Endowing Robots with Self-Modeling Abilities for Trustful Human-Robot Interactions

Cristiano Castelfranchi*, Antonio Chella†, Rino Falcone*, Francesco Lanza†, Valeria Seidita †
*Istituto di Scienze e Tecnologie della Cognizione (ISTC-CNR),
cristiano.castelfrancchi@istc.cnr.it, rino.falcone@istc.cnr.it
†Università degli Studi di Palermo,
antonio.chella@unipa.it, francesco.lanza@unipa.it, valeria.seidita@unipa.it

***Abstract*—Robots involved in collaborative and cooperative tasks with humans cannot be programmed in all their functions. They are autonomous entities acting in a dynamic and often partially known environment. How to interact with the humans and the decision process are determined by the knowledge on the environment, on the other and on itself. Also, the level of trust that each member of the team places in the other is crucial to creating a fruitful collaborative relationship. We hypothesize that one of the main components of a trustful relationship resides in the self-modeling abilities of the robot. The paper illustrates how employing the model of trust by Falcone and Castelfranchi to include self-modeling skills in the NAO humanoid robot involved in trustworthy interactions. Self-modeling skills are then implemented employing features by the BDI paradigm.**

***Index Terms*—Human-Robot Interaction; Trust; Multi-agent systems; BDI; JASON**

I. Introduction

Human-robot interaction (HRI) is the discipline investigating how to analyze and develop robots that interact with humans to pursue a common objective. Interaction is the process of working together to reach a goal and it can be viewed from different points of view and has various forms, from direct command and clear response to the ability of autonomously decide how to pursue a goal. Every robot applications present some kind of interactions with humans through explicit or implicit communications. In the case of autonomous robots operating as a teammate towards humans, humans provide the goal and the robot has to be able to maintain knowledge about the environment and the tasks to perform in order to decide whether adopting or delegating a task or an action.

Autonomy, proactivity, and adaptivity are the features to decide, at each moment, which activity has to be fruitfully performed for efficiently pursuing an objective. From a cooperative and social point of view - human-robot team interaction - this means to be able to decide which action to perform by itself and which one to delegate to another component of the team.

This decision cannot be imposed during the design process, for many reasons ranging from the composition of the environment to the characteristics of the interacting entities. The environment is always strongly dynamic and often unknown.

In the case of a team composed of only humans, the interaction with a teammate is based on the level of knowledge owned on the environment and on the "other". Especially, knowledge about the capabilities of the other, about the interpretation of the actions of the other concerning the shared goals and therefore also about the level of trust that is created towards the other. Trustworthiness is a parameter to be used for letting an entity decide which action to adopt or which to delegate.

In our work, we are analyzing the role of trust in the human-robot interactions and the integrated function of self-modeling and theory of mind for implementing human-robot interactions based on trust. In this paper, we focus on how to implement self-modeling in the NAO robot employing the BDI (belief, desires, intention [15] [3]) agent paradigm and the JASON framework [2] [1].

The final goal of our work is to implement interactions in teams of humans and robots so that collaboration is as efficient and reliable as possible. To do this, both entities involved in the interaction need to have a certain level of confidence in each other. Measuring trust in the other is made easier if he has full knowledge of his capabilities, or if he can understand his own limitations. The more one of the two entities is aware of its limitations and abilities, the more the other entity can establish a level of confidence and create a productive and fruitful interaction. That is the founding factor of our work.

The idea is to exploit practical reasoning in conjunction with a well-known model of trust [6] [10] to let the robot create a model of its actions and capabilities, hence some kind of self-modeling abilities. We claim that self-modeling is one of the essential components in trust-based interactions. Starting from the BDI practical reasoning cycle, we extend the deliberation process and the belief base representation in a way that allows the robot to decompose a plan in a set of actions strictly associated to the knowledge useful for performing each action. In this way, the robot creates and maintains a model of the "self" and can justify the results of its actions.

Justification is an essential result of self-modeling abilities application and at the same time is a useful means for improving trustful interactions.

The rest of the paper is organized as follows: in section II we illustrate the motivations of our work along with some basic concepts from trust theory and multi-agent systems domain useful for understanding the solution proposed in section III; in section IV we show how we employed our theory in a real case study; in section V we compare our work with some related works and finally in section VI we draw some discussions and conclusions.

II. The Trust Theory and Agents

Trust is a general term to explain what a human has in mind about how to rely on others. In literature, we can retrieve more than one definition of trust. These definitions often are partially or entirely related one with the others.

One of the most accepted definitions of trust is the one by Gambetta [12]: *Trust is the subjective probability by which an individual, A, expects that another individual, B, performs a given action on which its welfare depends.*

Trust is strongly related to the knowledge one has on the environment and on the other. Knowledge of the environment is often the result of some kind of measure of trust. Trust is seen both as a mental state and as a social attitude. Trust is related to the mental process leading to the delegation. The degree of trust is used to rationally decide whether or not to delegate an action to another entity, the classic "on behalf of". It is for this reason that we choose to use agent technology. A software agent [19] [20] is born to act in place of the human; all the theories and technologies about agents are born and have evolved around this pivotal point.

We refer to the work of Falcone and Castelfranchi [6] [10] [11] [8]. In [6] the authors consider:

- *trust* as *mental attitude* allowing to predict and evaluate other agents' behaviors;
- *trust* as a *decision* to rely on in other agent's abilities;
- *trust* as a *behaviour*, or an intentional act of entrusting.

Moreover, in [6], trust is considered as composed of a set of different figures that take part in a trust model:

- the *trustor* - is an "intentional entity" like a cognitive agent based on the BDI agent model that has to pursues a specific goal.
- the *trustee* - is an agent that can operate into the environment.
- the *context* - is a context where the trustee performs actions.
- τ - is a "causal process". It is performed by the trustee and is composed of a couple of act α and result p, g_X is surely included in p and sometimes it coincides with p.
- the *goal* g_X - is defined as $Goal_X(\mathbf{g})$.

The trust function can be defined as *the trust of a trustor agent in a trustee agent for a specific context to perform acts to realize the outcome result*. The trust model is described as a five-part figures relation:

$$TRUST(\text{X Y C } \tau \ g_X) \quad (1)$$

where X is the trustor agent, Y is the trustee agent. X's goal or briefly g_X is the most important element of this model. In some cases, the outcome result can be identified with the goal. For more insights on the *model of trust* and *the trust theory* refer to [6].

In this theory, trust is the mental counterpart of delegation. In the sense that trust denotes a specific mental state mainly composed of beliefs and goals, but it may be realized only through actions. Delegation is the result of a decision taken by the trustor to achieve a result by involving the trustee.

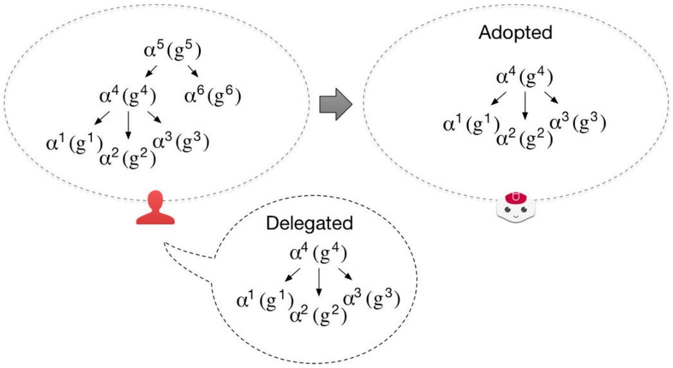


Fig. 1. Level of Delegation/Adoption, *Literal Help*

Several different levels of the delegation have been proposed in [7] and [9], they range from a situation in which the trustor directly delegates the trustee to case in which the trustee autonomously acts on behalf of the trustor.

In our work, we assume an interaction like a continuous operation of adoptions and delegations and we focus only on the *literal help* shown in Fig. 1.

In the literal help, a client (trustor) and a contractor (trustee) act together to solve a problem, the trustor asks the trustee to solve a sub-goal by communicating the trustee the set of actions (plan) and the related result. In the *literal help* approach, the *trustee* strictly adopts all the sub-goals the *trustor* assigns to him [7] [9]. This corresponds to the notion of behaving "on behalf of" that, as said, is one of the key ideas in the multi-agent systems paradigm. Agents' features, such as autonomy, proactivity and rationality are a powerful means that make trust-based agents ideal candidates to be used in applications such as human-robot interaction. By employing the multi-agent paradigm, we may design and develop a multi-agent system in which a certain number of agents is deployed in the robots involved in the application domain.

Our idea is to use the *belief-desire-intention* (BDI) paradigm [3]. The decision-making model underpinning BDI paradigm is known as *practical reasoning*. Practical reasoning is a *reasoning process* for actions, where agents' desires and agents' beliefs supply the relevant factor [4]. The practical reasoning, in human-terms, consists of two activities:

- *deliberation and intentions*;
- *means-ends reasoning*.

Each activity can be expressed as the ability to fix a behavior related to some intentions and deciding how to behave.

All these features of a BDI agent shall faithfully reflect all we need to realize a system based on the trust theory.

Fig. 2 shows the standard practical reasoning cycle of a BDI agent. In the following sections, we illustrate how we changed the reasoning to include self-modeling.

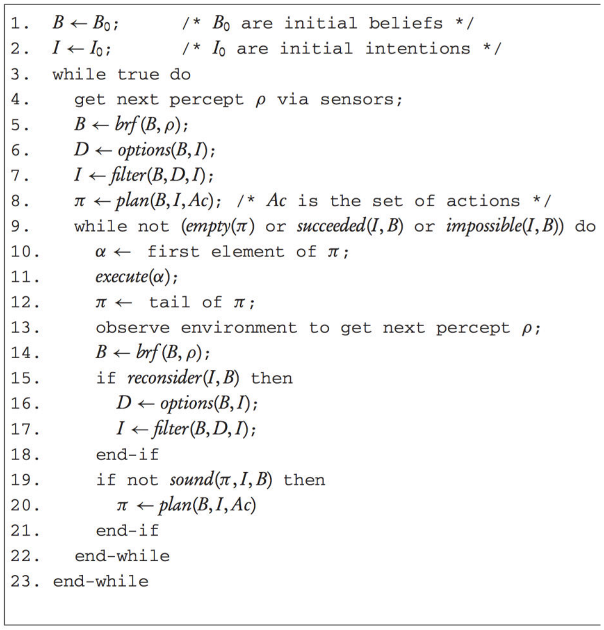
III. Self-modeling using BDI agents

How to design and implement a team of robots that possess a model of themselves, of their actions, behaviors, and abilities? And more, how to allow robots reason about themselves and infer information about their activities, such as why action has failed?

The idea we propose is to use the multi-agent paradigm and the BDI theories and techniques for analyzing trust-based interactions among robots and humans working in a partially unknown environment. We propose to employ the model discussed in [10] [6] and to integrate it with the traditional BDI working cycle [2] (see section II).

For employing this model of trust, we considered the robot as the trustee and the human as the trustor. Assuming that the human delegates a part of his goals to the robot, the level of trust the human has in the robot may be derived from the robot's ability to justify the outcome of its actions, especially in the case of failure. Indeed, self-modeling is the ability to create a model of several features realizing the self. Among them the knowledge of owns capabilities, in the sense that the agent is aware of what it is able to do, and the knowledge on which actions may be performed on every part of the environment. Justifying action is the result of reasoning about actions, it is a real implementation of the self-modeling ability of an agent (human or robot). For doing this, we propose to represent the robot's knowledge through actions and beliefs on those actions.

In particular, we claim that the module containing the justification of an action, or of behavior, should comprise components allowing to reason about the portion of knowledge useful for performing that action. This has to be made for each action of a complete plan. If an action is coupled with all the concepts it needs for being completed then the performer may know at each moment whether and why an action is going wrong and then it may motivate all eventual faults.



```
$B \leftarrow B_0$;        /* $B_0$ are initial beliefs */
$I \leftarrow I_0$;        /* $I_0$ are initial intentions */
while true do
    get next percept $\rho$ via sensors;
    $B \leftarrow brf(B, \rho)$;
    $D \leftarrow options(B, I)$;
    $I \leftarrow filter(B, D, I)$;
    $\pi \leftarrow plan(B, I, Ac)$;  /* $Ac$ is the set of actions */
    while not ($empty(\pi)$ or $succeeded(I, B)$ or $impossible(I, B)$) do
        $\alpha \leftarrow$ first element of $\pi$;
        $execute(\alpha)$;
        $\pi \leftarrow$ tail of $\pi$;
        observe environment to get next percept $\rho$;
        $B \leftarrow brf(B, \rho)$;
        if $reconsider(I, B)$ then
            $D \leftarrow options(B, I)$;
            $I \leftarrow filter(B, D, I)$;
        end-if
        if not $sound(\pi, I, B)$ then
            $\pi \leftarrow plan(B, I, Ac)$
        end-if
    end-while
end-while
```

Fig. 2. Practical reasoning taken from [2] .

This scenario is the result of the implementation of self-ability and contributes improving the trustful interaction. In the sense that trust, and then the attitude to adopt or delegate, may change accordingly. For instance, let us suppose a person sitting on his desk in a room having the goal of going out the room; this aim may be pursued by performing some simple actions like for instance standing up, heading to the door, opening the door with the key, going out. For each action the performer uses the knowledge he owns about the external environment and himself, about his own capabilities: he has to be able to stand up, he has to know that a key is necessary for opening the door and he has to possess that key and so on. Before and during each action the person continuously and iteratively checks and monitors if he can perform the action. This can be translated in: having the knowledge on all the conditions allowing an action to be undertaken and finished.

In section II, in the trust function, the mental state of the trust is achieved through actions, agent beliefs are implicit and do not appear as direct variables in the trust function. For the purpose of this work, we made beliefs explicit so that each action of the model corresponds to one belief. This choice allowed us to map the theory of trust with the BDI cycle and to regularly report the new BDI cycle to the implementation part including Jason.

We needed to introduce a new representation in the model of τ from [6].

$$TRUST(X\ Y\ C\ \tau\ g_X) \tag{2}$$

$$\text{where} \quad \tau = (\alpha, p) \quad \text{and} \quad g_X \equiv p; \tag{3}$$

By combining the trust theory model and the self-modeling approach, τ is a couple of a set of plans π_i and the related results p_i. Indeed, now the trust model may implement the BDI paradigm breaking down actions and results into a combination of various arrangements of plans and sub-results.

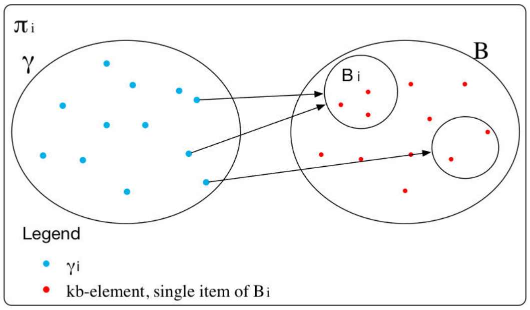


Fig. 3. Mapping actions onto beliefs (relation 4)

The model of τ is formalized as:

$$\tau = (\alpha, p) \quad \text{where} \quad \alpha = \bigcup_{i=1}^{n} \pi_i \quad \text{and} \quad p = \bigcup_{i=1}^{n} p_i \tag{4}$$

Moreover, each atomic plan π_i is the composition of action γ_i and the portion of belief base B_i for pursuing it; formalized as:

$$\pi_i = \gamma_i \circ B_i \Rightarrow \alpha = \bigcup_{i=1}^{n} (\gamma_i \circ B_i) \tag{5}$$

B_i is a portion of the initial belief base of the overall BDI system. The $\circ$ *operator* represents the composition between each action of a plan with a subset of the belief base (Fig. 3)

This theoretical framework has been implemented in a real robotic platform (the NAO-robot) exploiting Jason [2] and CArtAgO [16] for representing the BDI agents and the virtual environment. The environment model is created through the implementation of a perception module using NAO. Actions, into the real world, are performed using *CArtAgO Artifact* through *@Operation function*.

What happens while executing actions can be explained by referring to the BDI reasoning cycle. Once the robotic system has been, at a first stance, analyzed, designed and put in execution all the agents involved in the system acquire knowledge. They explore the belief base and all the initial goals they are responsible for (points 1. 2. 3. 4. - Fig. 2). Then, the module implementing the deliberation and means-and-reasoning (points 5. 6. 7. - Fig. 2) is enriched with a new function. Commonly at this point, while executing the BDI cycle, the tail of actions for each plan is elaborated to let the agent decide which action to perform. Since we are interested in the tail of actions and in all the knowledge useful for each action, we add a new function:

$$A_c \leftarrow \mathit{action}(B_{\alpha_i}, \mathit{Cap}) \tag{6}$$

where B_{α_i} and Cap are respectively portions of belief base related to the action α_i and the set of agent's capability for that action.

Agents execution and monitoring, implies the points 8. 9. 10. 11. 12 of the BDI cycle, that we enriched with a new portion of the algorithm able to identify the *impossible* (I,B) and ¬ *succeeded*(I,B) (ref. point 9.)

In this step the effective trust interaction takes place, here we may assume that the robot is endowed with the ability to re-planning, justifying and requesting supplementary information to the human being. Thus making the robot fully and trustfully autonomous and adaptive to each kind of situation it might face or learn depending on its capabilities and knowledge. The newly added functions, only for the case of *justification*, are shown in the following algorithm:

Algorithm 1:

```
foreach $\alpha_i$ do
    evaluate($\alpha_i$);
    $J \leftarrow$ justify($\alpha_i$,$B_{\alpha_i}$);
end
```

Fig. 4 details all the elements and the mapping process among beliefs, actions and plans.

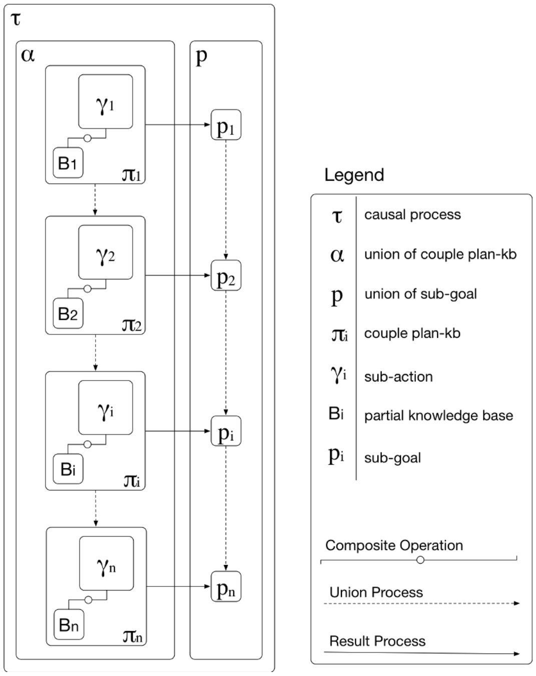


Fig. 4. A block-diagram representation of the causal process τ.

Summarizing, τ is the goal that a trustor decides to assign to a trustee; it means that a BDI agent is assigned the responsibility to perform all the actions γ_i included in τ. The BDI implementation using Jason and CArtAgO environments natively owns means for realizing the trust model, by implying:

- Jason Agent - is a BDI agent that allows managing the NAO robot through an AgentSpeak formalization and the related asl file [2] with the following:
  - ASL Beliefs - is the portion of asl file allowing to encode the agents' knowledge base through a set of beliefs. The set of beliefs includes all the knowledge about the external and the inner (the capabilities) environment of an agent;
  - ASL Rules - is a way that we use to represent beliefs that include norms, constraints and domain rules;
  - ASL Goals - is the asl file section devoted to encode the list of goals of the application domains (the list of desires in the BDI logic);
  - ASL Plans - is the section devoted to encoding the high logic inference to do actions ;
  - ASL Actions - is the actual part of the asl file that let agent commit actions hence a plan;
- CArtAgO Artifact - let the agent perform a set of actions into the environment. The environment is represented into CArtAgO virtual environment through all the beliefs acquired by NAO's perception module. Moreover, in *init function*, all the initial beliefs are imported from the jason agent file;
- CArtAgO @Operation - is used to implement the agent's actions in the environment.

Therefore, starting from:

- a reference model of environment and agents where the key point is to consider the agent (hence the robot);
- all the internal elements of agents as part of the environment;
- the BDI cycle;
- the theory of trust by Falcone and Castelfranchi; [6]

we implemented the trust model allowing to realizing self-modeling abilities in the agent.

In the following section, we validate this idea by developing a human-robot team employing the NAO robot and one human.

IV. Validation - the robot in action using JASON

The case study we show in this section focuses on a human-robot team whose goal is to carry a certain number of objects from a position to another in the room. The work to be done is intended to be collaborative and cooperative. Ideally, and this is part of the continuation of the present work, both the human and the robot know the overall goals of the system and communicate each other in order to commit or to delegate some goals. In this setup, we decided to simplify the example and considered only the case in which the robot is assigned (by code, thus simulating the command of the human) to pursue a specific goal, therefore the first type of delegation shown in section II.

The environment is composed of a set of objects marked with the landmarks useful for the NAO to work [1], the set of capabilities is made up basing on the NAO features, for instance, to be able to grasp a little box. The NAO is endowed with the capability of discriminating the dimensions of the box, and so on.

In this simplified case there is only one agent, the one managing the robot, which has the responsibility of carrying a specific object to a given position. The human, ideally the other agent of the system, indicates the object and its position.

Let us suppose to decompose the main goal (as shown in Fig. 5) *BoxInTheRigthPosition* in three sub-goals, namely *FoundBox*, *BoxGrasped ReachedPosition*. Let us consider the sub-goal *ReachedPosition*, two of the actions that allow pursuing this goal are: *goAhead* and *holdBox*[2].The NAO has to go ahead towards the objective and contemporarily hold the box. The beliefs associated with these actions refer to the concepts of the knowledge base these actions affect. In this case, one of the concept is the box, it has attributes like its dimension, its color, its weight, its initial position and so on. The approach we use for describing the environment results in a model containing all the actions that can be made on a box, for instance *holdBox*, and a set of predicates representing the beliefs for each object, for instance *hasVisionParameters* or *isDropped*. They lead to the beliefs *visionParameter* and *dropped* that are associated with the action *holdBox* through the relation number (5).

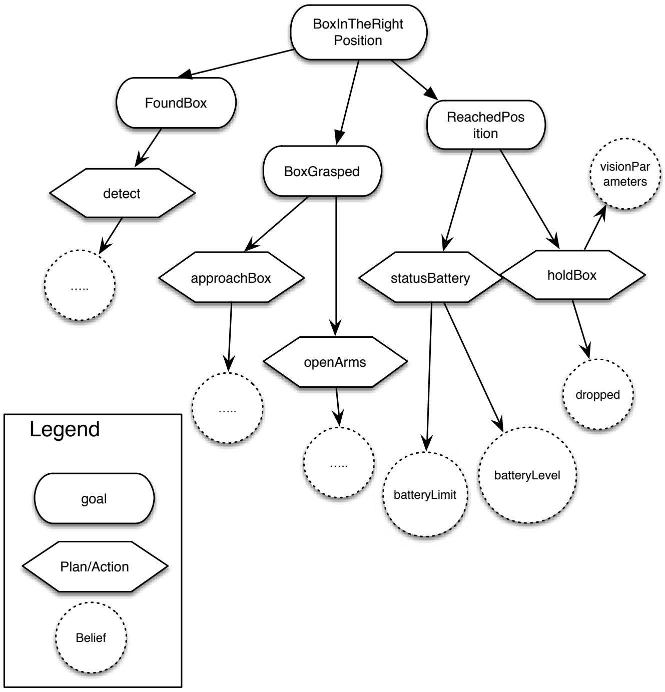


Fig. 5. A portion of the assignment tree for the case study

In the following a portion of code related to this part of the example.

Algorithm 2: Portion of code that implement the τ decomposition.

```
1 +!ReachedPosition: true ← goAhead; holdBox. [$\tau$];
2 +!goAhead: batteryLimit(X) & batteryLevel(Y) &
    $Y < X$ ← say(“My battery is exhaust. Please let me
    charge.”). [$\gamma_1^+$];
3 +!goAhead: batteryLimit(X) & batteryLevel(Y) &
    $Y \geq X$ ← execActions. [$\gamma_1^-$];
4 $B_1$: batteryLimit, batteryLevel ;
5 +!holdBox: dropped(X) & visionParameters(Y) &
    $X == false$ ← execAct(Y). [$\gamma_2^+$];
6 +!holdBox: dropped(X) & visionParameters(Y) &
    $X == true$ ← say(“The box is dropped.”). [$\gamma_2^-$];
7 $B_2$: dropped, visionParameters ;
```

It is worth to note that the model we developed does not change the way we implement the agent, but only adds a way to match knowledge to actions.

In Fig. 6 some pictures showing the execution of the case study with the NAO robot.

V. Related Work

Most of the work in the literature explores the concept of trust, how to implement it and how to use it, from an agent society, working in an open and dynamic environment, viewpoint. So literature mostly focuses on organizations in

[1] All the technological implications of using the NAO robot are out of the scope of this section.

[2] For space concerns in this paper we show only an excerpt of the whole AssignmentTree diagram, so only few explanatory belies for each action are reported.

which multiple agents must interact with each other and decide which action to take based on a certain level of trust in each other. In our case, while sharing the concept of an open and dynamic environment, we focus on the theme of man and robot and explore the two-way role of man-robot and robot-man, of trust in the interactions between them.

Among the approaches in the literature that focus on trust-based interactions in open and dynamic environments, here we briefly present and compare our approach with some existing ones that, in our opinion, embody the basic features of most trust approaches.

In [14] decision making is based on trust evaluation through a decision-theoretic model that allows controlling trust decision-making activities. The leading point of this works is to make agents able to evaluate trust. Some reputation mechanism enables trustor to make a better evaluation. Our work shares the same objectives but it focuses, we may say, at a different level of abstraction, we endow the agent with self-modeling abilities to give the trustor a means for delegating or making the action by himself. We propose this way as a higher autonomous form of interaction and cooperation.

In [18] the trust model is applied to the virtual organization and uses a probabilistic theory that considers parameter calculated from past interactions, if some information lacks or is inaccurate then the model relies on third parties. In our case instead, we pose the basis for giving the trustee the ability to ask for help when it does not possess the knowledge to perform the delegated action thus always letting the possibility to the trustor to evaluate. It is some kind of reverse logic, it is no longer the trustor who is concerned about assessing his trust in the trustee but it is the trustee who provides the means to do so.

In [13] is presented a trust model based on reputation, here FIRE allows creating a measure for the trust that can be used in different circumstances. This model overcomes the problem of evaluating trust in a dynamic environment where it is difficult to consolidate the knowledge the agent has on the environment. The model we propose, instead, is at this moment constrained by the fact that the trustor establishes a level of trust by observing the other agent. However, endowing the trustee with self-modeling abilities gives the trustor the possibility to evaluate the work of the other better. In the sense that the trustor must not only imagine and then evaluate what the trustee is doing, just by his beliefs but can be enriched by the explanations that are given by the trustee.

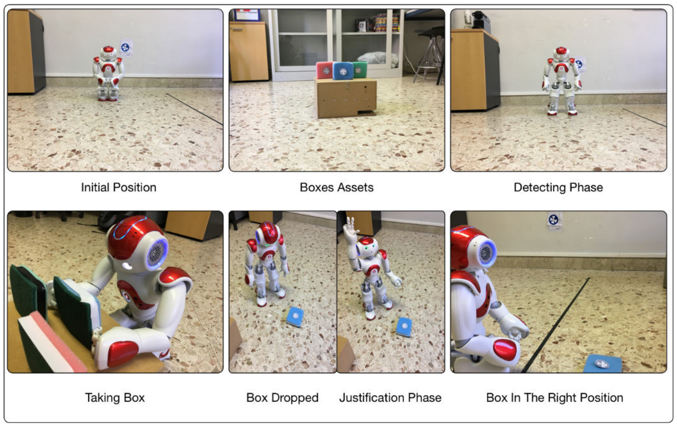


Fig. 6. The NAO working on the *BoxInTheRightPosition* goal and the justification

A different approach is proposed in [17], here the authors use a meta-analysis for establishing which features of the robot may affect trustworthy relationship form the point of view of the human. The robot is considered a participant to the team but not an active part of it, some kind of resource. From this work we may outline the main difference of our approach against all the others, we consider the trustee (agent, robot or whatever else) an active autonomous entity in the interaction.

VI. Discussions and Conclusions

In this work, we employed the trust model by Falcone and Castelfranchi for human-robot interaction in unknown and highly dynamic environments.

The primary goal of our work is to equip the robot with the self-modeling ability that allows it to be aware of its skills and failures. In this work, we made self-modeling explicit as the ability to justify oneself in the case of failure. In the future, we will extend the model with the ability to ask for help when the trustor's requests do not fall within the trustee's knowledge and the ability to autonomously re-planning.

The trust model has been integrated with a BDI-based part of the deliberation process to include self-modeling. The self-modeling ability is obtained by joining the plan a BDI agent commit to activating with the portion of knowledge base useful for it.

We chose and used JASON and CArtAgO because they natively support everything that is part of the BDI theory and besides allow us to implement, without significant changes to the agent language paradigm, all the elements of the new reference model for the environment we use.

The outcomes we use in the various phases are not binding; we are inspired by Tropos [5] for modeling goals, actions and capabilities. However, we might use whatever methodological approach giving a view of goals and their decomposition, and decomposition into plans in a way useful to match with the related knowledge base.

In the future, we are going to develop and implement the complete trust model that also implies the ability of an entity to understand what the other one is going to do. In this way, we aim at implementing human-robot interaction where each involved entity delegates or commits an action on the base of a kind of theory of mind of the other.

References

[1] Rafael H Bordini and Jomi F Hübner. BDI agent programming in agentspeak using jason. In *Proceedings of the 6th international conference on Computational Logic in Multi-Agent Systems*, pages 143–164. Springer-Verlag, 2005.

[2] Rafael H Bordini, Jomi Fred Hübner, and Michael Wooldridge. *Programming multi-agent systems in AgentSpeak using Jason*, volume 8. John Wiley & Sons, 2007.

[3] Michael Bratman. *Intention, plans, and practical reason*, volume 10. Harvard University Press Cambridge, MA, 1987.

[4] Michael E Bratman. What is intention. *Intentions in communication*, pages 15–32, 1990.

[5] Paolo Bresciani, Anna Perini, Paolo Giorgini, Fausto Giunchiglia, and John Mylopoulos. Tropos: An agent-oriented software development methodology. *Autonomous Agents and Multi-Agent Systems*, 8(3):203–236, 2004.

[6] Christiano Castelfranchi and Rino Falcone. *Trust theory: A socio-cognitive and computational model*, volume 18. John Wiley & Sons, 2010.

[7] Cristiano Castelfranchi and Rino Falcone. Delegation conflicts. *Multi-agent rationality*, pages 234–254, 1997.

[8] Cristiano Castelfranchi and Rino Falcone. Towards a theory of delegation for agent-based systems. *Robotics and Autonomous Systems*, 24(3-4):141–157, 1998.

[9] Rino Falcone and Cristiano Castelfranchi. The human in the loop of a delegated agent: The theory of adjustable social autonomy. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 31(5):406–418, 2001.

[10] Rino Falcone and Cristiano Castelfranchi. Social trust: A cognitive approach. In *Trust and deception in virtual societies*, pages 55–90. Springer, 2001.

[11] Rino Falcone and Cristiano Castelfranchi. Trust dynamics: How trust is influenced by direct experiences and by trust itself. In *Autonomous Agents and Multiagent Systems, 2004. AAMAS 2004. Proceedings of the Third International Joint Conference on*, pages 740–747. IEEE, 2004.

[12] Diego Gambetta. Can we trust trust? trust: Making and breaking cooperative relations, department of sociology, university of oxford, 2000.

[13] Trung Dong Huynh, Nicholas R Jennings, and Nigel R Shadbolt. An integrated trust and reputation model for open multi-agent systems. *Autonomous Agents and Multi-Agent Systems*, 13(2):119–154, 2006.

[14] Chris Burnett Timothy J Norman and Katia Sycara. Trust decision-making in multi-agent systems. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, 2011.

[15] Anand S Rao, Michael P Georgeff, et al. Bdi agents: from theory to practice. In *ICMAS*, volume 95, pages 312–319, 1995.

[16] Alessandro Ricci, Mirko Viroli, and Andrea Omicini. Cartago: A framework for prototyping artifact-based environments in mas. *E4MAS*, 6:67–86, 2006.

[17] Tracy Sanders, Kristin E Oleson, Deborah R Billings, Jessie YC Chen, and Peter A Hancock. A model of human-robot trust: Theoretical model development. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 55, pages 1432–1436. SAGE Publications Sage CA: Los Angeles, CA, 2011.

[18] WT Luke Teacy, Jigar Patel, Nicholas R Jennings, and Michael Luck. Travos: Trust and reputation in the context of inaccurate information sources. *Autonomous Agents and Multi-Agent Systems*, 12(2):183–198, 2006.

[19] Michael Wooldridge. *An introduction to multiagent systems*. John Wiley & Sons, 2009.

[20] Michael Wooldridge and Nicholas R Jennings. Intelligent agents: Theory and practice. *The knowledge engineering review*, 10(2):115–152, 1995.