogously for j (which is a factor in Proposition 1 above).

The utility of interactive belief of agent i, contained in i's type θ_i, is:

$$U_i(\theta_i) = \max_{m_{i,s} \in M} \Big\{ \sum_{is \in IS} b_{is}(s) ER_i(is, m_{i,s}) + \tag{13}$$
$$+\gamma \sum_{m_{i,r} \in M} P(m_{i,r}|b_i) \times$$
$$\times U_i(\langle SE_{\theta_i}(b_i, nil, m_{i,s}, nil, m_{i,r}), \hat{\theta}_i \rangle) \Big\}$$

$ER_i(is, m_{i,s})$ above is the immediate reward to i for sending $m_{i,s}$ given the interactive state is and is equal to $R_i(is, nil, nil, m_{i,s})$, ($i$'s reward can depend on the cost of sending $m_{i,s}$ as we mentioned before.)

Equation (13) defines the utility of an interactive belief in θ_i and is the Bellman equation for communicative behavior[15]. The agent's ability to compute optimal utility maximizing communicative actions is the basis for rational communicative behavior. The Bellman equation above describes the back-up operation during value-driven decision-theoretic search through an agent's interactive beliefs reachable by both agents' executing communicative acts. It is a decision-theoretic version of multi-agent epistemic planning [4, 5, 14, 18]. As we mentioned, the agents' ability to consider possible future communications (and in general interactions) could allow them to prevent themselves from being exploited.

[13] In [11] this assumption is called model non-manipulability.

[14] Recall that either message may be nil – silence can be informative and takes part in the belief update.

[15] As mentioned before, we account only for communicative behaviors here. Including impact of possibly simultaneous physical actions is left for future work.

An optimal message[16], $m^*_{i,s}$, agent i should send (assuming infinite time horizon criterion with discounting) is an element of the set of optimal messages, $OPT(\theta_i)$, defined as:

$$OPT(\theta_i) = \arg\max_{m_{i,s} \in M} \Big\{ \sum_{is \in IS} b_{is}(s) ER_i(is, m_{i,s}) + \tag{14}$$
$$+\gamma \sum_{m_{i,r} \in M} P(m_{i,r}|b_i) \times$$
$$\times U_i(\langle SE_{\theta_i}(b_i, nil, m_{i,s}, nil, m_{i,r}), \hat{\theta}_i \rangle) \Big\}$$

The right-hand side of Equation (14) for any particular message $m_{i,s}$ is the expected utility, $U_i(m_{i,s})$, of sending $m_{i,s}$, and similarly for agent j.

Equation (14) allows agents to predict which messages are rational for other agents. As for IPOMDPs above, there is a hard maximization criterion according to which i would model j as a strict optimizer and predict that j could send only messages in $OPT(\theta_j)$. Soft maximization defines the the probability of j sending $m_{j,s}$ as:

$$Pr(m_{j,s}|\theta_j) = \frac{exp[\lambda U_j(m_{j,s})]}{\sum_{a_j} exp[\lambda U_j((m_{j,s})]} \tag{15}$$

As before, when λ is 0 the choice is random and when λ is infinity the soft max criterion becomes hard max. Equation (15) treats agents as rational speakers and is central to Rational Speech Acts model [12]. Equation (15) quantifies an agent's sincerity by tying the message it decides to send to it's beliefs (contained in θ_j) by modeling it as a self-interested rational speaker. In other words, a speaker may be insincere and intend to deceive to further its interests, or the communication network (or speaker itself) may be compromised. As we mentioned agents can protect themselves from being lied to by out-thinking the other agent in terms of depth of the nested theories of mind and in terms of the time horizon. Further investigation of this topic is left for future work.

5.1.3 Literal Meaning

Given that the interactive state space IS contains nested agent models, the Bayesian pragmatics update in Proposition 1 descends down all on the levels of nesting of the theories of mind agent has in alternating invocations of the agents' belief update, τ. These simulations are characteristic of game-theoretic pragmatics [15]. The recursion bottoms out

[16]There could be more than one optimal message.

at level l with "flat" POMDP models of agents who do not have any explicit models of other agents. At that point messages have only their **literal meaning** (see [7, 8] and references therein.) While our approach builds on this previous work it leaves a number of unanswered questions, we discuss these briefly below.

We assume that an agent that does not model the other agent can still participate in exchange of messages. A **literal speaker** can simply generate a message (including nil) by choosing a subset of variables it uses to describe the physical state S and forming a message containing information about these variables. The message can then be broadcasted to "no one in particular" (NOIP). Similarly, a **literal listener**, upon receiving a message from NOIP can incorporate the content of the message into its beliefs about the physical state. To model why the speaker and the listener that do not model any other agent(s) would participate in information exchange we can turn to reinforcement learning. Intuitively, we want to model the possibility that sending messages to, and accepting them from, NOIP may have resulted in enhanced rewards in the past.

We propose two functions. G_α is the generation function, with a parameter α, used by a literal speaker to convert its beliefs about the world into a message it can then send: $G_\alpha(b^t) = m_s^t$. For example G could be a nondeterministic function which returns a nil message with probability $(1 - \alpha)$, and a message equal to a marginal over space b^t for a subset of most important variables used to describe IS with probability α. In that case α is a single number in $[0, 1]$ and the message generated would be a true reflection of the agent's beliefs. The value of α can be increased if an agent finds that sending sincere messages results in increased rewards. A richer parametrisation is needed to model generating insincere messages (see [6] for an example of insincere drongo bird's warning calls scaring other animals away from food.)

A combination function, C_β, is one that a literal listener can use to incorporate the incoming message m_r into its beliefs: $b^{t+1} = C_\beta(b^t, m_r^t)$. Given that m_r can be interpreted as a probability distribution a simple combination function is a weighted sum of prior belief and newly received message: $C_\beta = (1 - \beta)b^t + \beta m_r$. Again, the value of β can be increased as it turns out that relying on previously received messages leads to positive rewards.

Specification of more precise models of literal listener and speaker, their parametrisations, and parameter update based on reinforcement learning is left for future work.

6 Example

We consider a simple cooperative interaction between two agents engaged in a multi-agent Tiger game [17, 24] 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 (gold is always opposite the tiger.) 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 and perform a nil action, so the actions are: $A = \{OR, OL, L, nil\}$ 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 or perform the nil action, 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.[17] The observation function O (identical for both agents) specifies the accuracy of observations. As is usual we assume that tiger's growls are informative, with 85% accuracy, only if the agents listen (i.e., if tiger is of the left a listening agent will get a growl from the left, GL, with probability 0.85 and growl from the right, GR, with probability 0.15. If either of them opens the doors the growls have equal chance to come from left or right door and are thus completely uninformative.

What makes the interaction cooperative is the reward function R. We assume reward values following [24]; a joint action of both agents opening the door with the gold behind it results in the payoff of 20 to each agent, a joint action of both agents opening the door with the tiger behind it results in payoff of -50 each, but when they miscoordinate and each open different door they each get -100. Also, the nil actions are cost-free and listening costs -1. We will call this reward function a "friend" reward function, R_F. In general, of course, agents cannot directly observe another agent's reward function directly. Here we will assume that it is known the agents' reward function is R_F for simplicity, and that deception is ill advised.

To illustrate the Bayesian pragmatics in this setting we will assume the following scenario: Both agents i and j start off with no information about the position of the tiger and the initial beliefs of agents consist of probability of 0.5 assigned to state TL and 0.5 to TR. Also, both agents first listen, and then exchange messages containing their beliefs about tiger's position. We will analyze the scenario form the point of view of agent i who has a strategy level of $l = 1$, i.e., i models j using a "flat" POMDP. As i listens, it will update its belief over its interactive state $IS_{i,1} = \{(TL, \theta_{j,0}), (TR, \theta_{j,0}\}$ within which $j's$ model is a POMDP $\theta_{j,0} = \langle (0.5, 0.5), A, \Omega, T, O, R_F \rangle$. Here $(0.5, 0.5)$ is $j's$ belief about the tiger's position and the other elements as defined before. i's initial belief, b_i^0, over the two elements in $IS_{i,1}$ is $(0.5, 0.5)$ since i does not know the tiger's position and knows the model of agent j.

Note that if i were alone with the tiger, listened and received, say, growl form the

[17] In [11]agents can also hear the creek of an opened door but we neglect this here.

left, GL', its "flat" belief over $S = (TL, TR)$ would become $(0.85, 0.15)$.[18] Also, if a lone agent i would listen again, and receive second growl GL", it's updated belief over $\{TL, TR\}$ would be $(0.97, 0.03)$.[19]

Coming back to our scenario with both agents, initially both listen since opening the doors given they do not know anything about the tiger's location is ill advised. Again assume that agent i received growl from the left, GL. i now uses interactive update, Eq. (3) and there now four interactive states that, for i, have positive probabilities: $\{(TL, \langle(0.85, 0.15), \hat{\theta_{j,0}}\rangle), (TL, \langle(0$ $(TR, \langle(0.85, 0.15), \hat{\theta_{j,0}}\rangle), (TR, \langle(0.15, 0.85), \hat{\theta_{j,0}}\rangle)\}$, reflecting the fact that there are two physical states and two possible beliefs of agent j; one, $(0.85, 0.15)$ corresponding to j having received GL and the other corresponding to j getting GR. The probabilities i assigns to these four possible interactive states turn out to be: $(0.7225, 0.1275, 0.0225, 0.1275)$; this is $i's$ updated belief at time $t = 1$, b_i^1. Note that $i's$ marginalized belief over physical states $\{TL, TR\}$ at that time is $(0.85, 0.15)$.

Now we can consider agents' exchanging messages containing their true beliefs about tiger's location. Assume that at $t = 1$ i sends $m_{s,i}^1$ with the content "Probability of tiger being on the left is 0.85", that at time $t = 2$, i receives $m_{i,r}^2$ with the identical content from j, and there are only nil actions and observations at time $t = 1$. We use Bayesian pragmatics, Eq. (12), to compute b_i^1: In τ_{θ_j} the model of j in a "flat" POMDP so literal interpretation applies; assuming that $\beta = 1$ in agent $j's$ combination function, the b_j^2 becomes $(0.85, 0.15)$, but recall that it was also possible that $j's$ belief at $t = 1$ $(0.15, 0.85)$, in two of the interactive states i considered possible at $t = 1$. But, since $m_{i,r}^2$ was sent by j at time $t = 1$, and can be assumed to be sincere, i.e., $P(m_{i,r}^2|\theta_j^1)$ is 1 only for models of j in which $j's$ belief is $(0.85, 0.15)$ and zero for others. Consequently, $i's$ probability distribution over four interactive states listed above becomes $b_i^2 = \alpha(0.7225, 0.0, 0.0225, 0.0)$. α is the normalization constant; the normalized b_i^2 over the two possible states $\{(TL, \langle(0.85, 0.15), \hat{\theta_{j,0}}\rangle), (TR, \langle(0.85, 0.15), \hat{\theta_{j,0}}\rangle)\}$ becomes $(\frac{0.7225}{0.7225+0.0225}, \frac{0.0225}{0.7225+0.0225}) = (0.97, 0.03)$. This belief awards the **same** probability to TL as if i was alone and listened twice, as should be expected (see also [23].) Also, if i received a message from j stating that probability of TL is 0.15 instead, i's updated probability would be $(0.5, 0.5)$ – also the expected result. Finally, i's model of j's belief does **not** get updated; this is because, at that level, the message j receives is interpreted literally and just confirms js' prior belief.

[18] The updated belief in TL is $p(TL|GL') = \frac{p(GL'|TL)p(TL)}{p(GL')} = \frac{0.85 \times 0.5}{0.5} = 0.85$.

[19] $p(TL|GL', GL") = \frac{p(GL"|TL)p(TL|GL')}{p(GL"|GL')} = \frac{0.85 \times 0.85}{0.85^2 + 0.15^2}$

7 Conclusion

We presented an approach to Bayesian pragmatics and decision-theoretic planning for communicative acts by building on IPOMDPs. We added to IPOMDPs the capability for agents to exchange messages, derived Bayesian update of agents' beliefs due to messages and formulated Bellman optimality for communicative behavior.

References

[1] Robert J. Aumann. "Interactive Epistemology I: Knowledge". In: *International Journal of Game Theory* 28 (1999), pp. 263–300.

[2] J. L. Austin. *How to do Things with Words*. Clarendon Press, 1962.

[3] Robert Axelrod. *The Evolution of Cooperation*. Basic Books, 1984.

[4] Thomas Bolander and Mikkel Birkegaard Andersen. "Epistemic Planning for Single- and Multi-Agent". In: *Journal of Applied Non-Classical Logics* 21 (2011), pp. 9–34.

[5] Thorsten Engesser et al. "Cooperative Epistemic Multi-Agent Planning for Implicit Coordination". In: *Electronic Proceedings in Theoretical Computer Science (ISSN: 2075-2180) (DOI: http://dx.doi.org/10.4204/EPTCS.243.6)* 243 (2017).

[6] Tom P. Flower, Matthew Gribble, and Amanda R. Ridley. "Deception by Flexible Alarm Mimicry in an African Bird". In: *Science* 334 (2014), pp. 513–516.

[7] Michael C. Frank and Noah D. Goodman. "Predicting pragmatic reasoning in language games". In: *Science* 336 (2012), pp. 998–998.

[8] Michael Franke and Gerhardt Jager. "Probabilistic pragmatics, or why Bayes' rule is probably important for pragmatics". In: *Zeitschrift fur Sprachwissenschaft* 35 (2016), pp. 3–44.

[9] Chris Frith and Uta Frith. "Theory of mind". In: *Current Biology* 15.17 (2005), R644–R645. ISSN: 0960-9822. DOI: http://dx.doi.org/10.1016/j.cub.2005.08.041. URL: http://www.sciencedirect.com/science/article/pii/S0960982205009607.

[10] Vittorio Gallese and Alvin Goldman. "Mirror neurons and the simulation theory of mind-reading". In: *Trends in Cognitive Sciences* 2.12 (1998), pp. 493–501. ISSN: 1364-6613. DOI: http://dx.doi.org/10.1016/S1364-6613(98)01262-5. URL: http://www.sciencedirect.com/science/article/pii/S1364661398012625.

[11] Piotr Gmytrasiewicz and Prashant Doshi. “A Framework for Sequential Planning in Multiagent Settings”. In: *Journal of Artificial Intelligence Research* 24 (2005). http://jair.org/contents/v24.html, pp. 49–79.

[12] Noah D. Goodman and Daniel Lassiter. “Probabilistic Semantics and Pragmatics: Uncertainty in Language and Thought, A draft chapter”. In: *Wiley-Blackwell Handbook of Contemporary Semantics fffdfffdfffd second edition, https://web.stanford.edu/ ngoodman/papers HCS-final.pdf*. Ed. by Shalom Lappin and Chris Fox. 2014.

[13] H. P. Grice. “Meaning”. In: *Philosophical Review* LXVI (1957), pp. 377–388.

[14] W. V. D. Hoek and M. Wooldridge. “Tractable Multiagent Planning for Epistemic Goals”. In: *Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS-2002)*. 2002, 1167fffdfffdfffd1174.

[15] Gerhardt Jaeger. “Game-theoretical pragmatics”. In: *J. van Benthem and A. ter Meulen, eds., Handbook of Logic and Language, 2nd edition*. Elsevier, 2011, pp. 467–491.

[16] David Ettinger Philippe Jehiel. *Towards a Theory of Deception*. Tech. rep. UCLA Department of Economics, 2006.

[17] Leslie P. Kaelbling, Michael L. Littman, and Anthony R. Cassandra. “Planning and Acting in Partially Observable Stochastic Domains”. In: *Artificial Intelligence* 101.2 (1998), pp. 99–134.

[18] Filippos Kominis and Hector Geffner. “Beliefs in Multiagent Planning: From One Agent to Many”. In: *Proceedings of ICAPS*. 2015, pp. 157–155.

[19] Maria Konnikova. *The Confidence Game; Why we Fall for It Every Time*. Viking, 2016.

[20] Alan M. Leslie, Ori Friedman, and Tim P. German. “Core mechanisms in fffdfffdfffdtheory of mindfffdfffdfffd”. In: *Trends in Cognitive Sciences* 8.12 (2004), pp. 528–533. DOI: http://dx.doi.org/10.1016/j.tics.2004.10.001. URL: http://www.cell.com/trends/cognitive-sciences/comments/S1364-6613(04)00260-8.

[21] David Lewis. “Scorekeeping in a language game”. In: *Journal of Philosophical Logic* 8.1 (1979), pp. 339–359.

[22] Richard McKelvey and Thomas Palfrey. “Quantal Response Equilibria for Normal Form Games”. In: *Games and Economic Behavior* 10 (1995), pp. 6–38.

[23] Ranjit Nair et al. “Communication for Improving Policy Computation in Distributed POMDPs”. In: *Proceedings of the Agents and Autonomous Multiagent Systems (AAMAS)*. 2004.

[24] Ranjit Nair et al. “Taming Decentralized POMDPs: Towards Efficient Policy Computation for Multiagent Settings”. In: *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*. 2003.

[25] Frans Olihoek, Mathijs Spaan, and Nikos Vlassis. “DEC-POMDPs with Delayed Communication”. In: *Proceedings of MSDM 2007 May 15, 2007, Honolulu, Hawaifffdfffdfffdi, USA*. 2007.

[26] Ahti Pietarinen. *Game Theory and Linguistic Meaning*. Vol. 18. Elsevier: Current in the Semantics/Pragmatics Interface, 2007.

[27] D. Pynadath and M. Tambe. “The communicative multiagent team decision problem: Analyzing teamwork theories and models”. In: *Journal of AI Research (JAIR)* (2002).

[28] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 2010.

[29] Karen Shanton and Alvin Goldman. “Simulation theory”. In: *Wiley Interdisciplinary Reviews: Cognitive Science* 1.4 (2010), pp. 527–538. ISSN: 1939-5086. DOI: 10.1002/wcs.33. URL: http://dx.doi.org/10.1002/wcs.33.

[30] R. Stalnaker. “Assertion”. In: *Syntax and Semantics 9: Pragmatics*. Ed. by P. Cole. Academic Press, 1978.

[31] Adam Vogel, Christopher Potts, and Dan Yurafsky. “Implicatures and nested beliefs in approximate decentralized-pomdps”. In: ACL (2013).

[32] Feng Wu, Shlomo Zilbersein, and Xiaoping Chen. “Online planning for multi-agent systems with bounded communication”. In: *Artificial Intelligence* 167 (2011), pp. 487–511.

[33] Henk Zeevat. “Perspectives on Natural Language Semantics and Pragmatics”. In: *Bayesian Natural Language Semantics and Pragmatics*. Ed. by Schmitz Zeevat Henk and Hans-Christian. Springer, 2015, pp. 1–24.

Appendices

A Proof of proposition 1

$$b_i^t(is^t) = \alpha \sum_{is^{t-1}} b_i^{t-1}(is^{t-1}) \sum_{a_j^{t-1}} Pr(a_j^{t-1}|\theta_j^{t-1}) O_i(s^t, a^{t-1}, o_i^t) Pr(m_{i,r}^t|\theta_j^{t-1}, a_j^{t-1}) I(\hat{\theta}_j^{t-1}, \hat{\theta}_j^t) T_i(s^{t-1}, a^{t-1}, s^t)$$
$$\times \sum_{o_j^t} \tau_{\theta_j^t}(b_j^{t-1}, a_j^{t-1}, m_{j,s}^{t-1}, o_j^t, m_{j,r}^t, b_j^t) O_j(s^t, a^{t-1}, o_j^t)$$

$$\begin{aligned}
b_i^t(is^t) &= Pr(is^t|b_i^{t-1}, a_i^{t-1}, m_{i,s}^{t-1}, o_i^t, m_{i,r}^t) \\
&= \frac{Pr(is^t, b_i^{t-1}, a_i^{t-1}, m_{i,s}^{t-1}, o_i^t, m_{i,r}^t)}{Pr(b_i^{t-1}, a_i^{t-1}, m_{i,s}^{t-1}, o_i^t, m_{i,r}^t)} \\
&= \frac{Pr(is^t, o_i^t, m_{i,r}^t|b_i^{t-1}, a_i^{t-1}, m_{i,s}^{t-1}) Pr(b_i^{t-1}, a_i^{t-1}, m_{i,s}^{t-1})}{Pr(o_i^t, m_{i,r}^t|b_i^{t-1}, a_i^{t-1}, m_{i,s}^{t-1} Pr(b_i^{t-1}, a_i^{t-1}, m_{i,s}^{t-1})} \\
&= \alpha Pr(is^t, o_i^t, m_{i,r}^t|b_i^{t-1}, a_i^{t-1}, m_{i,s}^{t-1}) \\
&= \alpha \sum_{is^{t-1}} b_i^{t-1}(is^{t-1}) \sum_{a_j^{t-1}} Pr(is^t, o_i^t, m_{i,r}^t|a_i^{t-1}, a_j^{t-1}, m_{i,s}^{t-1}, is^{t-1}) Pr(a_j^{t-1}|a_i^{t-1}, is^{t-1}, m_{i,s}^{t-1}) \\
&= \alpha \sum_{is^{t-1}} b_i^{t-1}(is^{t-1}) \sum_{a_j^{t-1}} Pr(is^t, m_{i,r}^t|a^{t-1}, m_{i,s}^{t-1}, is^{t-1}) Pr(o_i^t|a^{t-1}, m_{i,s}^{t-1}, is^{t-1}) Pr(a_j^{t-1}|is^{t-1}, m_{i,s}^{t-1})
\end{aligned}$$

$$b_i^t(is^t) = \alpha \sum_{is^{t-1}} b_i^{t-1}(is^{t-1}) \sum_{a_j^{t-1}} Pr(is^t, m_{i,r}^t|a^{t-1}, m_{i,s}^{t-1}, is^{t-1}) Pr(o_i^t|a^{t-1}, m_{i,s}^{t-1}, is^{t-1}) Pr(a_j^{t-1}|is^{t-1}, m_{i,s}^{t-1}) \tag{16}$$

Now,

$$Pr(is^t, m^t_{i,r}|a^{t-1}, m^{t-1}_{i,s}, is^{t-1}) = \frac{Pr(is^t, m^t_{i,r}, a^{t-1}, m^{t-1}_{i,s}, is^{t-1})}{Pr(a^{t-1}, m^{t-1}_{i,s}, is^{t-1})}$$

for intentional model,

$$is^t = (s^t, \theta^t_j) = (s^t, b^t_j, \hat{\theta}^t_j)$$

$$\begin{aligned}
Pr(is^t, m^t_{i,r}|a^{t-1}, m^{t-1}_{i,s}, is^{t-1}) &= \frac{Pr(s^t, b^t_j, \hat{\theta}^t_j, m^t_{i,r}, a^{t-1}, m^{t-1}_{i,s}, is^{t-1})}{Pr(a^{t-1}, m^{t-1}_{i,s}, is^{t-1})} \\
&= \frac{Pr(b^t_j|s^t, \hat{\theta}^t_j, m^t_{i,r}, a^{t-1}, m^{t-1}_{i,s}, is^{t-1}) Pr(s^t, \hat{\theta}^t_j, m^t_{i,r}, a^{t-1}, m^{t-1}_{i,s}, is^{t-1})}{Pr(a^{t-1}, m^{t-1}_{i,s}, is^{t-1})} \\
&= Pr(b^t_j|s^t, \hat{\theta}^t_j, m^t_{i,r}, a^{t-1}, m^{t-1}_{i,s}, is^{t-1}) Pr(s^t, \hat{\theta}^t_j, m^t_{i,r}|a^{t-1}, m^{t-1}_{i,s}, is^{t-1}) \\
&= Pr(b^t_j|s^t, \hat{\theta}^t_j, m^t_{i,r}, a^{t-1}, m^{t-1}_{i,s}, is^{t-1}) Pr(s^t, \hat{\theta}^t_j|a^{t-1}, m^{t-1}_{i,s}, is^{t-1}) \\
&\quad Pr(m^t_{i,r}|a^{t-1}, m^{t-1}_{i,s}, is^{t-1}) \\
&= Pr(b^t_j|s^t, \hat{\theta}^t_j, m^t_{i,r}, a^{t-1}, m^{t-1}_{i,s}, is^{t-1}) Pr(\hat{\theta}^t_j|s^t, a^{t-1}, m^{t-1}_{i,s}, is^{t-1}) \\
&\quad Pr(s^t|a^{t-1}, is^{t-1}) Pr(m^t_{i,r}|a^{t-1}, m^{t-1}_{i,s}, s^{t-1}, \theta^{t-1}_j)
\end{aligned}$$

Due to MNM assumption,

$$Pr(\hat{\theta}^t_j|s^t, a^{t-1}, is^{t-1}) = I(\hat{\theta}^{t-1}, \hat{\theta}^t_j)$$

and,$m^t_{i,r} \perp (a^{t-1}_i, m^{t-1}_{i,s}, s^{t-1})|\theta^{t-1}_j, a^{t-1}_j$

$$Pr(is^t, m^t_{i,r}|a^{t-1}_i, m^{t-1}_{i,s}, is^{t-1}) = Pr(b^t_j|s^t, \hat{\theta}^t_j, m^t_{i,r}, a^{t-1}, m^{t-1}_{i,s}, is^{t-1}) I(\hat{\theta}^{t-1}_j, \hat{\theta}^t_j) T_i(s^{t-1}, a^{t-1}, s^t) Pr(m^t_{i,r}|\theta^{t-1}_j, a^{t-1}_j) \tag{17}$$

Now,

$$\begin{aligned}
Pr(b^t_j|s^t, \hat{\theta}^t_j, m^t_{i,r}, a^{t-1}, m^{t-1}_{i,s}, is^{t-1}) &= \sum_{o^t_j} Pr(b^t_j|s^t, \hat{\theta}^t_j, a^{t-1}, is^{t-1}, m^{t-1}_{i,s}, o^t_j, m^t_{i,r}) Pr(o^t_j|s^t, \hat{\theta}^t_j, a^{t-1}, is^{t-1}, m^{t-1}_{i,s}, m^t_{i,r}) \\
&= \sum_{o^t_j} Pr(b^t_j|s^t, \hat{\theta}^t_j, a^{t-1}, is^{t-1}, m^{t-1}_{i,s}, o^t_j, m^t_{i,r}) Pr(o^t_j|s^t, \hat{\theta}^t_j, a^{t-1}, m^{t-1}_{i,s}, m^t_{i,r}) \\
&= \sum_{o^t_j} \tau_{\theta^t_j}(b^{t-1}_j, a^{t-1}_j, o^t_j, b^t_j, m^{t-1}_{i,s}, m^t_{i,r}) O_j(s_t, a^{t-1}, o^t_j)
\end{aligned}$$

$$Pr(b^t_j|s^t, \hat{\theta}^t_j, m^t_{i,r}, a^{t-1}, m^{t-1}_{i,s}, is^{t-1}) = \sum_{o^t_j} \tau_{\theta^t_j}(b^{t-1}_j, a^{t-1}_j, o^t_j, b^t_j, m^{t-1}_{i,s}, m^t_{i,r}) O_j(s_t, a^{t-1}, o^t_j) \tag{18}$$

from eqn 1, 2 and 3,

$$b_i^t(is^t) = \alpha \sum_{is^{t-1}} b_i^{t-1}(is^{t-1}) \sum_{a_j^{t-1}} Pr(a_j^{t-1}|\theta_j^{t-1}) O_i(s^t, a^{t-1}, o_i^t) Pr(m_{i,r}^t|\theta_j^{t-1}, a_j^{t-1}) I(\hat{\theta}_j^{t-1}, \hat{\theta}_j^t) T_i(s^{t-1}, a^{t-1}, s^t)$$
$$\times \sum_{o_j^t} \tau_{\theta_j^t}(b_j^{t-1}, a_j^{t-1}, m_{j,s}^{t-1}, o_j^t, m_{j,r}^t, b_j^t) O_j(s^t, a^{t-1}, o_j^t)$$