Towards We-intentional Human-Robot Interaction using Theory of Mind and Hierarchical Task Network

Maitreyee[a] and Michele Persiani[b]
Department of Computing Science, Umeå University, Universitetstorget 4, Umeå, Sweden

Keywords: We-intention, Joint Intention, Joint Activity, Human-Robot Interaction, We-mode, I-mode, Hierarchical Task Network, Dialogue Interaction.

Abstract: Joint activity between human and robot agent requires them to not only form joint intention and share a mutual understanding about it but also to determine their type of commitment. Such commitment types allows robot agent to select appropriate strategies based on what can be expected from others involved in performing the given joint activity. This work proposes an architecture embedding commitments as we-intentional modes in a belief-desire-intention (BDI) based Theory of Mind (ToM) model. Dialogue mediation gathers observations facilitating ToM to infer the joint activity and hierarchical task network (HTN) plans the execution. The work is ongoing and currently the proposed architecture is being implemented to be evaluated during human-robot interaction studies.

1 INTRODUCTION

For successful joint activities, human and a robot agent take a we-intentional stance (Tuomela and Miller, 1988; Tuomela, 2005) aiming to act on a joint intention. In a day-to-day scenario such interactions about medication reminder, taking care of health and finding objects in the house may require participants to sometimes commit to a shared goal and other times pursue their private ones.

An example of a prior case is when the human and a robot agent deliberate by deciding on how to find an object like newspaper, reading glasses or their cell-phone, prepare a meal, or plan and clean the house together. The instance for the latter case would be when robot persuades human agent to take scheduled medication or manage daily physical well-being by rating any discomfort, taking some rest, or booking an appointment to seek medical care.

The research has already discussed the content required (Clodic et al., 2014; Levine and Williams, 2018; Vinanzi et al., 2021) for human-robot agent joint actions. However, another factor that has not received much attention is of the stance (Schweikard and Schmid, 2020) an agent should have to perform the joint activity (Grynszpan et al., 2019; van der Wel, 2015). A research question to ask for such a case is *How can agent's stance be modeled for deliberation and persuasion during a joint activity and what kind of mediation will be required to achieve it?*

To address the mentioned research gap we examine here different modes of we-intention (Tuomela and Miller, 1988; Tuomela, 2006) determining the stance of the agents with-in a theory of mind (ToM) framework. Furthermore, we use dialogue communication to increase the information that the robot has available when computing its ToM model. Finally, through hierarchical task networks (Georgievski and Aiello, 2015) we create probability distributions of plans to be executed as legible or optimal depending on the calculated information gain.

The proposed architecture requires a model of shared cognition and mind (Thellman and Ziemke, 2019; Bianco and Ognibene, 2019) allowing robot agent to attribute behavior to mental attitudes of humans or other artificial agents, similar to the concept from Theory of Mind (ToM) (Scassellati, 2002). In robotics ToM reasons the state of mind of other involved agents at different dispositions, where two most applicable ones' are first and second order (Hellström and Bensch, 2018). From Robot's stand point first order is robot's reasoning of the human agent and second order is robot's reason of what human reasons about the robot. Humans can be considered as rational agents with mental attitudes composed of their knowledge about the world, the motivation and their

[a] https://orcid.org/0000-0002-3036-6519
[b] https://orcid.org/0000-0001-5993-3292

Maitreyee, . and Persiani, M.
Towards We-intentional Human-Robot Interaction using Theory of Mind and Hierarchical Task Network.
DOI: 10.5220/0010722200003060
In *Proceedings of the 5th International Conference on Computer-Human Interaction Research and Applications (CHIRA 2021)*, pages 291-299
ISBN: 978-989-758-538-8; ISSN: 2184-3244

goals required to mediate in their complex and unpredictable environment.

A belief-Desire-Intention (BDI) (Rao et al., 1995) model can be used to represent agents with mental attitudes mentioned before. In this work, the agents are represented as BDI models on which the robot agent computes first and second order ToM. Interaction between robot and a human agent can be structured by a shared task domain. For example, in a health-care scenario a possible task is of taking pills where the robot should bring pills and/or water to human owner. It is therefore necessary that both human and robot comply with a structure both in terms of tasks and interactions that it allows. In this context we study how a robot agent can form a we-intention (Tuomela and Miller, 1988) with a human.

A we-intention is a type of aim intention, where the agents share a set of actions to be performed aimed by all the involved participants. We-intentional agents form a joint intention to perform an action together, namely, a joint action, in two different modalities of; we-mode when the agents act as group members "committed to collectively agreed upon goals" and i-mode when agents are part of a group "committed to their own private goal" (Tuomela, 2006; Tuomela, 2005). These two modes imply legible or optimal mechanisms for joint activity. Our contribution implies mechanisms for robot agent to create BDI based ToM for executing joint activity with humans. Which is realized by dialogue based observation gathering and planning of shared task using hierarchical task network.

Following Section 2 briefly introduces theory of mind, motivates we-intentions for joint activity and concludes with some related work. Then, Section 3 provides the formal architecture created for robots BDI ToM, dialogue mediation, task execution and hierarchical task network. The article concludes with an illustrative example and ongoing work in Section 4 and Section 5.

2 BACKGROUND AND RELATED WORK

This section provides relevant interpretation of Theory of Mind (ToM) and stance/commitment in We-intention (Tuomela and Miller, 1988) for the purpose of understanding this work.

2.1 Theory of Mind

A field recently being investigated to create models for belief reasoning is theory of mind (ToM) (Scassellati, 2002; Thellman and Ziemke, 2019). Theory of mind relates to the ability of agents to attribute mental states and beliefs to themselves or other agents, and of creating a point of view of a situation in terms of beliefs, goals and intentions that is different from their own but rather belonging to others. A first order theory of mind is expressed in the sentence "Bob thinks that Alice thinks X", or in other words Bob has an estimate of Alice's mental state, believing she's thinking X. Higher order theories deepen these levels of reasoning by extending the thinking chain. A second order reasoning would be "Carl thinks that [Bob thinks that Alice thinks X]"—with parenthesis added to highlight the recursion. In this case Carl holds an estimate of Bob's mental state. Arbitrary higher orders of reasoning follow the same incremental structure.

Theory of mind is considered to be at the cornerstone of shared intentionality (Tomasello et al., 2005). Human beings are able to understand each other and collaborate thanks to their innate will of sharing intentions and psychological states.

2.2 We-intentional Stance

Intentions that agents' mind create in order to perform actions and achieve goals as a collective can be defined and studied under the umbrella term of 'collective intentionality'. With explicit discussions in Phenomenology, Existential Philosophy and Sociological Theory collective intentionality gained attention in practical philosophy when introduced by Sellars (Wilfrid, 1980) as 'we-referential intentions' or simply called 'we-intentions'. He argues we-intention to have the following characteristics: *(a)* to be attitudes of individuals, *(b)* to involve a non-parochial attitude towards group the individuals have, and *(c)* that there are two types of intentions– primary i-intention and we derivative i-intention to joint intention.

Joint intention can be expressed as a case of performing joint actions by agents with joint agency (Tuomela and Miller, 1988), we-intention and a shared mutual belief. Central tenet of we-intention entails a participation intention to perform a joint action requiring agents to jointly intend to participate action-ally, that is, contributing to performing of joint actions. A we-intending agent has a rational belief that it can perform its part of a joint action A with some probability, and that it can perform A collectively with other agents with some non-zero probability.

Attitudes of we-intending agents allows joint intention to exist in either– I-mode or we-mode for a group as illustrated next.

2.2.1 We-mode of We-intention

An agent in we-mode, functions as a group member satisfying and committed to *group's* goal g in the joint intention by participating in joint action with the set of requirements (Tuomela, 2006): (a) agent A_i intends to do his part of g (b) agent A_i has a belief that other agents in the group will perform their parts of g and (c) that agent A_i beliefs that there is a mutual belief in the group that g will be effectuated to achieve the joint intention.

2.2.2 I-mode of We-intention

An agent pursuing i-mode, functions as part of a group satisfying the groups goal g being committed to its *private* goal pg in the joint intention by participating in joint action with the set of requirements (Bratman, 1993): (a) agent A_i intends to satisfy g, (b) agent A_j intends to satisfy g (c) A_i and A_j do so by meshing up their sub-plans of their private goal pg and (d) they have a mutual belief about (a), (b), and (c).

2.3 Joint Intention and We-intention in Human Robot Interaction (HRI)

(Clodic et al., 2014) highlights the components required for joint activities between human and robot. Their work discusses low level processing to infer intentions of other agents, to jointly direct the attention and to form a shared mutual belief about task in hand.

Some of the previous work (Rauenbusch and Grosz, 2003; Kamar et al., 2009) proposed team rationality for building collaborative multi-agent systems, for example, in (Rauenbusch and Grosz, 2003) the authors used Shared Plan (Grosz and Kraus, 1996) and *Propose Trees* to model collaboration as multi-agent planning problem, where a rational team will perform an action only if the benefits from performing an action is less than its cost.

Authors in (Devin et al., 2017) illustrated a block world scenario for human and robot agent to collaborate on a shared plan, with the focus on flexibility and non-intrusiveness. Extension of HATP (Lallement et al., 2014) (a hierarchical task network planner for human robot agent shared planning) was proposed with conflict resolution mechanism to account for flexible plan interaction and minimal dialogue interaction for non-intrusiveness. Pike (Levine and Williams, 2018) was introduced for robot agent to concurrently infer human intention and adapt to it. The robot agent reasons on specific combinations of controllable (actions that robot agent can choose to do) and uncontrollable (actions performed by the human, environment or other agents in the environment) possibilities for effective plans to jointly achieve a task. The combinations are constrained by ordering of actions, time constraint and unpredictable situations. A temporal plan network generates real-time plans for execution. Authors in (Vinanzi et al., 2021) assume trust and intention recognition as essential to collaborative joint activities with mutual understanding, interaction and coordination. Authors in this work differed from most of the research to determining robot agent's trust in human partner in performing joint activities. Their work performs simulated evaluation of learning and probabilistic cognitive architecture to perform joint activity by robot agent estimating intention of and trust in human agent.

Literature suggests that most of the work concerning joint activity in HRI models relate to intention recognition and human-aware planning, i.e., what should the robot must do at perception and action level without distinguishing the way in which a joint activity could be pursued. Some work relating to such modality has been investigated within the context of *agency*. During joint activity between humans the author in (van der Wel, 2015) studies the effect of we-mode processing of joint actions on the sense of control in participating human agents. The author studies dyads of human agents during a joint action with different roles validating we-mode stance with a high sense of control and performance of ones own as well as of the other. However, online intention negotiation reduced sense of control and shifted the stance of the dyad to i-mode.

3 METHODOLOGY

We assume for the robot to be already have an awareness of some of the goals that human might want to pursue as a joint activity with the robot. In a day-to-day health-care scenario human might want the robot to; remind them about taking medication on time, to help the human find objects such as newspaper, reading glasses or their cell phone, to make human aware of any persisting health issue they have been ignoring and to discuss possible solutions. Either of the human or the robot may initiate a dialogue pursuing a goal with a joint activity by first facing the prospective participant. The initiating agent then proposes a joint intention which could be co-created by other agent resulting into a we-/I-mode behavior or breakdowns because the other agent might be engaged in or prioritizes another activity or lacks the knowledge and/or functions to participate. To this end we only

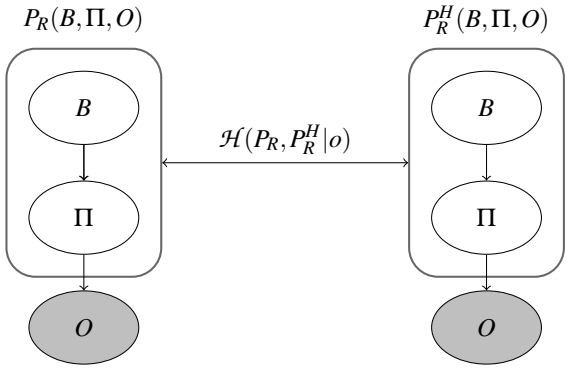

Figure 1: Robot model and second-order theory of mind as equivalent Bayesian Networks. The cross-entropy H measures the difference between robot's and user's estimated state of mind a posteriori of the observations o gathered through the dialogue with the user.

formulate co-creation of joint intention. Co-creation of joint intention is realized with a ToM where, the robot agent computes an information gain between its own and human agent's belief-desire-intention (BDI) models. The information gain then directs the agent to carryout the plan legibly or optimally.

Next sections illustrates the procedure robot uses to (1) create its ToM model; (2) gather observations by dialogue mediation; (3) regularize an HTN plan constrained by human expectation; and (4) execute the plan legibly or optimally.

The formal procedure in the text uses following symbols: P_R denotes probability distribution for robot, P_R^H is probability distribution on human computed by the robot. $b \in B$ denotes set of beliefs, G is the goal and $o \in O$ is the set of observations gathered from the dialogue interaction. $\pi \in \Pi$ denotes the set of task networks. $\mathcal{H}$ is used here to denote cross-entropy between human and robot's intentions.

3.1 Theory of Mind- Human and Robot as BDI Models

The general architecture to plan and executes joint tasks in we-mode is shown in Figure 1. It is composed of two Belief-Desire-Intention (BDI) models in a configuration that describes a second-order theory of mind between robot and human agent. On the left side is the BDI model that the robot uses to plan the interaction. Its joint distribution is:

$$P_R(B,\Pi,O) = P_R(O|\Pi)P_R(\Pi|B)P_R(B) \qquad (1)$$

where $P_R(B)$ is the probability distribution of beliefs and desires (here put together for simplicity). Sampling from $P_R(B)$ yields a candidate belief and desire of the robot, that it can use to compute its intention. $P_R(\Pi|B)$ is the intention model, that computes intentions in the form of plans. Sampling $P_R(\Pi, B = b)$ yields a candidate robot plan that is consistent with b. Finally, $P_R(O|\Pi)$ is the observation model, that provides the robot knowledge of how its intentions are perceived, either in the form of a dialogue or physical actions.

On the right side is instead the second-order theory of mind of the human agent that the robot uses to compute his expectations on the interaction. Again, the theory of mind is a BDI model and defined as P_R. Its joint probability distribution is defined as:

$$P_R^H(B,\Pi,O) = P_R^H(O|\Pi)P_R^H(\Pi|B)P_R^H(B) \qquad (2)$$

For simplicity, we set P_R and P_R^H as equivalent Bayesian networks. This relies on the assumption of having a correct human agent model and that it matches the model used by the robot, that can be achieved by training the human agent before the interactions. Notice that while the BDI models are structurally the same, the random variables could be differently distributed between them, allowing to model uncertainty in the estimate of the human agent's beliefs or intentions.

Crucial to the functioning of the proposed theory of mind model is the distance measure $\mathcal{H}(P_R,P_R^H)$, that the robot uses to compute the distance between its own and the human agent's probabilistic BDI models. We refer to this degree of matching between the two models as the general *understanding* that the human agent has of the interaction (Hellström and Bensch, 2018).

A we-mode interaction implies that the human agent is consistently understanding what's happening during the interaction i.e the state of mind of the interaction. Therefore, to operate in we-mode from the robot side, means to perform in a way that maintains ToM to be "synchronized". This can for example be achieved by producing observations increasing the understanding of the interaction. In this context, the Information Gain (IG) that observations produce is:

$$IG(P_R,P_R^H|o) = \mathcal{H}(P_R,P_R^H) - \mathcal{H}(P_R,P_R^H|o) \qquad (3)$$

Information gain will be used to guide the execution of robot's plan during the two main phases of interactions, namely the dialogue phase, where robot and human agent mediates the task (who will do what) of their joint activity, and the execution phase, where they jointly perform the agreed upon activity.

3.2 Dialogue Interaction and Gathering Observations

In order to mediate what they will do, human agent and robot perform a dialogue to decide on a set of commitments. These commitments are expressed in the form of action frames that are used to constrain the execution of the planner. Table 2 shows an example dialogue and how utterances are associated through action frames.

As proposed in previous research (Persiani and Hellström, 2020), we utilize Semantic Role Labeling to classify action frame from utterances. For example, an utterance "Can you remind me to take my pills in fifteen minutes?" could get parsed into the semantic frame *verb*: remind, *object*: pills, *recipient*: me, while a classifier maps the semantic frames to a corresponding task. For example, if the task domain defines an action frame **remind ?agent ?recipient ?object ?time**, the classifier could map the utterance to a partially specified action frame **remind robot0 human0 pills0 15minutes**, where **robot0 human0 pills0 15minutes** are identifiers belonging to the planning instance. Not all arguments of the action frame are to be specified, and the missing ones are found during the subsequent inference steps.

We utilize a finite state machine of dialogue acts representing the utterances in each turn to model a dialogue scenario about medication as illustrated in 2. A set of dialogue acts were selected for the purpose of this work from (Harry et al., 2017)–*Greeting, GoodBye, Question, Offer, CounterOffer, AcceptOffer, RejectOffer, Inform.* To organize the utterances such that they fulfill the joint activity, Table 1 contains the effects of dialogue acts with respect to three sets: θ is a set of questions, inform or offers R and O are respectively the sets of rejected and accepted commitments. Either of the agents can initiate a dialogue with a greeting and end it with a goodbye, without any effects on commitments. A *question* or an inform instantiates an action frame that one of the agent wants to achieve during a joint activity. An *offer* provides an agent with a possible action that can be performed during the joint activity. A *counter offer* is an action a_1 not accepted and an alternative a_2 is instead proposed. An *Accept* and *Reject* can be used to accept or reject proposed commitments.

At the end of the dialogue the set $o = \{a_0, ..., a_n\}$ identifies the action frames accepted through the dialogue. This set is utilized to constrain the planning procedure as we will next describe.

3.3 Task Execution: Legibly or Optimally

After the dialogue the set of mediated action frames are utilized for constraining inference using the theory of mind architecture. Two main type of inference are required for obtaining a we-mode interaction: intent recognition, that from the action frames finds the most likely intention being expressed; and legible behavior, which uses the found intention to change the original robot plan toward more legible versions.

Intent recognition is performed by solving the following equation:

$$g_H, \pi_H = \underset{b \in B, \pi \in \Pi}{argmax} P_R^H(b, \pi | o) \propto P_R^H(b, \pi, o) \quad (4)$$

that corresponds to finding the most likely goal and plan that the second-order theory of mind yields while being constrained by o. In other words, the human agent intention is found by simulating its model under the constraints.

Legible behavior is obtained by making the robot plan consistent to human agent's intention. This is expressed by the following equation:

$$b_R, \pi_R = \underset{b \in B, \pi \in \Pi}{argmax} P_R(\pi | o) + \alpha \mathcal{H}(P_R^H(b_H, \pi_H) \parallel P_R(b, \pi)) \quad (5)$$

The right part regularizes the planner towards intentions that are expected by the human agent model, ie. similar to g_H and π_H.

3.4 Implementation through Hierarchical Task Network

We implemented the BDI models $P_R(O, \Pi, B)$ and $P_R^H(O, \Pi, B)$ by specifying planning instances using the Hierarchical Task Network formalism (HTN). A planning instance for the robot is obtained by specifying the tuple $\langle \mathcal{P}_R, \mathcal{T}_R, \mathcal{M}_R, \mathcal{A}_R, I_R, \mathcal{G}_R, O_R \rangle$. Where I_R is the set of ground predicates in the initial state, $\mathcal{G}_R$ the goal task to be decomposed into primitive actions, O_R is the set of objects available to ground the available predicates $\mathcal{P}_R$, $\mathcal{T}_R$ is the set of tasks available to the planner, $\mathcal{M}_R$ the set of methods to decompose tasks in sub-tasks, $\mathcal{A}_R$ the set of available primitive actions. Refer to (Bercher et al., 2014) for a complete description of HTN formalism. Similarly, the second order theory of mind model has components $\langle \mathcal{P}_R^H, \mathcal{T}_R^H, \mathcal{M}_R^H, \mathcal{A}_R^H, I_R^H, \mathcal{G}_R^H, O_R^H \rangle$.

To implement the equivalent Bayesian networks if Figure 1 we set the descriptive components of the planning instances of robot and theory of mind to be

Table 1: Dialogue acts for the SDS that allows mediation of actions and goals with respect to the sets of offered, rejected and accepted commitments.

Dialogue Act	Precondition	Effect
$Question, x, \hat{g}$	$\emptyset$	$g = \hat{g}$
$Offer, x, a$	$a \notin \theta \wedge a \notin R \wedge a \notin R$	$a \in \theta$
$CounterOffer, x, a_1, a_2$	$a_1 \in \theta\ a_2 \notin R \wedge a_2 \notin O$	$a_1 \notin \theta \wedge a_2 \in \theta$
$Accept, x, a$	$a \in \theta$	$a \notin \theta \wedge a \in O$
$Reject, x, a$	$a \in \theta$	$a \notin \theta \wedge a \in R$
$Inform, x, a_1 ... a_n$	$a_1 ... a_n \in \theta$	$a_1 ... a_n \notin \theta \wedge a_1 ... a_n \in O$

Table 2: Robot and human mediating joint intention about medication committed to mutually agreed upon group goal (we-mode).

	Sentence	Speech act	Action frame
H	Hello Robo	Greeting	
H	Can you remind me to take my pills in fifteen minutes?	Question	**remind robot0 human0 pills 15minutes**
R	Yes sure!	Accept	
R	Do you want me to bring you some water as well?	Offer	**bring robot0 human0 water ?when**
H	Yes, that would be great.	Accept	
R	Okay! I will remind you to take pills in fifteen minutes and bring you some water.	Inform	
H	Thanks a lot!	GoodBye	
R	No problem!	GoodBye	

equivalent. ie. $\mathcal{P}_R = \mathcal{P}_R^H, \mathcal{T}_R = \mathcal{T}_R^H, \mathcal{M}_R = \mathcal{M}_R^H, \mathcal{A}_R = \mathcal{A}_R^H$ and $O_R = O_R^H$, with the only probabilitic parts being I_R, $\mathcal{G}_R$, I_R^H and $\mathcal{G}_R^H$. We can realize the probability distribution over the possible HTN instances describing the robot state through a combination of Bernoulli distribution for the beliefs I_R, and a categorical distribution for the possible goals $\mathcal{G}_R$ (the same for I_R^H and $\mathcal{G}_R^H$ respectively).

$$P_R(B) = P_R(I)P_R(G) \tag{6}$$

$$\begin{gathered} P_R(I;\theta_R) = \Pi_i P(p_i \in I_R;\theta_{p_i}) \\ P(p_i \in I_R) = \theta_i \end{gathered} \tag{7}$$

$$\begin{gathered} P_R(G;\theta_R) = P(G|\{g_0,...,g_m\}) \\ P(G = \langle \mathcal{G}_R \rangle | \{g_0,...,g_m\}) = \theta_j \\ \sum_j \theta_j = 1 \end{gathered} \tag{8}$$

Sampling a belief from $P_R(B)$ yields a initial state and a goal task for the HTN planner. The planning model $P_R(\Pi|B)$ is implemented by a planner of choice compatible with the underlying HTN requirements. For our experiments we are using the PANDA(Bercher et al., 2014) planner.

As also commonly proposed in research on planning, we model the human agent as expecting rational behavior from the robot ie. by associating higher probabilities to cheap plans. Therefore, the probability of a plan $P_R(\Pi = \pi|B = b)$ is defined as a function of rationality:

$$P_R(\Pi = \pi|B = b) = \alpha exp\{\tau(|\pi_{opt,b}| - |\pi|)\} \tag{9}$$

where $|\pi_{opt,b}|$ is the length of an optimal plan for a belief b, while $|\pi|$ the length of the considered plan. α is a normalizing constant, τ a temperature parameter. Eq. 9 gives high likelihood to plans with lower cost, with maximum likelihood associated to the plans of the same cost of the optimal plan. Sampling from the planning model can for example be done through Diverse Planning techniques (Katz and Sohrabi, 2020). $P_R^H(B)$ and $P_R^H(\Pi|B)$ are similarly defined.

4 ILLUSTRATIVE EXAMPLE AND ONGOING WORK

To better show the functioning of the proposed architecture we present an illustrative scenario with Examples 3 and 4. We assume a household scenario, where a robot agent cohabits with human supporting him in daily tasks. Among these tasks, the robot could be asked to perform the joint activity of cleaning the house, where they plan and execute the tasks to; clean windows, mirrors and other surfaces, vacuum the floor, or organize the house.

Depending on human's expectation from the robot, the proposed scenario can unfold into follow-

Table 3: Robot and human mediating joint intention about cleaning the house committed to mutually agreed upon group goal (We-intention in we-mode)

	Sentence	Speech act	Action frame
H	Hello	Greeting	
H	Can you help me with cleaning the house now?	Question	**clean robot0 house0 now**
R	Yes sure! What would you like me to do?	Accept	
H	I would like you to go and vacuum the room once I have dusted their surfaces.	Offer	**vacuum robot0 room0 floor afterDusting**
R	Okay!	Accept	
R	I will vacuum the floors once you have finished dusting the room. Is that correct?	Inform Question	
H	Yes, Thanks.	Accept	
R	No Problem.	GoodBye	

ing two dialogues shown in Table 3 and Table 4. Where, in the first one human asks the robot to help him clean the house by performing one task of vacuuming the rooms given the condition that robot should vacuum each room after human has cleaned the surfaces of that room. In the second case, human instructs the robot to perform the tasks of cleaning surfaces and vacuuming the two rooms (living and guest room) in the house on its own, while he could clean the other rooms. A set of commitments for the corresponding joint activity (cleaning the house) is created using the action frames for **vacuum** action in Example 3 and, **vacuum** and **clean surface** actions in Example 4. Depending on the dialogue scenario that human chose, action frames will be used to guide the execution of the joint activity in the following ways: firstly, as constraints while computing the intention of the human. The human's intention is computed to know what the human expects (we-mode in Example 3 and i-mode in Example 4) from the joint activity. Since in the example dialogues the human uses *Offer* speech act to set the goal task, what remains to compute is the plan part of the intention. To do so the human model P_R^H is simulated giving the set of action frames as constraints (Eq. 4) ie. only intention containing the mentioned action frames are considered.

After finding the human intention the robot should compute its own intention. It can perform the activity either in I-mode or in we-mode. Robot computes optimal plan for operating in I-mode for the dialogue Example 4 while the plan is regularized to match the expectation of the human in dialogue Example 3. Such plans for this work is created using HTN as illustrated in the previous Section 3.4.

This work reports on ongoing research and currently we are implementing the different components for realizing the proposed architecture from Activity Theoretical perspective (Vygotsky, 1978; Kaptelinin and Nardi, 2006; Leontiev, 1978). Activity theory facilities the study of human activity at the intersection of human psychology, social relationships and institutions (Roth, 2014). We modify the proposed BDI model in our architecture with elements–(subject, object, tool, division of labor, rules, and community) from Engeström's Activity System Model (ASM) (Engeström, 1999), for dialogues between robots and humans in a health-care setting. As a first step of validation, a qualitative study was conducted with 20 human participants, who watched recordings of human and robot interacting in a Wizard of Oz setup. The study aimed to understand how human's perceived HRI interactions for joint activity in we-/I-mode of we-intention. Qualitative results highlighted that behavior, role, relationship and dialogue strategies were perceived as intuitive, intelligent and human-like, however, the robot was expected to adapt and improve to have more empathetic behavior. The results from our qualitative study will now be used to inform the development of our architecture.

5 CONCLUSIONS

With this position paper we proposed a theory-of-mind architecture for joint human-robot activities. Joint activities are first mediated through a dialogue, allowing human and robot to specify what will be their commitments inside the joint activity. Then, from the robot's perspective, a joint activity can be executed in I-mode or we-mode respectively. While operating in I-mode, the robot committed towards its private goal executes the joint activity in the most optimal way given its capabilities, and without considering the fact that it is collaborating with a human. This can be desirable is some scenarios, such as when robot and human are committed to the same goal operating in we-mode instead, a legibility requirement is introduced to make the robot plan similar to the hu-

Table 4: Robot and human mediating joint intention about cleaning the house committed to their own private goals (we-intention in i-mode).

	Sentence	Speech act	Action frame
H	Hi, Can you help me with cleaning the house now?	Greeting and Question	**clean robot0 human0 house now**
R	Yes sure! What would you like me to do?	Accept	
H	I would like you to dust the surfaces and vacuum floors of the living room and guest room, while.	Offer	**dust,vacuum robot0 room0 living room1 guest**
R	Yes sure!	Accept	
R	I will clean the living and guest room.	Offer	
H	Thanks, that would be great.	Accept	
R	No problem!	GoodBye	

man's expectations.

We proposed the formalization of the architecture, together with an illustrative example showing its function in a simple scenario where human and robot mediate the task of bringing the human medications. Future work will conduct Human Robot Interaction (HRI) studies to evaluate the proposed architecture.

REFERENCES

Bercher, P., Keen, S., and Biundo, S. (2014). Hybrid planning heuristics based on task decomposition graphs. In *Seventh Annual Symposium on Combinatorial Search*.

Bianco, F. and Ognibene, D. (2019). Transferring adaptive theory of mind to social robots: Insights from developmental psychology to robotics. In *Social Robotics*, pages 77–87. Springer International Publishing.

Bratman, M. E. (1993). Shared intention. *Ethics*, 104(1):97–113.

Clodic, A., Alami, R., and Chatila, R. (2014). Key elements for joint human-robot action. In *Robo-Philosophy*, volume 273 of *Sociable Robots and the Future of Social Relations*, pages 23–33, Aarhus, Denmark. IOS Press Ebooks.

Devin, S., Clodic, A., and Alami, R. (2017). About decisions during human-robot shared plan achievement: Who should act and how? In *Social Robotics*, pages 453–463, Cham. Springer International Publishing.

Engeström, Y. (1999). Activity theory and individual and social transformation. *Perspectives on activity theory*, 19(38):19–38.

Georgievski, I. and Aiello, M. (2015). Htn planning: Overview, comparison, and beyond. *Artificial Intelligence*, 222:124–156.

Grosz, B. and Kraus, S. (1996). Collaborative plans for complex group action. *Artificial Intelligence*.

Grynszpan, O., Sahaï, A., Hamidi, N., Pacherie, E., Berberian, B., Roche, L., and Saint-Bauzel, L. (2019). The sense of agency in human-human vs human-robot joint action. *Consciousness and cognition*, 75:102820.

Harry, B., Volha, P., David, T., and Jan, A. (2017). *Dialogue Act Annotation with the ISO 24617-2 Standard*, pages 109–135. Springer International Publishing.

Hellström, T. and Bensch, S. (2018). Understandable robots - what, why, and how. *Paladyn, Journal of Behavioral Robotics*, 9(1):110–123.

Kamar, E., Gal, Y., and Grosz, B. J. (2009). Incorporating helpful behavior into collaborative planning. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. Springer Verlag.

Kaptelinin, V. and Nardi, B. A. (2006). *Acting with Technology: Activity Theory and Interaction Design*. The MIT Press, USA.

Katz, M. and Sohrabi, S. (2020). Reshaping diverse planning. In *AAAI*, pages 9892–9899.

Lallement, R., de Silva1, L., and Alam, R. (2014). Hatp: An htn planner for robotics. In *Planning and Robotics (PlanRob), ICAPS Workshop*, volume 2, page 8, Portsmouth, USA. AAAI Press.

Leontiev, A. N. (1978). *Activity, consciousness, and personality.* Prentice-Hall, Moscow, Russia.

Levine, S. J. and Williams, B. C. (2018). Watching and acting together: Concurrent plan recognition and adaptation for human-robot teams. *J. Artif. Int. Res.*, 63(1):281–359.

Persiani, M. and Hellström, T. (2020). Intent recognition from speech and plan recognition. In *International Conference on Practical Applications of Agents and Multi-Agent Systems*, pages 212–223. Springer.

Rao, A. S., Georgeff, M. P., et al. (1995). Bdi agents: from theory to practice. In *ICMAS*, volume 95, pages 312–319, USA. MIT Press.

Rauenbusch, T. W. and Grosz, B. J. (2003). A decision making procedure for collaborative planning. In *Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems*, AAMAS '03, page 1106–1107. Association for Computing Machinery.

Roth, W.-M. (2014). *Activity Theory*, pages 25–31. Springer New York, New York, NY.

Scassellati, B. (2002). Theory of mind for a humanoid robot. *Autonomous Robots*, 12:13–24.

Schweikard, D. P. and Schmid, H. B. (2020). Collective Intentionality. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2020 edition.

Thellman, S. and Ziemke, T. (2019). The intentional stance toward robots: conceptual and methodological considerations. In *The 41st Annual Conference of*

the Cognitive Science Society, July 24-26, Montreal, Canada, pages 1097–1103.

Tomasello, M., Carpenter, M., Call, J., Behne, T., and Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, 28(5):675–691.

Tuomela, R. (2005). We-intentions revisited. *Philosophical Studies*, 125(3):327–369.

Tuomela, R. (2006). Joint intention, we-mode and i-mode. *Midwest Studies In Philosophy*, 30(1):35–58.

Tuomela, R. and Miller, K. (1988). We-intentions. *Philosophical Studies*, 53(3):367–389.

van der Wel, R. P. (2015). Me and we: Metacognition and performance evaluation of joint actions. *Cognition*, 140:49–59.

Vinanzi, S., Cangelosi, A., and Goerick, C. (2021). The collaborative mind: intention reading and trust in human-robot interaction. *iScience*, 24(2):102130.

Vygotsky, L. S. (1978). Mind in society: The development of higher mental processes (e. rice, ed. & trans.).

Wilfrid, S. (1980). On reasoning about values. In *American Philosophical Quarterly*, volume 17, page 81–101. JSTOR.