Journal of Artificial Intelligence Research 68 (2020) 753-776 Submitted 02/2020; published 08/2020

How to Do Things with Words: A Bayesian Approach

Piotr Gmytrasiewicz PIOTR@UIC.EDU
Computer Science Department and AI Laboratory,
University of Illinois at Chicago,
Chicago, Illinois, 60607, USA

Abstract

Communication changes the beliefs of the listener and of the speaker. The value of a communicative act stems from the valuable belief states which result from this act. To model this we build on the Interactive POMDP (IPOMDP) framework, which extends POMDPs to allow agents to model others in multi-agent settings, and we include communication that can take place between the agents to formulate Communicative IPOMDPs (CIPOMDPs). We treat communication as a type of action and therefore, decisions regarding communicative acts are based on decision-theoretic planning using the Bellman optimality principle and value iteration, just as they are for all other rational actions. As in any form of planning, the results of actions need to be precisely specified. We use the Bayes' theorem to derive how agents update their beliefs in CIPOMDPs; updates are due to agents' actions, observations, messages they send to other agents, and messages they receive from others. The Bayesian decision-theoretic approach frees us from the commonly made assumption of cooperative discourse – we consider agents which are free to be dishonest while communicating and are guided only by their selfish rationality. We use a simple Tiger game to illustrate the belief update, and to show that the ability to rationally communicate allows agents to improve efficiency of their interactions.

1. Introduction

The idea of interacting and communicating with machines like we do with people is a very appealing one for AI and is important for its applications. Our work is predicated on the idea that interaction in general, and communication in particular, should be approached using existing Bayesian and decision-theoretic techniques, allowing, say, autonomous vehicles to navigate a partially observable and non-deterministic environment. The reason these techniques are applicable is that interactive settings are also non-deterministic and partially observable. Clearly, a unified framework applicable to both single-agent as well as multi-agent environments is desirable to enable an AI system to seamlessly operate in both settings.

Consider what would happen if, before you left your hotel in, say, Palo Alto in the morning, your friend Joe checked the weather and told you that it's 45 degrees F outside. Would you trust/believe him? What, precisely, happens to the state of the world, your state of knowledge about the state of the world, and what you know about what Joe knows as the result of Joe's announcement? How does your prior knowledge about the temperature (say that likely it's 60, plus minus 5, degrees F) factor into this? What if you think Joe's weather app may be set for Chicago not Palo Alto? What if you see Joe putting on his shorts and t-shirt? 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 issues?

Our ability to answer the questions above is crucial to creating robust AI systems that can interact and communicate with humans and other systems. Figure 1 depicts the summary of the

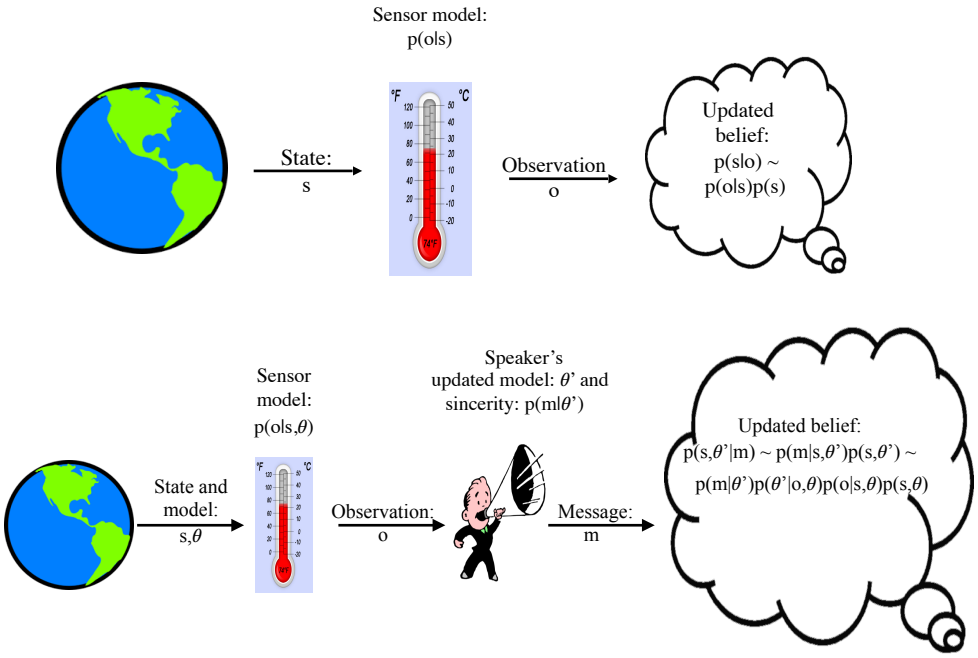

Figure 1: Top: Schematic representation of Bayesian update due to imperfect observation. Updated belief is proportional to the product of the prior belief in a state, $p(s)$, and the sensor (accuracy) model, $p(o \mid s)$. Better sensor accuracy results in a more informed posterior belief about the state.
Bottom: Simplified representation of Bayesian update due to communication. The listener has a model of the speaker, θ, and a prior over the state and θ, $p(s, \theta)$. The listener models the accuracy of the speaker's sensor, $p(o \mid s, \theta)$, the speakers' belief update due to observation, $p(\theta' \mid o, \theta)$, and the dependence, $p(m \mid \theta')$, between the speaker's updated belief (included in θ') and the message produced, quantifying the speaker's sincerity.

Bayesian update of an agent's belief due to observation and due to communication. In both cases, an agent's belief about the state of the world is updated based on the model of how well the state "explains" what is observed and heard. In the upper half of Figure 1, the model of the accuracy of the sensor is used to update an agent's prior belief, $p(s)$, about the state of the world using Bayes' theorem. In the case of communication, in the lower half of Figure 1, a model, θ, of the speaker is needed because it is the speaker who generates the message, m. The update, in this simple case, multiplies the prior, $p(\theta, s)$, the model of the speaker's sensor, $p(o \mid s, \theta)$, the relationship between the speaker's updated belief (included in θ') and the observation, $p(\theta' \mid o, \theta)$, and the dependence of the message, m, generated by the speaker and the updated model, $p(m \mid \theta')$. Below we derive the dependence between m and θ', which quantifies sincerity, by modeling the speaker as rational.

Belief update is a key notion allowing planning with communicative actions because it is the specification of the consequences of such actions. In Markov decision processes (MDPs) and their partially observable variant (POMDPs), consequences of physical actions are specified in the transition function, T, which designates the possible and likely states resulting from any action in any initial state (Kaelbling, Littman, & Cassandra, 1998; Russell & Norvig, 2010; Smallwood & Sondik, 1973). Interestingly, while the actions are physical, the planning process in POMDPs, called value iteration, takes place in the space of the agent's beliefs (Kaelbling et al., 1998; Russell & Norvig, 2010). Our paper extends this approach and reports on the Bayesian approach to specifying the consequences of, and planning for, communicative acts interwoven with physical actions and observations during an interaction with other agents and the physical world. Planning is also based on value iteration and it takes place in the space of the agent's *interactive* beliefs (Aumann, 1999; Doshi & Gmytrasiewicz, 2009; Gmytrasiewicz & Doshi, 2005), which are beliefs about the world and about other agents, including their beliefs. Communicative actions are unlike physical actions since the function of communication is to change the agents' beliefs (Goodman & Frank, 2016; Goodman & Lassiter, 2014; Lewis, 1979; Stalnaker, 1978), not to change the physical environment. On one hand communication is action (Austin, 1962) executed by a speaker, but on the other hand, it is perception for the listener. Like beliefs in POMDPs, interactive beliefs acquire value because they enable beneficial physical actions. Communicative acts, in turn, are valuable because they lead to valuable state of interactive beliefs which, for example, facilitate beneficial coordinated action.

Our approach builds on previous work on POMDPs (Kaelbling et al., 1998; Russell & Norvig, 2010; Smallwood & Sondik, 1973) and IPOMDPs (Doshi & Gmytrasiewicz, 2009; Gmytrasiewicz & Doshi, 2005). When it comes to communication, we build on classical work of Austin "How to Do Things with Words" (Austin, 1962), on approaches that make a cooperative assumption, like Grice's maxim of quality stating that everything that is said is true and can be believed (Grice, 1975), and on the game-theoretic and Bayesian approaches to pragmatics (Goodman & Lassiter, 2014; Goodman & Frank, 2016; Pietarinen, 2007; Vogel, Potts, & Jurafsky, 2013; Zeevat, 2015) which analyze the interests of the sender of the information to interpret it – we follow this approach here. Communication and action among fully cooperative agents in DEC-POMDPs was also investigated but in that framework agents are assumed to communicate their most recent observations, so the issue of pragmatics does not arise (Olihoek, Spaan, & Vlassis, 2007; Pynadath & Tambe, 2002; Wu, Zilbersein, & Chen, 2011). We consider a more general case of agents being uncertain about the reward functions of other agents. As we show (see Figure 1), agents having explicit models of others is necessary for Bayesian approach to interpreting communication (i.e., its pragmatics) and actions of others.

We propose a principled approach to interaction and communication based on Bayesian decision theory and decision-theoretic planning. We build on interactive POMDPs (Doshi & Gmytrasiewicz, 2009; Gmytrasiewicz & Doshi, 2005) 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 been called the theory of mind. Its usefulness while interacting with others has long been established in psychology, linguistics, philosophy, economics and AI (Albrecht & Stone, 2018; Aumann, 1999; Brandenburger & Dekel, 1993; Dennett, 1986; Frith & Frith, 2005; Gallese & Goldman, 1998; Gmytrasiewicz & Doshi, 2005; Grice, 1975; Leslie, Friedman, & German, 2004; Shanton & Goldman, 2010; Narayanan, 1988; Pietarinen, 2007; Stahl & Wilson, 1994, 1995; Vogel et al., 2013). We limit our attention to intentional models of other agents that represent agents' preferences and beliefs about the world and about other agents' beliefs; i.e., interactive beliefs

(Aumann, 1999; Brandenburger & Dekel, 1993; Doshi & Gmytrasiewicz, 2009; Gmytrasiewicz & Doshi, 2005). 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 others terminate at or below 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 are free to choose one as needed. A good illustration of this is the "Cheryl's Birthday" problem[1]. It reads as follows:

> Albert and Bernard just became friends with Cheryl, and they want to know when her birthday is. Cheryl gives them a list of 10 possible dates: May 15 , 16 or 19, June 17 or 18, July 14 or 16, and August 14, 15 or 17. Cheryl then tells Albert and Bernard separately the month and the day of her birthday respectively. Then Albert says: "I don't know when Cheryl's birthday is, but I know that Bernard doesn't know too". Then Bernard: "At first I didn't know when Cheryl's birthday is, but I know now." Now Albert states: "Then I also know when Cheryl's birthday is." So when is Cheryl's birthday?

To solve the puzzle, the audience needs to create models of participants, Albert and Bernard, and their models of each other (and of Cheryl) to interpret the information shared about what is known, and not known, to participants, and to figure out Cheryl's birthday. But if only Cheryl was less opaque and simply informed us (the audience) of her birthday directly, the nested models of Albert and Bernard would not be needed.[2]

In CIPOMDPs, an agent may have many different possible models of others, and of their theories of mind, nested to different levels (all equal or less than the agent's strategy level). This allows an agent to model other agents' possibly different levels of sophistication, represented by their own strategy levels. No agent is able to model other agents as more sophisticated than itself. Our assumption that the agent's theories of mind are finitely-nested ensures that their models are computable; i.e., that they generate a prediction as to the other agents' actions in finite time. Infinitely-nested theories of mind are of central importance in game theory (Binmore, 1990; Fudenberg & Tirole, 1991) and in epistemic game theory (Perea, 2012). Infinitely-nested models are, in special cases, solvable using Nash equilibrium analysis (Fudenberg & Tirole, 1991). Infinitely-nested models express the state of agents' belief (or knowledge) called common belief (knowledge) about each others' beliefs, preferences, and rationality. It has been argued that these epistemic states are not practically achievable (Halpern & Moses, 1990).

We use Bayes' update to describe the results of agents' communicative and physical actions on agents' beliefs. This formalizes game-theoretic pragmatics (see Franke and Jager (2016), Vogel et al. (2013), Pietarinen (2007) and references therein) and implements a Bayesian approach to language pragmatics (Zeevat, 2015); but see Goodman and Lassiter (2014) for an alternative approach. As we show further below, the update of agents' beliefs due to communication, action, and observation recursively descends l levels of theories of mind down to "flat" POMDP models where, we propose, the communicative acts are processed according to their literal meaning.

1. See `https://en.wikipedia.org/wiki/Cheryl's_Birthday` (search for "Cheryl's Birthday" on Wikipedia in case link is outdated).
2. To solve this particular puzzle it is sufficient to consider just the sets of possibilities, also called information sets, as opposed to also considering agents' probability distributions over the possible states, i.e., their beliefs.

When it comes to communication in particular (see Figure 1), the Bayesian update combines the listener's prior information with what the listener knows about the speaker's sincerity (i.e., whether the content of the message reflects the speaker's beliefs), and with what the listener knows about how informed the speaker is (i.e., whether the speaker's beliefs reflect reality). The most challenging part is handling the speaker's sincerity, quantified as $p(m \mid \theta)$. Our computation of $p(m \mid \theta)$ follows the rational speech act paradigm (Frank & Goodman, 2012; Goodman & Lassiter, 2014; Goodman & Frank, 2016) according to which speakers choose speech acts that maximize their expected utility. We identify the model, θ, as a CIPOMDP. We show that Bayes update frees us from the cooperative assumption (Grice, 1975) that agents are sincere and believe what they hear; we elaborate on this further below.[3]

The Bayesian update illustrated in Figure 1 is drastically simplified. The actual belief update, which we derive below, accounts for changes of an agent's beliefs not only due to a message it got from another agent but also due to its own observation, physical action, and a message it sent. All of these update what the agent knows about the world, about other agents, and about their state of knowledge. Although it complicates things, it is important to include physical actions and agents' rewards in this analysis. They get at the crucial issue of the motivation of the sender of the message, and allow us to quantify the value of communicative acts as enabling coordination of physical actions and resulting desirable changes in the physical environment. Treating communicative actions in isolation would not enable us to model these aspects and the analysis would be incomplete.

We use the tiger game (Kaelbling et al., 1998) below to show an example of the Bayesian update due to communication. Our analysis reveals that if the listener knows the speaker is sincere and knows the speakers' observation capabilities, then agents communicating their beliefs may be equivalent to them sharing their observations (Nair, Pynadath, Yokoo, Tambe, & Marsella, 2004). A related approach to ours that uses communicative interactive decision diagrams (Tian, Luo, & Huang, 2013; Tian, Luo, Zeng, & Wu, 2016) also assumes that agents' observations are exchanged, but uses the cooperative assumption and does not include planning for communicative actions. In a similar vein (Zhou & Luo, 2012) define a communication model based on IPOMDPs, Com-I-POMDP, but they do not model the communicative act as dictated by rationality of the speaker, as in our approach here.

Finally, we describe an approach to decision-theoretic planning, i.e., forward search through the space of interactive beliefs agents can use to plan the most advantageous communicative and physical action sequences. Planning is based on the Bellman optimality principle and is analogous to value iteration in POMDPs (Kaelbling et al., 1998; Russell & Norvig, 2010; Smallwood & Sondik, 1973) and IPOMDPs (Doshi & Gmytrasiewicz, 2009; Gmytrasiewicz & Doshi, 2005). Decision-theoretic planning backs up values of future reachable states while the agents optimize by making their utility-maximizing choices. As we mentioned, the agents' capability to examine possible futures during planning is important because it shows how physical and communicative actions can bring about valuable states of agents' interactive knowledge and coordinate physical actions.[4]

3. In case of Cheryl's Birthday the communications from Albert and Bernard were assumed to be sincere. Bayesian update then simply eliminates birth dates previously considered possible which the communicating participants declare to be impossible.

4. Another family of decision-theoretic techniques applied to dialog planning is also based on POMDPs (Young, Gasic, Thomson, & Williams, 2013) but it differs from our approach in that it does not model the interlocutor's beliefs and rationality.

Careful planning can also shield agents from being taken advantage of by insincere speakers. Consider the previous example in which Joe pranks us by lying about the temperature in Palo Alto this morning and gets us to come out of the hotel dressed in a warm coat and a hat to 60 degrees F. Joe may find it hilarious and get a short term reward, but a more distant future may include the possibility of 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 their opponent for defection (Axelrod, 1984). On the other hand, if we know that we will not have a chance to get Joe back for lying we may want to 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 Konnikova (2016), Robinson (2010) for some related reading, Jehiel (2006) for a game-theoretic approach, Whaley (2016) for a military perspective, and Isaac and Bridwell (2017) for ethical and philosophical analysis. Investigating the interaction between rationality and dishonesty opens up a host of complex and nuanced issues, but we leave a more thorough investigation of our framework's applicability to deception for future work.

Decision-theoretic planning in CIPOMDPs inherits a number of important properties from POMDPs: The belief (i.e., probability distribution) over the interactive state space is a sufficient statistic for all histories of actions, observations and messages, the value functions are piece-wise linear and convex in the agents' interactive beliefs, the Bellman backup operation is a contraction, and the value iteration converges. We show these properties formally below.

2. Interactive POMDPs

We briefly describe interactive POMDPs first since our approach to communication builds on them. IPOMDPs extend POMDPs to interactive settings (Doshi & Gmytrasiewicz, 2009; Gmytrasiewicz & Doshi, 2005). We consider only two agents, i and j, for simplicity (the framework generalizes to many agents). A finitely-nested interactive POMDP for agent i is defined as the following 6-tuple:

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

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 of agent i. Here we consider only one class of models which are the kth (less than l) level *intentional, i.e., rational* models of agent j: $\theta_{j,k} = \langle b_{j,k}, A_j, \Omega_j, T_j, O_j, R_j \rangle$[5], where $b_{j,k}$ is agent j's belief nested to the level k: $b_{j,k} \in \Delta(IS_{j,k})$.[6] The intentional model, $\theta_{j,k}$, also called an agent's *type* (Harsanyi, 1967), 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. Note that agent i may have many models of agent j, each nested down to a different level but always lower than l. $IS_{i,l}$ can be defined inductively:

5. We neglect the optimality criterion for simplicity.
6. $\Delta(S)$ is a set of probability distributions over set S.

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

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[7], $O_i : S \times A \times \Omega_i \rightarrow [0, 1]$ is the observation function[8], $R_i : S \times A \rightarrow R$ is the reward function.

Given all the definitions above, the IPOMDP belief update, derived in (Gmytrasiewicz & Doshi, 2005), is:

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

Unlike plain belief update in POMDPs, 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, i, needs to update its beliefs about the models of j 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, is an element of the set of optimal actions, $OPT(\theta_i)$, defined as:

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

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 $\Sigma_{a_j} R_i(is, a_i, a_j) P(a_j \mid \theta_j)$. The utility (value), $U(b_i)$,

7. Note that actions are assumed not to influence agents' models directly. This is called model non-manipulability assumption (Gmytrasiewicz & Doshi, 2005).
8. Note the assumption that agent cannot observe other agents' models directly, called model unobservability assumption (Gmytrasiewicz & Doshi, 2005).

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 Equation (4) 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[9], and is a decision-theoretic version of epistemic planning (Bolander & Andersen, 2011; Engesser, Bolander, Mattmüller, & Nebel, 2017; Hoek & Wooldridge, 2002; Kominis & Geffner, 2015). The 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 (i.e., without arg max) is the expected utility of this action to agent i, $U_{\theta_i}(a_i)$. Equation (4), when applied to other agents, allows agents to predict which actions other agents are likely to execute. Predicting other agents' actions is of course the main objective of modeling them, and it enters into belief update in Equation (3). According to the hard maximization criterion, agent i could model agent j as a strict optimizer and predict that j could execute only actions in $OPT(\theta_j)$ with probability $P(a_j \mid \theta_j) = \frac{1}{|OPT(\theta_j)|}$. A more relaxed criterion is soft maximization (McKelvey & Palfrey, 1995), for which the probability of j executing a_j is:

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

where λ is the rationality parameter. When λ is 0 the choice is random, and when λ is infinity the soft max criterion becomes hard maximization.[10] Note that computing others' expected actions according to the soft maximization above could be more complex than that involved in hard maximization because one needs to compute the expected utilities, and probabilities, associated with all of agent j's actions, not only with the ones in the optimal set $OPT(\theta_j)$.

The advantage of modeling other agents as rational in IPOMDPs lies in that it allows one to derive their expected actions and coordinate with them better. It can be quantified as an increase in the value functions obtained in IPOMDPs compared to POMDPs that treat other agents as noise (Gmytrasiewicz & Doshi, 2005). That increase in performance is obtained at the computational cost of modeling others. When there are no other agents to model, IPOMDPs reduce to POMDPs.

The principle of modeling other agents as rational optimizers is central to Bayesian pragmatics of communicative behavior based on the Rational Speech Acts model (Goodman & Frank, 2016; Goodman & Lassiter, 2014). We use it in our formalism below.

3. Communicative IPOMDPs

Communicative IPOMDPs (CIPOMDPs) build on IPOMDPs but include additionally a communication language, an action of sending a message, m_s, and an additional observation - a message that could be received, m_r. Either message can be nil. A CIPOMDP for agent i is defined as a 7-tuple:

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

9. Analogously to search through physical state space for MDPs, and through single agent belief space in POMDPs.

10. Although λ is not part of the other agent's model and is not updated using Bayes' theorem, we think it could be. For example, λ could be interpreted as the agent's computational capability, included in a model of this agent and updated given what is observed about the agent's behavior. We leave this issue as an avenue for future research.

where $\mathbb{M}_i$ is a set of messages agent i can send to and receive from other(s) (both m_s and m_r above are in $\mathbb{M}_i$). All of the other elements are as defined analogously to IPOMDPs except that agent models also include appropriately subscripted $\mathbb{M}$, and the reward function has an additional argument; $R_i : S \times A_i \times \mathbb{M}_i \rightarrow R$, so it can also depend on the messages i sends (messages can be costly.) We assume that $\mathbb{M}$'s for different agents overlap and the common set of messages constitute the language of communication, called $\mathbb{M}$, which the agents share. We leave the exact specification of $\mathbb{M}$ for future work but we make an assumption that each message in $\mathbb{M}$ 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 its veracity. Note that while $\mathbb{M}$ 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. Thus, we relax the principle of cooperative discourse, i.e., the maxim of quality (Grice, 1975), but model the agents' ability to discount the information they receive from others by modeling that it may be in their interest to be dishonest.

Also note that agents' interactive beliefs are quite rich: they contain variables describing the state of the physical world, others' beliefs, actions of other agents, and others' beliefs about others, including their actions. Further, since agents perform planning and project their interactive belief states into the future they can exchange messages about what they believe about the future, others' future beliefs, their beliefs about others' intended future actions, etc. We leave for further research messages containing beliefs about the past, queries, imperatives and other types of communicative acts (Searle, 1975).

3.1 Belief Update

We now move on to the Bayesian belief update during interaction and communication. For simplicity we consider two agents i and j and drop the variables related to nesting levels of models. Agents are assumed to operate in a perceive-think-act loop (Russell & Norvig, 2010) and, as they proceed through time, accumulate information from all of their observation, message, and action histories in their updated beliefs about the world and about other agents. At every time step, agents observe and receive messages, decide what to do, and then execute actions and send messages to others. Call the message i sent at time $t-1$, $m_{i,s}^{t-1}$, and the message i received at time t, $m_{i,r}^{t}$. All messages are in $\mathbb{M}$.

The update below is analogous to the belief update in IPOMDPs (Equation (3)), which updates the probability of interactive states given the previous action and current observation: $P(is^t \mid b_i^{t-1}, a_i^{t-1}, o_i^t)$. The belief update in CIPOMDPs has to update the probability of the interactive state also due the message sent at the previous time step and received at the current time: $P(is^t \mid b_i^{t-1}, a_i^{t-1}, m_{i,s}^{t-1}, o_i^t, m_{i,r}^t)$:

Proposition 1 *Belief Update:*

$$b_i^t(is^t) = P(is^t \mid b_i^{t-1}, a_i^{t-1}, m_{i,s}^{t-1}, o_i^t, m_{i,r}^t) = \alpha \sum_{is^{t-1}:\hat{\theta}_j^{t-1}=\hat{\theta}_j^t} b_i^{t-1}(is^{t-1}) \times$$
$$\sum_{a_j^{t-1}} Pr(m_{i,r}^t, a_j^{t-1} \mid \theta_j^{t-1}) O_i(s^t, a^{t-1}, o_i^t) T_i(s^{t-1}, a^{t-1}, s^t) \times \quad (7)$$
$$\sum_{o_j^t} \tau_{\theta_j^t}(b_j^{t-1}, a_j^{t-1}, m_{i,r}^t, o_j^t, m_{i,s}^{t-1}, b_j^t) O_j(s^t, a^{t-1}, o_j^t)$$

Proof of Proposition 1 appears in the Appendix. The above update treats messages sent as actions and messages received as additional observations (but containing probability distributions). 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 would have to maintain probabilities describing accuracy of transmission: $P(m_{j,s}^{t-1} \mid m_{i,r}^t)$ (probability of what j sent given what i received) and $P_i(m_{j,r}^t \mid m_{i,s}^{t-1})$ (probability of what j received given what i sent) for all possible $m_{j,s}^{t-1}$ and $m_{j,r}^{,t}$.

As we mentioned, the update in Proposition 1 is analogous to the belief update in IPOMDPs in Equation (7) when it comes to actions and observations. With respect to communication, Proposition 1 illustrates the difference between sending messages and executing physical actions: While both are actions, the message sent, $m_{i,s}^{t-1}$ does not participate in changes to the physical state – it only enters into the belief update for the other agent, $\tau_{\theta_j^t}$. As we mentioned before (see Figure 1), the update combines three elements. First, the updated belief depends on the agent's prior belief $b_i^{t-1}(is^{t-1})$. Second, the term $P(m_{i,r}^t, a_j^{t-1} \mid \theta_j^{t-1})$ is equal to $P(m_{j,s}^{t-1}, a_j^{t-1} \mid \theta_j^{t-1})$ due to perfect transmission. It 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[11]. We assume that agents are sincere to the extent that it pays off for them; we define this further below in Equation (11). Third, Proposition 1 includes the dependence of agent j's belief and the state of the world included in the interactive state $is = \langle s, \theta_j \rangle$, both at time t and $t-1$. More explicitly, the joint probability of a particular state s and θ_j, $P(\theta_j \mid s)P(s)$, represents the degree to which i estimates that j is informed about the true state of the world, as in Figure 1. i maintains its estimate of j's beliefs and their update through the τ_{θ_j} term; if j's observation function, contained in θ_j, is accurate, i would expect that j's beliefs accurately reflect the state of the world, s.

In summary, Proposition 1 generalizes the 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.

Proposition 1 also 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. In (Gmytrasiewicz & Doshi, 2005) this assumption is called model non-manipulability. It can be relaxed to allow agents to modify other agents' observation functions or action capabilities by, say, disabling their sensors or actuators.

11. The message j sent may also depend on j's action a_j^{t-1}.

We provide an example of Bayesian update in a multi-agent version of the well-known Tiger game environment (Kaelbling et al., 1998) below.

Proposition 2 *Sufficiency:*
In a Communicative IPOMDP of agent i, i's belief, i.e., the probability distribution over the set $S \times \Theta_j$, is a sufficient statistic for the past history of i's observations.

Proposition 2 is a consequence of the Bayesian belief update, in Proposition 1, analogously to POMDPs (Smallwood & Sondik, 1973) and IPOMDPs (Gmytrasiewicz & Doshi, 2005). It states that histories of agent's observations, actions, and messages sent and received, no matter how long, can be compactly summarized as a probability distribution over $S \times \Theta_j$. The proposition holds since the belief at any time, t, depends only on the agent's previous belief, action and message sent, and current observation and message received. This means that the information contained in any history is accumulated in the current belief during each belief update taking place in the perceive-think-act loop (Gmytrasiewicz & Doshi, 2005; Kaelbling et al., 1998; Russell & Norvig, 2010). The sufficient statistic over the interactive state space is analogous to one in DEC-POMDPs (Olihoek, 2013) defined over the agents' joint observation-action histories.

3.2 Decision-Theoretic Planning for Communication and Interaction

Given that belief update in CIPOMDPs[12] is analogous to belief update in IPOMDPs, we similarly proceed to define the Bellman equation which includes communicative actions and hard and soft maximization criteria quantifying speaker's sincerity.

The belief update defined in Proposition 1 can again be represented as a function SE so that the new belief is: $b_i^t = SE(b_i^{t-1}, a_i^{t-1}, m_{i,s}^{t-1}, o_i^t, m_{i,r}^t)$. Then $\tau_{\theta_i}(b_i^{t-1}, a_i^{t-1}, m_{i,s}^{t-1}, o_i^t, m_{i,r}^t, b_i^t)$ is defined as 1 when b_i^t is equal to $SE(b_i^{t-1}, a_i^{t-1}, m_{i,s}^{t-1}, o_i^t, m_{i,r}^t)$ and 0 otherwise, and analogously for j – it is a factor in Proposition 1 above.

Now the utility of an interactive belief of agent i, contained in i's type θ_i, can be defined analogously to POMDPs and IPOMDPs:

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

$ER_i(is, m_{i,s}, a_i)$ 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 $\Sigma_{a_j} R_i(is, a_i, a_j, m_{i,s}) P(a_j \mid \theta_j)$, ($i$'s reward can depend on the cost of sending $m_{i,s}$, as we mentioned before.) The term $P(m_{i,r}, o_i \mid b_i, a_i)$ needs to be computed by considering that i's observation and message it receives depend on actions and messages sent by agent j:

$$P(m_{i,r}, o_i \mid b_i, a_i) = \sum_{is} b(is) \sum_{a_j} O_i(s, a_i, a_j, o_i) Pr(m_{i,r}, a_j \mid \theta_j) \tag{9}$$

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

and where s and θ_j on the right-hand side are elements of is, and $Pr(m_{i,r}, a_j \mid \theta_j)$ is given below in Equation (11).[13]

Equation (8) defines the utility of an interactive belief, b_i, in θ_i, and is the Bellman optimality equation for interactive (physical and communicative) behavior. As we mentioned, the agent's ability to compute the optimal utility maximizing interactive behavior is the basis for rational interaction. The Bellman equation describes the back-up operation during value-driven search (i.e., planning) 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 (Bolander & Andersen, 2011; Engesser et al., 2017; Hoek & Wooldridge, 2002; Kominis & Geffner, 2015). As we noted, the agents' ability to consider possible future interactions is crucial in guarding against exploitation by insincere speakers.

An optimal message-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 pairs, $OPT(\theta_i)$, defined as:

$$OPT(\theta_i) = \arg\max_{(m_{i,s},a_i)} \Big\{ \sum_{is \in IS} b_{is}(s) ER_i(is, m_{i,s}, a_i) + \gamma \sum_{(m_{i,r},a_i)} P(m_{i,r}, o_i \mid b_i, a_i) U_i(\langle SE_{\theta_i}(b_i, a_i, m_{i,s}, o_i, m_{i,r}), \hat{\theta}_i \rangle) \Big\} \tag{10}$$

The right-hand side of the Equation (10) for any particular message-action pair $(m_{i,s}, a_i)$ is $i's$ expected utility, $U_{\theta_i}(m_{i,s}, a_i)$, of sending $m_{i,s}$ and executing action a_i, and similarly for agent j.

The equation above, applied to other agent(s), allows one to predict which messages and actions they will execute since these actions are rational for them. As in the case of IPOMDPs, the hard maximization criterion, according to which agent i could model agent j as a strict optimizer, predicts that j would only perform interactive actions in $OPT(\theta_j)$, each with probability $P(m_{j,s}, a_j \mid \theta_j) = \frac{1}{|OPT(\theta_j)|}$. The soft maximization relaxes this prediction and defines the the probability of j sending $m_{j,s}$ and performing a_j as:

$$Pr(m_{j,s}, a_j \mid \theta_j) = \frac{exp[\lambda U_{\theta_j}(m_{j,s}, a_j)]}{\sum_{(m_{j,s},a_j)} exp[\lambda U_{\theta_j}((m_{j,s}, a_j)]} \tag{11}$$

As before, when λ is 0 the choice is random and when λ is infinity the soft maximization criterion becomes the hard max. Equation (11) treats agents as rational and, when it comes to communication, is central to the Rational Speech Acts model (Goodman & Frank, 2016; Goodman & Lassiter, 2014). Equation (11) predicts that rational agents are more likely to perform actions that maximize their utility. As such, it quantifies an agent's sincerity by tying the message it decides to send to its beliefs (contained in θ_j) by modeling it as a self-interested rational speaker. But, a speaker may be insincere and intend to deceive to further its interests. As we mentioned, agents and AI systems 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. We leave detailed investigation of this topic for future work.

Equation (8) defines value iteration for CIPOMDPs. It turns out that CIPOMDPs inherit some desirable properties from IPOMDPs and POMDPs. The following two theorems establish this:

13. Recall our assumption that $m_{i,r} = m_{j,s}$, i.e., i receives what j sends.

Theorem 1 (Convergence of Value Iteration):
For any finitely-nested CIPOMDP, the value iteration algorithm starting from any well-defined value function converges to a unique fixed-point.

The outline of the proof is in the Appendix. Analogously to POMDPs and IPOMDPs, the above theorem makes use of the contraction property of the back-up operation defined in Equation (8). As is the case for POMDPs (Russell & Norvig, 2010) and IPOMDPs (Gmytrasiewicz & Doshi, 2005), the error in the iterative estimates, U^n, of the value of agent's belief contained in model θ_i, compared to the optimal value for infinite time horizon, U^*, i.e., $||U^n - U^*||$, is reduced by a factor at least γ on each iteration. Consequently, the number of iterations needed to reach an error of at most ϵ is:

$$N = \lceil \log(2R_{max}/\epsilon(1-\gamma))/\log(1/\gamma) \rceil \tag{12}$$

where R_{max} is the upper bound of the reward function.

Theorem 2 (PLWC of Value Function):
For any finitely-nested CIPOMDP, U is piece-wise linear and convex.

The proof of Theorem 2 is outlined in the Appendix; it is analogous to one due to Smallwood and Sondik (Smallwood & Sondik, 1973) for POMDPs and proceeds by induction. Establishing this property allows us to decompose the CIPOMDP value function into a set of alpha vectors, each of which represents a policy tree. The PWLC property enables us to work with sets of alpha vectors rather than perform value iteration in the agent's continuous belief space. The difference in CIPOMDPs, when compared to POMDPs, is that agent's beliefs are over physical states of the world as well as over the possible models of the other agents.

3.3 Literal Meaning

Given that the interactive state space, IS, contains nested models of agents, the Bayesian update in Proposition 1 descends down all of the l levels of nesting of the theories of mind the agent has in alternating invocations of the agents' belief update, τ. These simulations are characteristic of game-theoretic pragmatics (Jaeger, 2011). The recursion terminates 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** (Frank & Goodman, 2012; Franke & Jager, 2016). Our approach builds on this previous work but it also leaves a number of unanswered questions which we discuss briefly below.

We assume that an agent that does not model any 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 broadcast to "no one in particular" (NOIP) – because the speaker does not model any other agent. Similarly, a *literal listener*, upon receiving a message from NOIP, can incorporate the content of the message into its beliefs about the physical state in a way that does not use Bayes' update. 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 obtained in the past (Chandrasekaran, Doshi, Zeng, and Chen (2017) use a closely related technique to allow for coordinated action in ad-hoc teams).

To model this 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 the physical state, with probability α. In this case α is a single number in $[0, 1]$ and the message generated is 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 Flower, Gribble, and Ridley (2014) for an example of insincere drongo birds' 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.

The specification and update of α and β parameters based on reinforcement learning is left for future work – a promising approach is described in (Chandrasekaran et al., 2017). The generation and belief update due to communication for agents that do not model others we propose here is clearly not Bayesian decision-theoretic. It takes us into an uncharted territory of combining a Bayesian and non-Bayesian belief updates which we will investigate further in our future work.

3.4 Complexity

The complexity of solving a finitely-nested CIPOMDP is related to the complexity of IPOMDPs (Gmytrasiewicz & Doshi, 2005) and POMDPs (Papadimitriou & Tsitsiklis, 1987). Given agents' communication language, $\mathbb{M}$, CIPOMDPs have an effective action space of $A \times \mathbb{M}$ and so cannot be any easier to solve than IPOMDPs. To solve a finitely-nested IPOMDP with a number of models of other agent(s) bounded by a number M requires solution of M^l POMDPs. Thus, solving CIPOMDPs is PSPACE-hard for finite time horizons and undecidable for infinite horizons, just like for POMDPs.

The space of messages $\mathbb{M}$ may be continuous, as could be the space of actions A. Approximations used in this case are described in (Seiler, Kurniawati, & Singh, 2015; Thrun, Burgard, & Fox, 2005) – these are also applicable to cases which include communication.

4. Example

We consider a simple cooperative interaction between two agents engaged in a version of a multi-agent Tiger game (Kaelbling et al., 1998; Nair, Roth, Yokoo, & Tambe, 2003) 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

$\langle a_i, a_j \rangle$	TL	TR
OR,OR	20	-50
OL,OL	-50	20
OR,OL	-100	-100
OL,OR	-100	-100
L,L	-2	-2
L, OR	9	-101
OR, L	9	-101
L, OL	-101	9
OL, L	-101	9

Table 1: Reward Function, R_F, for Cooperative Multi-agent Tiger Game (Nair et al., 2003). The agent's payoffs depend on the states, TL and TR, and on the actions of both agents.

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.[14] 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 it 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 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 miss-coordinate 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. In our example here we assume that it is known that the agents' reward function is R_F, and that the above parameters are present in all of the agents' intentional models, θ, for simplicity. In such case deception is ill advised; we discuss what happens when one relaxes this assumption at the end of this section.

To illustrate the CIPOMDP belief update and rational behavior in this setting we further assume the following: Both agents, i and j, start off with no information about the position of the tiger and the initial beliefs of agents assign probabilities of 0.5 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 from the point of view of agent i who has a strategy level of $l = 2$, i.e., i models j as an CIPOMDP with strategy level of $k = 1$. This means i models j modeling i as flat POMDP. As i listens, it will update its belief over its interactive state $IS_{i,2}$ in which initially two elements have non-zero probabilities: $(TL, \theta_{j,1})$ and $(TR, \theta_{j,1})$. In each of these, $j's$ model, $\theta_{j,1}$, is a CIPOMDP with interactive state space $IS_{j,1}$ with

14. In (Gmytrasiewicz & Doshi, 2005) agents can also hear the creak of an opened door but we neglect this here.

two elements having non-zero probabilities: $(TL, \theta_{i,0})$ and $(TR, \theta_{i,0})$. Here, j's model of i is a POMDP: $\theta_{i,0} = \langle (0.5, 0.5), A, \Omega, T, O, R_F \rangle$, where $(0.5, 0.5)$ is $i's$ belief over the tiger's position, and the other elements are as defined before. i's initial belief, b_i^0, over the two non-zero probability elements in $IS_{i,2}$ is $(0.5, 0.5)$ since i does not know the tiger's position and is certain about the model of agent j.

Note that if agent j was not present and i listened and received, say, growl from the left, GL', then its "flat" belief over $S = \{TL, TR\}$ would become $(0.6, 0.4)$.[15] Also, if a lone agent i would listen again and receive second growl from the left door, GL'', then its updated belief over $\{TL, TR\}$ would become $(0.69, 0.31)$.[16] Yet another growl from left at the following time step would lead i to believe that tiger is on the left with probability 0.77. Further updates due to subsequent GLs would increase this probability to 0.83, 0.88, 0.92, and 0.94; with the last value obtaining after seven consistent growls from left at time step $t = 8$. It is instructive to compare these probabilities with ones updated due to similar observations but additionally due to messages exchanged in Table 2.

Coming back to our scenario with both agents, say that initially both listen, since opening the doors given agents do not know anything about the tiger's location is ill advised, and that both agents do not send messages. Say that agent i received growl from the left, GL. i uses CIPOMDP belief update, and there are 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 over tiger's position; one, $(0.6, 0.4)$, computed by i simulating j's belief update, τ_{θ_j}, in Equation (7), corresponds to j having received GL, and the other corresponds 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's. Note that i's marginalized belief over physical states $\{TL, TR\}$ at $t = 1$ is $(0.6, 0.4)$ as should be expected[17]. This update gives the same result for IPOMDP as well as CIPOMDP belief updates because agents did not exchange messages.

Now we can consider what could happen further. First, say that at time $t = 2$ each agent performs another listen action. Let's suppose i again gets GL. Now, at time $t = 3$, there are six interactive states i assigns non-zero probabilities to: $\{(TL, \langle (0.69, 0.31), \hat{\theta_{j,1}} \rangle), (TL, \langle (0.5, 0.5), \hat{\theta_{j,1}} \rangle),$
$(TL, \langle (0.31, 0.69), \hat{\theta_{j,1}} \rangle), (TR, \langle (0.69, 0.31), \hat{\theta_{j,1}} \rangle), (TR, \langle (0.5, 0.5), \hat{\theta_{j,1}} \rangle), (TR, \langle (0.31, 0.69),$
$\hat{\theta_{j,1}} \rangle)\}$. The probabilities i assigns to these six interactive states turn out to be: $(0.25, 0.33, 0.11, 0.05,$ $0.15, 0.11)$ and i's marginalized belief over physical states, TL and TR, is $(0.69, 0.31)$. At this point ($t = 3$), i's optimal action would be to listen. i also knows j won't open the door, and i knows j knows i won't open the door because of two levels of nested modeling in this case.

We can now compare the above to a second extension of this scenario which includes communication. Say the agents not only perform listening actions at $t = 2$, but also send messages containing their true beliefs about tiger's location. So, at $t = 2$, i listens again but also sends message $m_{i,s}^2$ with the content "I believe that probability of tiger being on the left is 0.6"[18], and

15. The updated belief in TL is $p(TL \mid GL') = \frac{p(GL'|TL)p(TL)}{p(GL')} = \frac{0.6 \times 0.5}{0.5} = 0.6$.
16. $p(TL \mid GL', GL'') = \frac{p(GL''|TL)p(TL|GL')}{p(GL''|GL')} = \frac{0.6 \times 0.6}{0.6^2 + 0.4^2}$
17. We marginalized j's belief over i's possible beliefs for simplicity. If represented explicitly, i's belief would reveal that it knows that j considers both $(0.6, 0.4)$ and $(0.4, 0.6)$ as i's possible beliefs.
18. Recall that 0.6 is the probability of TL after the first observation of GL, so the message is truthful.

Time Step	Obs.	m_r	Updated Belief	Action	m_s
t = 1			0.5	L	NIL
t = 2	GL	NIL	0.6	L	"p(TL)=0.6"
t = 3	GL	"p(TL)=0.6"	0.77	L	"p(TL)=0.77"
t = 4	GL	"p(TL)=0.77"	0.88	L	"p(TL)=0.88"
t = 5	GL	"p(TL)=0.88"	0.94	OR	NIL

Table 2: The evolution of agent i's belief through time along one possible future. At each time step, agent i gets an observation, receives a message, m_r, updates its belief (only the marginalized probability of tiger being on the left, p(TL), is shown), executes an action, and sends a message. If it were not for the message exchange, it would take up to time step 8 for agent i to assign probability 0.94 to TL, assuming all seven observations were growl from the left, GL.

that at time $t = 3$, i receives $m^3_{i,r}$ with the identical content, which j sent at $t = 2$. Assuming i again gets growl left as observation at time $t = 3$, we use the CIPOMDP belief update equation to get the updated belief of i after the second observation and the message from j. The content of the message sent by a sincere speaker means that the term $P(m^3_{i,r}, L \mid \theta^2_j)$ in Equation (7) is 1 only for models of j in which $j's$ belief at time $t = 2$ is $(0.6, 0.4)$ and zero for others. After receiving the message and observing GL, the probabilities i assigns to four interactive states $\{(TL, \langle(0.77, 0.23), \hat{\theta_{j,1}}\rangle), (TL, \langle(0.6, 0.4), \hat{\theta_{j,1}}\rangle), (TR, \langle(0.77, 0.23), \hat{\theta_{j,1}}\rangle),$
$(TR, \langle(0.6, 0.4), \hat{\theta_{j,1}}\rangle)\}$ become $(0.46, 0.31, 0.09, 0.14)$. As expected, i's marginalized belief over physical states is $(0.77, 0.23)$.

Note that i does not interpret the message it received literally; i updates the probability it assigns to state TL based on hearing two growls from the left and based on message it got from j stating that j, at time $t = 1$, assigns probability 0.6 to TL. The result is i's belief in TL equal to 0.77. This is the same belief i would have if it were alone and listened at time $t = 2$ and got GL at time $t = 3$ as well, for three consecutive growls from the left. Thus, getting a message from j allowed i to update beliefs about the state of the physical world beyond what was contained in the observations it made – the shared information, in this simple case, is equivalent to an additional observation. Note also that at time $t = 3$, i is again uncertain about what j believes about the tiger position. i assigns the probability of 0.55 that j's belief in TL is 0.77 as well. But i considers it also possible that j assigns probability 0.6 to TL. This is due to j's second observation possibly being growl from the right (j's update would then be based on first GL, second GR, and the message j got from i.)

Let us now see what i knows about how j models i. Recall that i's model of j's model of i is a "flat" POMDP so we apply the literal listener model and the combination function C_β. We will assume that β parameter is 1 for simplicity. Thus, no matter what i thinks that j thinks about what i believes about the tiger location before i receives the message (from NOIP), i's updated belief will assign 0.6 probability to the tiger being on the left. Given these probabilities, j knows that i would not open any doors and would listen. Given this conclusion, i would know that j would not open any doors either. Hence the best option for i is also to listen at the time step $t = 2$.

The further evolution of agent i's beliefs along one possible sequence of observations, actions and messages is depicted in Table 2. It shows that i reaches a belief allowing it to open the right door at the fifth time step. This is a fairly profitable sequence, allowing i to avoid paying the costs

of listening, and expecting a positive reward associated with getting the gold at time $t = 5$. Without communication, i would have to wait until time $t = 8$ to reach the belief 0.94 that the tiger is located behind the left door, as we mentioned before.

The profitable possible future depicted in Table 2 imparts positive values to messages agents can exchange at times 2, 3, and 4, along that path as described by Bellman Equation (8). The value of opening the right door at time $t = 5$ can be computed from Table 1 as: $U(OR) = 20p(TL)p(a_j = OR) - 50p(TR)p(a_j = OR) - 100p(a_j = OL) + 9p(TL)p(a_j = L) - 101p(TR)p(a_j = L)$. This expression illustrates the fact that, in order to be optimal, agent i needs to keep track of both its belief in the tiger location and of the model of agent j and its anticipated actions. If i assigns sufficiently large probability to TL and sufficiently small probability that j will open the left door then i opens the right door. This happens at time step $t = 5$ in the exchange in Table 2. Thus, messages i receives from j not only allow i to become more certain about tiger location but also allows i to eliminate the possibility that j may miss-coordinate by opening the left door resulting in the negative payoff of -100.

Also, let us note that the Bayesian method we present automatically keeps track of dependencies between messages and their information content to avoid "double counting" of evidence. Consider the exchange of messages in Table 2. One may suspect that the information about the tiger location i obtained at time $t = 2$ due to its observation of tiger growl from the left, which is contained in a message sent to j at the same time step and used by j to update its beliefs at time $t = 3$, is counted again when i receives a message from j at time $t = 4$. However, this is not the case because the Bayesian update i uses naturally keeps track of j's beliefs and reasons for their update, and avoids unwarranted double counting.

Finally, we briefly consider what happens in the above example if sincerity were not assumed. Say that i suspects that j may have a payoff function different from the friendly one depicted in Table 1. i would then have an additional set of j's models containing a frame, say $\hat{\theta}'_{j,1}$, of j for which the sincerity term $P(m^3_{i,r}, L \mid \theta'_{j,1})$ in Equation (7) is **not** zero for other possible beliefs of j at time $t = 2$. So, it could be that $m^3_{i,r}$, stating that the probability of the tiger being on the left is 0.6 is an attempt by j (who initially actually received growl from the right at the first time step) to mislead and to induce i to open the right door. i may estimate that it could be profitable for j to attempt to mislead because, in this case, i models j modeling i as a "flat" POMDP and thus a literal (and possibly gullible) listener. The resulting belief update is of course different from the one described above, and it includes i's updating the likelihood that j is not a friend due to its messages being different from i's own observations. The complexities of this situation are beyond the scope of this paper, but our framework provides a principled way to study the strategy and rationality of deception. Possibly the most interesting question is whether we can provide a sort of a Bayesian lie detector to differentiate lies from truth. We will report on this separately in our future work.

5. Conclusion and Future Work

We presented a Bayesian approach to pragmatics of communicative acts and a decision-theoretic planning approach for communication and interaction by building on interactive POMDPs. We added to IPOMDPs the capability for agents to exchange messages, and we derived Bayesian update of agents' beliefs due to message exchange, which can be simultaneous with physical actions and observations. We formulated the Bellman optimality for interactive behavior. This allows an agent or an AI system to use the machinery for optimal sequential planning developed within the POMDP

framework for interactive and communicative behaviors. Planning in CIPOMDPs explores possible future states of agent's interactive beliefs, and the expected values of these beliefs, to determine which actions and communicative behaviors lead to best expected future payoffs.

Our approach to communication does not make the common assumption that agents are cooperative in their communicative behavior (Grice, 1975). The capability of agents and AI systems to plan for deception, and to be able to detect and guard from deceptive behaviors of others, opens a number of ethical issues. Some are discussed in (Isaac & Bridwell, 2017), where considerations related to Asimov's Laws of Robotics are elucidated.

The approach we presented leaves a number of important avenues open for future work. First, a more precise specification of the agent communication language, $\mathbb{M}$, is needed. Further, formalizing the plethora of communicative acts used in human communication (Searle, 1975) is open. These include questions, and imperatives. Second, the generation of all possible insincere messages needs to be further described and parametrised. Third, the value iteration and its variants used to plan interactive and communicative behaviors may need to operate in continuous-action space since possible messages span the space of probability distributions (unless we discretize the set $\mathbb{M}$). The existing work on continuous-action POMDP techniques (Seiler et al., 2015; Thrun et al., 2005) are promising in this regard. Finally, a principled investigation of rational deceptive communicative behavior based on the Bayesian update we derived in Proposition 1 is a fascinating avenue for future research. It should include comparison of our approach to game-theoretic techniques (Jehiel, 2006; Kamenica & Gentzkow, 2017), and it is a subject of our ongoing work.

Appendix

Proof of Proposition 1

We prove Proposition 1 analogously to how the Bayesian update is derived in POMDPs and IPOMDPs (Gmytrasiewicz & Doshi, 2005), except that we account for message exchange. Proposition 1 states:

$$b_i^t(is^t) = P(is^t \mid b_i^{t-1}, a_i^{t-1}, m_{i,s}^{t-1}, o_i^t, m_{i,r}^t) = \alpha \sum_{is^{t-1}:\hat{\theta}_j^{t-1}=\hat{\theta}_j^t} b_i^{t-1}(is^{t-1}) \sum_{a_j^{t-1}} Pr(m_{j,s}^{t-1}, a_j^{t-1} \mid \theta_j^{t-1}) \times \tag{13}$$

$$O_i(s^t, a^{t-1}, o_i^t) T_i(s^{t-1}, a^{t-1}, s^t) \sum_{o_j^t} \tau_{\theta_j^t}(b_j^{t-1}, a_j^{t-1}, m_{j,s}^{t-1}, o_j^t, m_{j,r}^t, b_j^t) O_j(s^t, a^{t-1}, o_j^t)$$

We begin with applying definition of conditional probability:

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

We now introduce variable is^{t-1}, i.e., the interactive state, and agent j's action, a_j^{t-1}, at time $t-1$, and use conditional independence of i's observation and state and message received:

$$
\begin{aligned}
b_i^t(is^t) = \alpha \sum_{is^{t-1}} b_i^{t-1}(is^{t-1}) \sum_{a_j^{t-1}} Pr(is^t, o_i^t, m_{i,r}^t \mid a_i^{t-1}, a_j^{t-1}, m_{i,s}^{t-1}, is^{t-1}) \times \\
Pr(a_j^{t-1} \mid a_i^{t-1}, is^{t-1}, m_{i,s}^{t-1}) \\
= \alpha \sum_{is^{t-1}} b_i^{t-1}(is^{t-1}) \sum_{a_j^{t-1}} Pr(is^t, m_{i,r}^t \mid a^{t-1}, m_{i,s}^{t-1}, is^{t-1}) \times \\
Pr(o_i^t \mid a^{t-1}, m_{i,s}^{t-1}, is^{t-1}) Pr(a_j^{t-1} \mid is^{t-1}, m_{i,s}^{t-1})
\end{aligned} \tag{14}
$$

where a stands for agents' joint action.

Now, we work on term $Pr(is^t, m_{i,r}^t \mid a^{t-1}, m_{i,s}^{t-1}, is^{t-1})$ and expand the intentional model $is^t = (s^t, \theta_j^t) = (s^t, b_j^t, \hat{\theta}_j^t)$:

$$
\begin{aligned}
Pr(is^t, m_{i,r}^t \mid a^{t-1}, m_{i,s}^{t-1}, is^{t-1}) &= \frac{Pr(s^t, b_j^t, \hat{\theta}_j^t, m_{i,r}^t, a^{t-1}, m_{i,s}^{t-1}, is^{t-1})}{Pr(a^{t-1}, m_{i,s}^{t-1}, is^{t-1})} \\
&= \frac{Pr(b_j^t \mid s^t, \hat{\theta}_j^t, m_{i,r}^t, a^{t-1}, m_{i,s}^{t-1}, is^{t-1}) Pr(s^t, \hat{\theta}_j^t, m_{i,r}^t, a^{t-1}, m_{i,s}^{t-1}, is^{t-1})}{Pr(a^{t-1}, m_{i,s}^{t-1}, is^{t-1})} \\
&= Pr(b_j^t \mid s^t, \hat{\theta}_j^t, m_{i,r}^t, a^{t-1}, m_{i,s}^{t-1}, is^{t-1}) Pr(s^t, \hat{\theta}_j^t, m_{i,r}^t \mid a^{t-1}, m_{i,s}^{t-1}, is^{t-1}) \\
&= Pr(b_j^t \mid s^t, \hat{\theta}_j^t, m_{i,r}^t, a^{t-1}, m_{i,s}^{t-1}, is^{t-1}) Pr(s^t, \hat{\theta}_j^t \mid a^{t-1}, m_{i,s}^{t-1}, is^{t-1}) \times \\
&\quad Pr(m_{i,r}^t \mid a^{t-1}, m_{i,s}^{t-1}, is^{t-1}) \\
&= Pr(b_j^t \mid s^t, \hat{\theta}_j^t, m_{i,r}^t, a^{t-1}, m_{i,s}^{t-1}, is^{t-1}) Pr(\hat{\theta}_j^t \mid s^t, a^{t-1}, m_{i,s}^{t-1}, is^{t-1}) \times \\
&\quad Pr(s^t \mid a^{t-1}, is^{t-1}) Pr(m_{i,r}^t \mid a^{t-1}, m_{i,s}^{t-1}, s^{t-1}, \theta_j^{t-1})
\end{aligned}
$$

Assuming that agent's frames do not change (this is called model non-manipulability assumption in (Gmytrasiewicz & Doshi, 2005)): $Pr(\hat{\theta}_j^t \mid s^t, a^{t-1}, is^{t-1}) = I(\hat{\theta}^{t-1}, \hat{\theta}_j^t)$ (i.e., it is 1 if $\hat{\theta}^{t-1}$ equals $\hat{\theta}_j^t$ and 0 otherwise). Further, given that $m_{i,r}^t$ is independent of $(a_i^{t-1}, m_{i,s}^{t-1}, s^{t-1})$ given $\theta_j^{t-1}, a_j^{t-1}$, we have:

$$
\begin{aligned}
Pr(is^t, m_{i,r}^t \mid a_i^{t-1}, m_{i,s}^{t-1}, is^{t-1}) = &Pr(b_j^t \mid s^t, \hat{\theta}_j^t, m_{i,r}^t, a^{t-1}, m_{i,s}^{t-1}, is^{t-1}) I(\hat{\theta}_j^{t-1}, \hat{\theta}_j^t) \times \\
&T_i(s^{t-1}, a^{t-1}, s^t) Pr(m_{i,r}^t \mid \theta_j^{t-1}, a_j^{t-1})
\end{aligned} \tag{15}
$$

Now we introduce agent j's observation and use conditional independence of j's observation and i's messages:

$$
\begin{aligned}
Pr(b_j^t \mid s^t, \hat{\theta}_j^t, m_{i,r}^t, a^{t-1}, m_{i,s}^{t-1}, is^{t-1}) &= \sum_{o_j^t} Pr(b_j^t \mid s^t, \hat{\theta}_j^t, a^{t-1}, is^{t-1}, m_{i,s}^{t-1}, o_j^t, m_{i,r}^t) \times \\
&\quad Pr(o_j^t \mid s^t, \hat{\theta}_j^t, a^{t-1}, is^{t-1}, m_{i,s}^{t-1}, m_{i,r}^t) \\
&= \sum_{o_j^t} Pr(b_j^t \mid s^t, \hat{\theta}_j^t, a^{t-1}, is^{t-1}, m_{i,s}^{t-1}, o_j^t, m_{i,r}^t) \times \\
&\quad Pr(o_j^t \mid s^t, \hat{\theta}_j^t, a^{t-1}, m_{i,s}^{t-1}, m_{i,r}^t) \\
&= \sum_{o_j^t} \tau_{\theta_j^t}(b_j^{t-1}, a_j^{t-1}, o_j^t, b_j^t, m_{i,s}^{t-1}, m_{i,r}^t) O_j(s_t, a^{t-1}, o_j^t)
\end{aligned}
$$

So that:

$$Pr(b_j^t \mid s^t, \hat{\theta}_j^t, m_{i,r}^t, a^{t-1}, m_{i,s}^{t-1}, is^{t-1}) = \sum_{o_j^t} \tau_{\theta_j^t}(b_j^{t-1}, a_j^{t-1}, o_j^t, b_j^t, m_{i,s}^{t-1}, m_{i,r}^t) O_j(s_t, a^{t-1}, o_j^t) \tag{16}$$

From Equations (14) (15) and (16) we get:

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

which is equivalent to Proposition 1.

Proof of Theorem 1 (Sketch): Proof is analogous to one for IPOMDPs (Gmytrasiewicz & Doshi, 2005) and POMDPs (Smallwood & Sondik, 1973), except that the backup operator, H, is given by Equation (8). H is a contraction, with the expressions simplifying analogously. Theorem 1 then follows directly from the Contraction Mapping theorem (Stokey & Lucas, 1989).

The proof of Theorem 2, the property of piece-wise linearity and convexity (PWLC), of the CIPOMDP value function, is also analogous to corresponding proofs for POMDPs (Hausktecht, 1997; Smallwood & Sondik, 1973) and IPOMDPs (Gmytrasiewicz & Doshi, 2005). Proof proceeds by induction over the planning horizon.

Acknowledgments

The author is grateful to Sarit Adhikari, of UIC CS Department, for help with the proofs and the examples, to Natalie Parde, also from CS at UIC, for helpful comments, and to anonymous reviewers.

References

Albrecht, S. V., & Stone, P. (2018). Autonomous agents modelling other agents: A comprehensive survey and open problems. *Artificial Intelligence*, *258*, 66–95.

Aumann, R. J. (1999). Interactive epistemology I: Knowledge. *International Journal of Game Theory*, pp. 263–300.

Austin, J. L. (1962). *How to do Things with Words*. Clarendon Press.

Axelrod, R. (1984). *The Evolution of Cooperation*. Basic Books.

Binmore, K. (1990). *Essays on Foundations of Game Theory*. Blackwell.

Bolander, T., & Andersen, M. B. (2011). Epistemic planning for single- and multi-agent. *Journal of Applied Non-Classical Logics*, *21*, 9–34.

Brandenburger, A., & Dekel, E. (1993). Hierarchies of beliefs and common knowledge. *Journal of Economic Theory*, *59*, 189–198.

Chandrasekaran, M., Doshi, P., Zeng, Y., & Chen, Y. (2017). Can bounded and self-interested agents be teammates? Application to planning in ad hoc teams. *Autonomous Agents and Multi-Agent Systems*, *31*(4), 821–860.

Dennett, D. (1986). Intentional systems. In Dennett, D. (Ed.), *Brainstorms*. MIT Press.

Doshi, P., & Gmytrasiewicz, P. (2009). Monte carlo sampling methods for approximating interactive pomdps. *Journal of AI Research*, *34*, 297–337.

Engesser, T., Bolander, T., Mattmüller, R., & Nebel, B. (2017). Cooperative epistemic multi-agent planning for implicit coordination. *Electronic Proceedings in Theoretical Computer Science (ISSN: 2075-2180) (DOI: http://dx.doi.org/10.4204/EPTCS.243.6)*, *243*.

Flower, T. P., Gribble, M., & Ridley, A. R. (2014). Deception by flexible alarm mimicry in an african bird. *Science*, *334*, 513–516.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*, 998–998.

Franke, M., & Jager, G. (2016). Probabilistic pragmatics, or why Bayes' rule is probably important for pragmatics. *Zeitschrift fur Sprachwissenschaft*, *35*, 3–44.

Frith, C., & Frith, U. (2005). Theory of mind. *Current Biology*, *15*(17), R644 – R645.

Fudenberg, D., & Tirole, J. (1991). *Game Theory*. MIT Press.

Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences*, *2*(12), 493 – 501.

Gmytrasiewicz, P., & Doshi, P. (2005). A framework for sequential planning in multiagent settings. *Journal of Artificial Intelligence Research*, *24*, 49–79. http://jair.org/contents/v24.html.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences,*, *20*(11), 34–42.

Goodman, N. D., & Lassiter, D. (2014). Probabilistic semantics and pragmatics: Uncertainty in language and thought, a draft chapter. In Lappin, S., & Fox, C. (Eds.), *Wiley-Blackwell Handbook of Contemporary Semantics — second edition, https://web.stanford.edu/ ngoodman/papers/Goodman-HCS-final.pdf*.

Grice, H. P. (1975). Logic and conversation. In Cole, P., & Morgan, J. (Eds.), *Studies in Syntax and Semantics III: Speech Acts*, pp. 41–58. Academic Press.

Halpern, J. Y., & Moses, Y. (1990). Knowledge and common knowledge in a distributed environment. *Journal of the ACM*, *37*(3), 549–587.

Harsanyi, J. C. (1967). Games with incomplete information played by 'Bayesian' players. *Management Science*, *14*(3), 159–182.

Hausktecht, M. (1997). *Planning and control in stochastic domains with imperfect information*. Ph.D. thesis, MIT.

Hoek, W. V. D., & Wooldridge, M. (2002). Tractable multiagent planning for epistemic goals. In *Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS-2002)*, pp. 1167–1174.

Isaac, A., & Bridwell, W. (2017). White lies on silver tongues: Why robots need to deceive (and how).. In P, Lin, K. A., & Jenkins, R. (Eds.), *Robot Ethics 2.0*. Oxford University Press.

Jaeger, G. (2011). Game-theoretical pragmatics. In *J. van Benthem and A. ter Meulen, eds., Handbook of Logic and Language, 2nd edition*, pp. 467–491. Elsevier.

Jehiel, D. E. . P. (2006). Towards a theory of deception. Tech. rep., UCLA Department of Economics.

Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, *101*(2), 99–134.

Kamenica, E., & Gentzkow, M. (2017). Bayesian persuasion with multiple senders and rich signal spaces. *Games and Ecomomic Behavior*, *104*, 411–429.

Kominis, F., & Geffner, H. (2015). Beliefs in multiagent planning: From one agent to many. In *Proceedings of ICAPS*, pp. 157–155.

Konnikova, M. (2016). *The Confidence Game; Why we Fall for It Every Time*. Viking.

Leslie, A. M., Friedman, O., & German, T. P. (2004). Core mechanisms in 'theory of mind'. *Trends in Cognitive Sciences*, *8*(12), 528 – 533.

Lewis, D. (1979). Scorekeeping in a language game. *Journal of Philosophical Logic*, *8*(1), 339–359.

McKelvey, R., & Palfrey, T. (1995). Quantal response equilibria for normal form games. *Games and Economic Behavior*, *10*, 6–38.

Nair, R., Pynadath, D., Yokoo, M., Tambe, M., & Marsella, S. (2004). Communication for improving policy computation in distributed pomdps. In *Proceedings of the Agents and Autonomous Multiagent Systems (AAMAS)*.

Nair, R., Roth, M., Yokoo, M., & Tambe, M. (2003). Taming decentralized pomdps: Towards efficient policy computation for multiagent settings. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*.

Narayanan, A. (1988). *On Being a Machine*. Ellis Horwood.

Olihoek, F. (2013). Sufficient plan-time statistics for decentralized pomdps. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, pp. 302–308.

Olihoek, F., Spaan, M., & Vlassis, N. (2007). Dec-pomdps with delayed communication. In *Proceedings of MSDM 2007 May 15, 2007, Honolulu, Hawai'i, USA*.

Papadimitriou, C. H., & Tsitsiklis, J. N. (1987). The complexity of markov decision processes. *Mathematics of Operations Research*, *12*(3), 441–450.

Perea, A. (2012). *Epistemic Game Theory*. Cambridge University Press.

Pietarinen, A. (2007). *Game Theory and Linguistic Meaning*, Vol. 18. Elsevier: Current in the Semantics/Pragmatics Interface.

Pynadath, D., & Tambe, M. (2002). The communicative multiagent team decision problem: Analyzing teamwork theories and models. *Journal of AI Research (JAIR)*.

Robinson, J. (2010). *There's a Sucker Born Every Minute: A Revelation of Audacious Frauds, Scams, and Cons – How to Spot Them, How to Stop Them*. Penguin.

Russell, S., & Norvig, P. (2010). *Artificial Intelligence: A Modern Approach*. Prentice Hall.

Searle, J. R. (1975). A taxonomy of illocutionary acts. In *Language, Mind and Knowledge*, pp. 344–369. Minnesota Studies in the Philosophy of Science.

Seiler, K. M., Kurniawati, H., & Singh, S. P. N. (2015). An online and approximate solver for pomdps with continuous action space. In *IEEE International Conference on Robotics and Automation (ICRA)*, p. 2290–2297.

Shanton, K., & Goldman, A. (2010). Simulation theory. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(4), 527–538.

Smallwood, R. D., & Sondik, E. J. (1973). The optimal control of partially observable Markov decision processes over a finite horizon. *Operations Research*, pp. 1071–1088.

Stahl, D. O., & Wilson, P. W. (1994). Experimental evidence on players' models of other players. *Journal of Economic Behavior and Organization*, *25*, 309–327.

Stahl, D. O., & Wilson, P. W. (1995). On players' models of other players: Theory and experimental evidence. *Games and Economic Behavior*, *10*, 218–254.

Stalnaker, R. (1978). Assertion. In Cole, P. (Ed.), *Syntax and Semantics 9: Pragmatics*. Academic Press.

Stokey, N. L., & Lucas, R. E. (1989). *Recursive Methods in Economic Dynamics*. Harvard Univ. Press.

Thrun, S., Burgard, W., & Fox, D. (2005). *Probabilistic Robotics*. MIT Press.

Tian, L., Luo, J., & Huang, Z. (2013). Communication based on interactive dynamic influence diagrams in cooperative multi-agent systems. In *Proceedings of the The 8th International Conference on Computer Science and Education (ICCSE 2013), Colombo, Sri Lanka*, pp. 56–61.

Tian, L., Luo, J., Zeng, Y., & Wu, H. (2016). Modeling and algorithms for multiagent communication through interactive dynamic influence diagrams. *Applied Artificial Intelligence*, *30*(4), 352–377.

Vogel, A., Potts, C., & Jurafsky, D. (2013). Implicatures and nested beliefs in approximate decentralized-pomdps. *ACL*.

Whaley, B. (2016). *Practise to Decieve: Learning Curves of Military Deception Planners*. Naval Institute Press.

Wu, F., Zilbersein, S., & Chen, X. (2011). Online planning for multi-agent systems with bounded communication. *Artificial Intelligence*, *167*, 487–511.

Young, S., Gasic, M., Thomson, B., & Williams, J. D. (2013). Pomdp-based statistical spoken dialogue systems: a review. *Proceedings of IEEE*, *101*, 1160–1179.

Zeevat, H. (2015). Perspectives on natural language semantics and pragmatics. In Zeevat, Henk, S., & Hans-Christian (Eds.), *Bayesian Natural Language Semantics and Pragmatics*, pp. 1–24. Springer.

Zhou, L., & Luo, J. (2012). A communication model for interactive pomdps. *2012 7th International Conference on Computer Science Education (ICCSE)*, 169–174.