Honda Research Institute EU
**Honda Research Institute Europe GmbH**
https://www.honda-ri.de/

# Theory of Mind based Communication for Human Agent Cooperation

## Moritz Bühler, Thomas H Weisswange

### 2020

**Preprint:**

This is an accepted article published in IEEE International Conference on Human-Machine Systems. The final authenticated version is available online at: https://doi.org/10.1109/ICHMS49158.2020.9209472

Accepted Publication to the 2020 IEEE International Conference on Human-Machine Systems
September, 2020, Rome, Italy

# Theory of Mind based Communication for Human Agent Cooperation

Moritz C. Buehler
*TU Darmstadt, Germany*
https://orcid.org/0000-0002-4484-4315

Thomas H. Weisswange
*Honda Research Institute Europe GmbH*
https://orcid.org/0000-0003-2119-6965

***Abstract*—For human agent cooperation, reasoning about the partner is necessary to enable an efficient interaction. To provide helpful information, it is important not only to account for environmental uncertainties or dangers but also to maintain a sophisticated understanding of each other's mental state, a theory of mind. Sharing every piece of information is not a good idea, as some may be irrelevant at time or already known, leading to distraction and annoyance. Instead, an agent will have to estimate the novelty and relevance of information for the receiver, to trade off the cost of communication against potential benefits.**

**We propose the concept of theory of mind based communication as principled formulation to ground an agents cooperative communication on an understanding of the receiver's mental states to support her awareness and action selection. Therefore we formulate the problem of whether, when and what information to share as a sequential decision process with the human belief as central source of uncertainty. The agent's communication decision is obtained online during interaction by combining a second level Bayesian inference of human belief with planning under uncertainty, evaluating the influence of communication on the human belief and her future decisions. We discuss the resulting behavior on an illustrative communication scenario with different uncertain state aspects that an observing agent can communicate to the actor.**

***Index Terms*—Human agent interaction, Communication, Theory of Mind, Human Belief, POMDP, Planning under Uncertainty**

## I. Introduction

With the improvement of manipulation and processing capabilities of technical systems like robots, the interaction with a human becomes an interesting focus. However many technical systems are used like tools, they wait for human commands to execute or provide measurements that the human can read and interpret. However, this interpretation of interaction implies limits on the achievable support especially in more complex situations. Inter-dependencies become important and interaction factors known from human human cooperation, like trust or awareness for situation and partner, have to be considered. An intelligent cooperative system will have to take into account both, the human's action but also her mental state to support her efficiently in achieving her goals.

Communication is a key element of cooperation, that allows for sophisticated teaming. It is used to exchange information, about perceived external objects as well as internal states as plans, to coordinate individual behaviors towards a common goal. Sharing information supports other's awareness of the current situation which is needed for a good decision making.

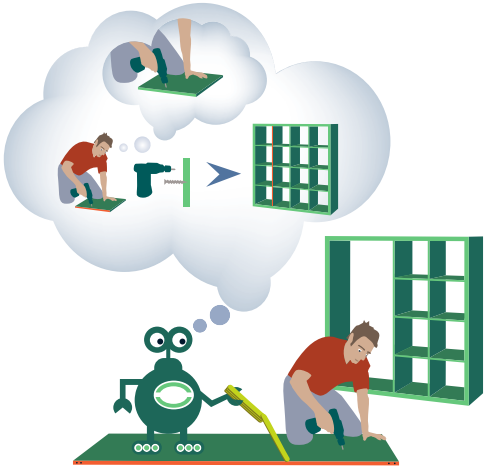

Fig. 1. Human is unaware about the plank's asymmetry. Artificial agent infers it from her behavior, anticipates the final result and decides to provide the missing information.

But information exchange requires perception and processing resources including attention mechanisms for both partners. Receiving too much information might overload and distract from other important aspects. Therefore, it is usually not advisable to share all available information with the partner but rather reason about its relevance to decide when and what to communicate. This does not only depend on the state of the environment, but importantly on the partner's current knowledge, awareness and goals. If one understands a partner's behavior, one can reason about her knowledge to detect a lack of awareness for the current situation and to support her with the right information. Figure 1 illustrates an example for the complexity of communication in cooperation. Humans create a sophisticated model of other's mental processing, a Theory of Mind (ToM). They infer goals and beliefs to explain their behavior [1]. This capability of understanding others supports communicative decisions, by estimating the relevance that specific information possesses for a partner [2]. An autonomous agent that wants to assist a human partner effectively, will need a similar understanding of human mental reasoning to improve the cooperative performance while avoiding information overload and annoyance. This is especially important due to the open or implicit delegation setting where the agent itself has to take initiative to support

978-1-7281-5871-6/20/$31.00 © 2020 IEEE

the human partner when it is necessary, providing "over help" [3].

In this paper, we will introduce the concept of theory of mind based communication. We formulate a cooperative problem of what and when to communicate as the optimization of a cooperative reward in a decision theoretic framework as a POMDP where we include environment state as well as the human belief with respect to multiple task relevant aspects. The belief is inferred online using the observation of human perception and action and used to evaluate the potential effects of communication on subsequent human decisions based on this belief. This results in an informed communication decision to provide support to a human when detecting a lack of relevant information while avoiding unnecessary disturbances.

*a) Related Work:* Many different research streams have tackled the issue of communication planning and human state estimation. Classical signaling games [4] consider Information transmission in semi-cooperative settings, where equilibria are strongly dependent on the knowledge and alignment of individual utilities. Multi agent systems that target explicit cooperation usually use a joint policy that is globally optimized. This can include decisions when to communicate about own states or goals to synchronize [5] or in which situations to best request such information [6] if dealing with conflicting space constraints. When different aspects or types of information are available it is necessary to also select what observation to communicate by explicitly comparing the impact on the joint policy [7]. Additionally, it is possible to include learned human preferences for certain information into the selection [8]. Respecting overlapping perception, Foerster et al. [9] use the definition of common knowledge between all agents to optimize for joint policies that favor actions which implicitly reveal parts of each agent's private knowledge to the group. Rabinowitz et al [12] trained an ensemble of neural networks to predict aspects of policy and reward of different types of interaction partners. The networks were first trained with supervision, and afterwards able to generate accurate results after observation of few action examples.

These approaches rely on prior coordination and the commitment to an explicit joint policy. However, we can not assume that humans and artificial agents synchronize their strategies beforehand, nor that every human will follow the exact same strategy. Instead, the human agent cooperation setting resembles an ad hoc cooperation, requiring a flexible and fast adapting agent behavior [10]. Communication in ad hoc settings is analyzed in [11], coordinating actions of a new agent with an existing team through inferring the type of each team member to include this into decision making.

Regarding the inference of different mental states of a (human) partner, many specific methods are proposed. Work in language understanding benefits from inference of intention using speech context [13]. Generally, intentions are useful for short-term assistance, e.g. with a robot [14]. For long-term assistance, inverse reinforcement learning [15] has proven to be of interest, as it infers the goal respectively reward function from human behavior. This information can also be used to integrate interaction effects of own actions into behavior selection [16] or perform optimal cooperation through an explicitly shared reward function [17]. Knowing the goal, [18] estimates a human's understanding of environmental dynamics at the start of a demanding control task. Afterwards the human control action can be compensated for detected biases.

All of these aspects (and some more) can be described jointly in the human's belief state. Inspired by the theory of mind, Baker et al. [19] introduce a generative model for human action selection based on belief and fit an inverted model retrospectively to an observed human action sequence to explain her behavior. Poeppel et al. [20] use a number of probabilistic human models with different complexities to explain similar behavior traces. Each model made assumptions about which parts of the human belief (goal, environment layout) were known or uncertain. Our previous work [21] focused on doing Bayesian belief inference in interactive settings where it is important to estimate a human belief online from only few actions.

In the following, we formulate the problem of sharing appropriate information to a cooperation partner. We combine planning ideas from multi agent approaches with model based inference of the human belief online to evaluate the effect of explicit communication on joint performance.

## II. Problem formulation

We consider the general cooperative situation that a human acts to reach a certain goal within an uncertain environment and a cooperative artificial agent supports the human in her task. Here, we focus on informative communication actions of the agent. An approach that explicitly models an agent's task actions within a human's belief can be found in previous work [21]. Information exchange is directed from agent to the human without additional strategic inference processes on the underlying agent intentions (as is done e.g. in [22]).

We interpret the human as a goal directed agent in a first Partially Observable Markov Decision Process (POMDP). At each discrete time step she decides for an action $a_H$ changing the environmental state $s$ to a next state $s'$ according to the transition function $T(s, a, s') = p(s'|s, a)$. The human does not know the environmental state $s$ but perceives limited information as observation $o_H$ generated according to the observation function $O_H(s, o_H) = p(o_H|s)$. The human wants to achieve the goal encoded in the reward $R_H(s, a_H, s')$ that she receives each time step, by choosing her actions $a_H$ to maximize the cumulated expected future reward $E[\sum_{\tau=t}^{T_{end}} R(s^\tau, a_H^\tau, s^{\tau+1})]$. Since the human does not know the true environmental state $s$, she has to reason about it based on past observations and actions. This is formulated as probability distribution over the unknown current state, called the human belief $b_H(s)$. Besides the environmental state, the human can also be uncertain about transition function $T$, observation function $O_H$ or reward $R_H$. By parameterizing these uncertainties we can extend the state to include these parameters which shifts all uncertainty to the extended state and allows for a simpler formulation.

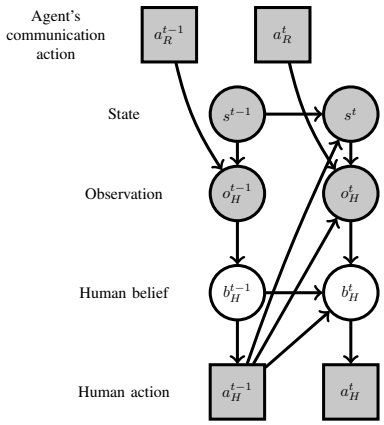


Fig. 2. Directed graphical model for the artificial agent's decision problem regarding communication actions $a_R$. Filled nodes are observable.

The artificial agent faces a second POMDP. The agent observes the human acting in the environment and selects one of several communication actions $a_R$ to share parts of its knowledge with the human, leading to an additional human observation. These communication actions may describe different state aspects e.g. regarding different objects or different ways of communicating (implicit or explicit, reliable or noisy). A cost of communication $R_{\text{comm}}(s, a_R) \leq 0$ is necessary to represent the corresponding disturbance. The artificial agent is fully cooperative in the sense of [17], meaning it receives the same reward as the human while considering the communication cost: $R_R(s, a_H, a_R, s') = R_H(s, a_H, s') + R_{\text{comm}}(s, a_R)$. The optimal action is computed by optimizing the expected cumulative reward

$$a_R^* = \arg\max_{a_R^t} E[\sum_{\tau=t}^{T_{end}} R_R(s^\tau, a_H^\tau, a_R^\tau, s^{\tau+1})] \tag{1}$$

As the cumulated future return mainly depends on the human behavior, it is necessary to predict the human decisions as well as the effect of the agent's action, which are connected through the human belief. The agent state $s_R$ consists of the environmental state, the human belief and the current human action, $s_R = (s, b_H, a_H)$. However, for the artificial agent, the human belief is unknown and has to be inferred from observation $o_R$.

Figure 2 shows the causal interdependencies of both POMDPs as probabilistic graphical model from the artificial agent's perspective.

For action selection, the agents should evaluate their expected future reward depending on their behavior. This can be described through an action value function $Q(b, a)$ which assigns each action to its expected future return, given current belief,

$$Q(b^t, a^t) = E[\sum_{\tau=t}^{T_{end}} R(s^\tau, a^\tau, s^{\tau+1})]. \tag{2}$$

Calculating optimal solutions for a POMDP, respectively the exact evaluation of the action value function $Q$, is intractable besides for small problems. Instead, approximative methods are proposed to estimate the action value function, for example tree based online algorithms that can flexibly cope with larger POMDPs [23]. Starting with the current belief, a finite search tree is expanded over possible future actions and observations, until a certain depth in time or number of nodes is reached.

In the following, we present the human model used to understand the human behavior and infer the human belief which is subsequently used for planning the effects of communication.

## III. Human Model and Inference of Human Belief

For the inference of the hidden mental states of the human, we need a model for her cognitive processing, perception and decision making. Based on the human trajectories $(a_H, o_H)^{0\dots t}$, we can invert this model to estimate her belief.

*a) Human Model:* When the human interacts with its environment, she perceives observations $o_H$ containing information about the environmental state. According to a Bayesian update, the new human belief becomes

$$p(s|o_H) \sim p(o_H|s)p(s), \tag{3}$$

with the observation likelihood $p(o_H|s)$.

To decide for her next action, the human will evaluate her current belief $b_H$ according to the action value function (2). To allow for suboptimal or noisy decisions, we assume a softmax action selection policy by the human,

$$\begin{aligned} p(a_H|b_H) &= \text{softmax}(Q_H(b_H, a_H)) \\ &\sim \exp(\tau Q_H(b_H, a_H)), \end{aligned} \tag{4}$$

where $\tau$ characterizes the degree of rationality of action selection. A human action will lead to the transition of environmental state, $T(s, a_H, s')$, which the human will account for. Hence the human belief will be updated to

$$p(s'|a_H) = \sum_s T(s, a_H, s')p(s). \tag{5}$$

*b) Belief representation and inference:* Inferring the human's belief is a second level inference since the belief itself is the result of the human's state inference. Inferring a general probability distribution over continuous probabilities becomes intractable even for small state spaces. As a further restriction, we need the human belief during the interaction to decide for the agent's action online. Therefore, we approximate the full inference through a parametrized distribution for the second order belief as proposed in prior work [21].

The human belief $b_H(s)$ assigns a probability to each possible state $s$. For the agent's belief about the belief we consider a Dirichlet distribution $b_H \sim Dir(b_H|\alpha)$, and the approximate inference of human belief is achieved by calculating the parameters $\alpha$ of the Dirichlet distribution. This parametrized distribution is flexible enough to describe relevant second order belief configurations. Further, in most practical application, the environmental state contains several independent aspects, like positions of different agents and objects, or states of objects, $s = (s_1 \dots s_k)$. To avoid the combinatorial increase in the total number of states, these can be assumed to factorize as

we proposed in [21], as long as correlations effects e.g. introduced through relative perception are negligible. Otherwise, additional forward backward propagations should be applied.

The human action serves as sparse feedback for the agent's belief estimation since it is directly caused by the human belief. When observing a human action $a_H$, the human belief estimate has to be updated to respect her action decision, (4),

$$p(b_H|a_H) \sim p(a_H|b_H)p(b_H). \tag{6}$$

For the human observation and her respect for state transition, we can directly use equations (3) and (5) to update the agent belief. For all these updates, we sample the human belief and use moment matching to rematch the new belief estimate to a Dirichlet distribution.

## IV. Communication decisions

The inferred human belief should serve as the basis, to decide in a principled manner for possible information actions. A communication of the artificial agent will influence the human belief and, via the human decision making, effect her action, the environmental state and the common reward. Therefore, the agent has the possibility to improve the joint performance with appropriate communication decisions.

In a cooperative setting with communication cost, information sharing is only beneficial, if the human belief is incorrect or uncertain. However, it further depends on the relevance of information aspects for the current situation, i.e. the human's situation awareness, since parts of the state might be irrelevant for the current action evaluation. The agent should estimate the potential impact of different information actions on the human decisions.

*a) Costs and effects of communication:* The artificial agent can choose among several communication actions $a_R \in A_R = \{a_\emptyset, a_1, \dots a_k\}$, including to not communicate ($a_\emptyset$) or to inform the human about specific aspects of the state, e.g. different positions or object states. The cost for communication, $R_{\text{comm}}(s, a_R) \leq 0$ (zero in the case of no communication) is used to respect the time delay of task completion due to the human's information processing.

The communication action $a_R$ generates an additional observation for the human $o_H$, leading to a human belief update as in (3) according to the observation likelihood $p(o_H|s, a_R)$. In the following, we consider that the agent's communication actions transmit distinct state aspects with action $a_R = a_i$ sharing information about the $i$th aspect $s_i$. For our example, we assume a reliable information transmission, i.e. the observation likelihood is $p(o_H|s, a_R = a_i, (o_H)_i = s_i) \approx 1$. For practical applications with specific communication modes, the communication likelihood can be taken from communication models (e.g. [24]).

*b) Planning:* To optimize the cumulated expected reward, the artificial agent needs to plan the future effects of communication actions $a_R$ on its state $s_R$, which includes the human mental states.

A transition of agent state $s_R$ concatenates several processes, namely the communication influence on human belief, the human decision, the environmental state transition and the human receiving and processing a new observation,

$$\begin{aligned} p(s'_R|s_R, a_R) &= p(s', b'_H, a'_H|s, b_H, a_H, a_R) \\ &= \sum_{o'_H, b_{H-}} \underbrace{p(b'_H|a_H, b_{H-}, o'_H)}_{\text{belief update}} \underbrace{p(o'_H|s')}_{\text{human perception}} \\ &\quad \underbrace{p(s'|s, a'_H)}_{\text{state trans.}} \underbrace{p(a'_H|b_{H-})}_{\text{human decision}} \underbrace{p(b_{H-}|a_R, b_H)}_{\text{comm. effect}} \end{aligned} \tag{7}$$

where $b_{H-}$ is the intermediate human belief after communication. This transition function is central for planning the effects of communication and to evaluate the agent's action value to compute the best action according to (1).

## V. Example and Discussion

To illustrate our method, we consider an example scenario, that includes the possibility for cooperative communication with uncertainty. It has similarities to the examples used in [19], [20], [25] but includes communication options for different state aspects. One agent, the "human", moves in a grid world to reach some goal position, shown in Fig. 3 bottom left. She has three available actions, move forward, turn left and turn right (transition with certainty if next position is accessible). Some of the grid cells are inaccessible ("walls", black) and one location is a "door" that can be closed or open (grey). The episode ends when the human reaches the goal located at one of two possible positions ($g_1$, or $g_2$). Every movement leads to a reward of $R_H = -1$. The human has limited perception, observing only local information. She can perceive the accessibility of the cells in front, to the left and to the right (blue triangle) with a certainty of 90% for each. She does know the overall wall configuration (map) but can be uncertain about her location and orientation as well as door state and goal position, leading to a state space of 352 (22 position $\times 4$ directions $\times 2$ goal positions$\times 2$ door states). We factor the state into three groups, combined position and orientation, the door state, and the goal location.

A second, artificial agent, the "speaker" can observe the entire situation, $o_R = (s, a_H, o_H)$, leaving the human belief as uncertain state aspect. Each time step, the speaker has the option to reliably share one aspect of the state to the human at a constant communication cost, $R_{com} = -1.5$ for all $a_R \neq a_\emptyset$. It optimizes the assistive reward function to trade-off communication cost against and human reward. For both agents we implement a rational POMDP strategy, combining belief update with greedy action selection based on a depth 2 planning tree together with an MDP based leaf evaluation. The speaker uses a sample size of 100 for the human belief inference.

To demonstrate certain behaviors we will discuss concrete scenarios for the above setting. We vary the prior human belief and also the initial speaker belief of human belief to generate different illustrative situations. As alternative, standard approaches consider communication when the human deviates from a nominal trajectory, either with a warning, by proposing the next action, or by sharing every available

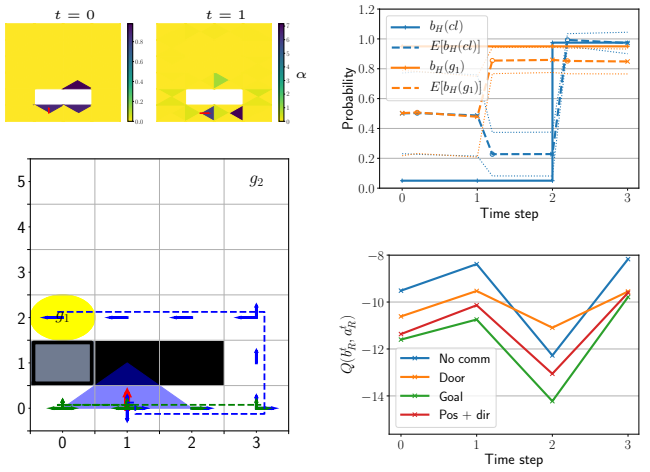


Fig. 3. Scenario: What to communicate. Bottom left: Starting position (red arrow), relative field of view (blue triangle), wall fields (black), door (grey), goal (yellow). Trajectories with (blue) and without (green) communication. Top: Inferred belief position (left) goal $g_1$ and door closed (right). Bottom right: Speaker's action values.

information. As another option, a ToM equipped speaker could always communicate when a false or uncertain human belief is detected independently of its impact (similar to [21]).

*a) What to communicate:* Let's first consider a case, where the human starts in location $l = (1, 0, \text{north})$ with goal location $g_1$ ($l_g = (0, 2)$) and door closed (see Fig. 3 bottom left). The human goal location prior is correct, however she does have a false belief for the door state (beliefs it to be open) and is uncertain about its position between 3 options, $(1, 0, \text{north}), (2, 0, \text{north}), (2, 2, \text{south})$. The speaker has a uniform belief for human goal and door belief and is ignorant about the human belief for the 3 starting positions.

Without speaker interaction, the human acts suboptimal due to the false belief for the door aspect (green path in Fig. 3, bottom left). After the human agent turns right, she perceives an observation increasing the belief for state (1, 0) and (2, 0) significantly. Since she has a high probability for the door to be open, she again turns right to take the shorter path (although it is blocked). When she reaches location $(0, 0)$ the next behavior depends on the ratio of prior door belief to perception certainty and the length of the alternative path. When her probability for a closed door is large enough, she turns and approaches the goal on the longer but open path through (3, 0).

When the speaker observes the second right turn of the human, after time step 1, it infers with high probability that the human door belief is false (Fig. 3 top right) and the goal belief is right. Consequently the speaker's plan evaluation expects a large positive effect (compared to the communication cost) from communicating the door state, which will avoid that the human tries to pass through the closed door, (3 bottom right). The speaker will provide information about the door state to the human, who then takes the possible path through the right passage (blue trajectory in Fig. 3 bottom left).

In this scenario, the "what" of information is important. A method without belief inference could also detect the human unawareness due to the deviation from the optimal path (the "when"). Warning or proposing the optimal next action however would be less explicit and less helpful in this situation, since the human could not distinguish between door state error and goal location error. The human's behavior revealed precise information about her belief. Informing the human about the true door state gives her all necessary information to act appropriate in this situation while it takes less effort than sharing all available information.

*b) When to (not) communicate:* Our approach also provides benefits for the question when to communicate. There are situations, where the human deviates from the optimal path but it is not beneficial to intervene, or dangerous situations where it makes sense to inform the human when she is still on the track but might miss some important information for the future.

Consider the human starting in location $(1, 0, \text{west})$ knowing position and goal location $l_g = (0, 2)$ but being unaware about the closed door (see Fig. 4 bottom left). Since the human is uncertain about the door state, she will move forward and receive an observation about it. Although this behavior is a deviation from the optimal MDP policy, interrupting the human is not necessary because she already perceives the observation that the door is closed (Fig. 4 top right). After her first step, she will be aware of the situation and follow the optimal path, without a need for any intervention.

Theory of mind based communication intrinsically tolerates noisy human actions. A human may deviate from the optimal path despite having all relevant information due to approximate planning or stochasticity in decision or action execution. This is respected in the human policy formulation (4). The speaker will only intervene, when a false or uncertain human belief is estimated as likely cause of the deviation. On the other hand, if the speaker is uncertain about a human belief which might lead to a very bad human action (danger), it will inform the human about this aspect although she is still on the optimal path.

Although the speaker may know about a false human belief, it can decide not to communicate the true state. This is the case, if the state aspect is irrelevant for the current situation (hence an unnecessary disturbance) or when communication would lead to negative effects. Consider the human starting again in location $(1, 0, \text{north})$ but having the false belief about the goal location to be at $(3, 5)$ instead of the true $(0, 2)$, while she has no idea about the door, which is actually closed. If the speaker would know about the false goal belief and communicate the true goal state immediately, a rational human would gather the door state first, which is worse then before and not in the interest of the cooperative speaker.

*c) Additional interesting communication strategies:* In this last scenario, one could argue for telling the human that the door is closed. If the human would reason about the speaker's decision making, she could conclude that the door state is relevant which implicitly tells her, that the goal location should be $(0, 2)$. To allow for such strategic reasoning,

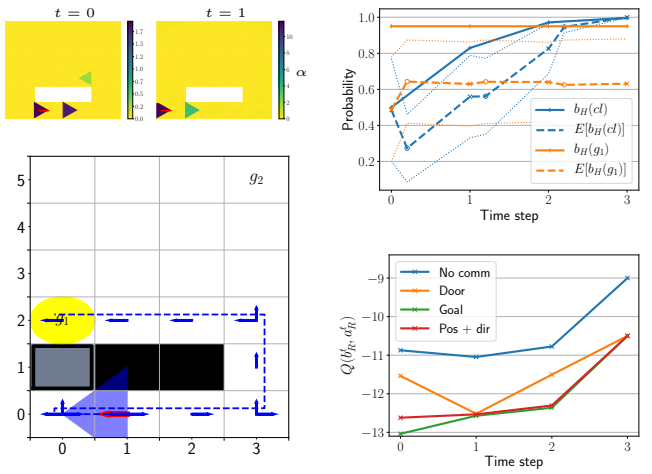


Fig. 4. Scenario: When to not communicate. Above: Inferred human belief. Bottom: start configuration and trajectory (left), speaker's action values (right)

we could extend our method with an inference model of the speaker's belief in the human model. However this will increase complexity and processing load for both agents with a higher sensibility for misunderstandings.

When considering to not tell the human about a false belief, one could also discuss "white lies" which means to communicate a wrong state when it is more cost effective but still has the same positive effect on the human behavior [26]. This would be possible by giving the speaker the choice to communicate something in deviation to its own knowledge. However such behavior can effect the human's trust and represents an ethical controversy.

## VI. Conclusion

We presented a concept to decide for cooperative communication based on an understanding of the mental state of a human partner. An uncertain setting is considered, where an artificial agent assistively supports a human with information constrained by a cost of communication. We infer the human belief by observing human actions and observations. This belief of human belief is part of our POMDP formulation for the agent to evaluate the potential effects of communication actions on the task progress. To our knowledge we described for the first time a general approach to select communication actions considering a human receiver's external as well as mental situation by combining inference of human belief with decision making under uncertainty. We illustrated the resulting behavior with an example scenario demonstrating the principled communication trade off and discussed its benefits compared to other communication strategies.

As we have shown, a theory of mind based modeling of a human partner offers various opportunities for more efficient and intuitive human agent interaction.

## References


[1] H. Wimmer and J. Perner, "Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception," *Cognition*, vol. 13, no. 1, 1983.

[2] F. G. Happé, "Communicative competence and theory of mind in autism: A test of relevance theory," *Cognition*, vol. 48, no. 2, 1993.

[3] C. Castelfranchi and R. Falcone, "Towards a theory of delegation for agent-based systems," *Robotics and Autonomous Systems*, 1998.

[4] M. J. Osborne, *An introduction to game theory*. Oxford University Press, 2004.

[5] C. V. Goldman and S. Zilberstein, "Optimizing information exchange in cooperative multi-agent systems," in *2nd international joint conference on Autonomous agents and multiagent systems - AAMAS '03*, 2003.

[6] F. S. Melo, M. T. J. Spaan, and S. J. Witwicki, "QueryPOMDP: POMDP-Based Communication in Multiagent Systems," in *EUMAS 2011: European Workshop on Multi-Agent Systems*, 2012.

[7] M. Roth, R. Simmons, and M. Veloso, "What to Communicate? Execution-Time Decision in Multi-agent POMDPs," in *Distributed Autonomous Robotic Systems 7*, 2006.

[8] R. Chitnis, L. P. Kaelbling, and T. Lozano-Perez, "Learning What Information to Give in Partially Observed Domains," in *Proceedings of The 2nd Conference on Robot Learning*, 2018.

[9] J. N. Foerster, F. Song, E. Hughes, N. Burch, I. Dunning, S. Whiteson, M. Botvinick, and M. Bowling, "Bayesian Action Decoder for Deep Multi-Agent Reinforcement Learning," *Proceedings of the 36th International Conference on Machine Learning*, 2019.

[10] S. V. Albrecht and S. Ramamoorthy, "A Game-theoretic Model and Best-response Learning Method for Ad Hoc Coordination in Multiagent Systems," in *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2013.

[11] S. Barrett, N. Agmon, N. Hazon, S. Kraus, and P. Stone, "Communicating with unknown teammates," in *ECAI*, 2014.

[12] N. Rabinowitz, F. Perbet, F. Song, C. Zhang, S. M. A. Eslami, and M. Botvinick, "Machine Theory of Mind," in *Proc. 35th International Conference on Machine Learning*, 2018.

[13] S. Young, M. Gasic, B. Thomson, and J. D. Williams, "POMDP-Based Statistical Spoken Dialog Systems: A Review," *Proceedings of the IEEE*, vol. 101, no. 5, 2013.

[14] O. Görür, B. S. Rosman, G. Hoffman, and S. Albayrak, "Toward integrating Theory of Mind into adaptive decision-making of social robots to understand human intention," 2017.

[15] A. Y. Ng and S. J. Russell, "Algorithms for Inverse Reinforcement Learning," in *Proceedings of the Seventeenth International Conference on Machine Learning (ICML)*, 2000.

[16] D. Sadigh, S. S. Sastry, S. A. Seshia, and A. Dragan, "Information gathering actions over human internal state," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016.

[17] D. Hadfield-Menell, A. Dragan, P. Abbeel, and S. Russell, "Cooperative inverse reinforcement learning," in *Advances in Neural Information Processing Systems*, 2016.

[18] S. Reddy, A. D. Dragan, and S. Levine, "Where Do You Think You're Going?: Inferring Beliefs about Dynamics from Behavior," in *Advances in Neural Information Processing Systems 31 (NIPS 2018)*, 2018.

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

[20] J. Pöppel and S. Kopp, "Satisficing models of bayesian theory of mind for explaining behavior of differently uncertain agents: Socially interactive agents track," in *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems*, 2018.

[21] M. C. Buehler and T. H. Weisswange, "Online inference of human belief for cooperative robots," in *International Conference on Intelligent Robots and Systems (IROS)*, 2018.

[22] W. Yoshida, R. J. Dolan, and K. J. Friston, "Game Theory of Mind," *PLoS Comput Biol*, vol. 4, no. 1210, 2008.

[23] S. Ross, J. Pineau, S. Paquet, and B. Chaib-draa, "Online Planning Algorithms for POMDPs," *J. Artif. Intell. Res.*, vol. 32, 2008.

[24] A. Dragan and S. Srinivasa, "Integrating human observer inferences into robot motion planning," *Autonomous Robots*, vol. 37, no. 4, 2014.

[25] R. Lowe, Y. WU, A. Tamar, J. Harb, O. P. Abbeel, and I. Mordatch, "Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments," in *Advances in Neural Information Processing Systems*, 2017.

[26] T. Chakraborti and S. Kambhampati, "Algorithms for the Greater Good! On Mental Modeling and Acceptable Symbiosis in Human-AI Collaboration," 2018.