KEY ELEMENTS FOR JOINT HUMAN-ROBOT ACTION

Aurélie CLODIC (*), Rachid ALAMI (*), Raja CHATILA (**)

* CNRS, LAAS, 7 avenue du colonel Roche, F 31400 Toulouse, France and Univ de Toulouse, LAAS, F 31400 Toulouse, France
** Sorbonne Universités, UPMC, Univ Paris 06, UMR 7222, Institut des Systèmes Intelligents et de Robotique, F-75005, Paris, France and CNRS, UMR 7222, Institut des Systèmes Intelligents et de Robotique, F-75005, Paris, France.

Abstract For more than a decade, the field of human-robot interaction has generated many valuable contributions of interest to the robotics community at large. The field is vast, going all the way from perception to action and decision. In the same time, research on human-human joint action has become a topic of intense research in cognitive psychology and philosophy, bringing elements and even architecture hints to help our understanding of human-human joint action. In this paper, we would like to analyse some findings from these disciplines and connect them to the human-robot joint action case. This work is for us a first step toward the definition of a framework dedicated to human-robot interaction.

Keywords Action, Joint action, Architecture for Social Robotics, Human Robot Interaction

INTRODUCTION

For more than a decade, the field of human-robot interaction has generated many valuable contributions of interest to the robotics community at large. The field is vast, going all the way from perception (e.g., tactile or visual) to action (e.g., manipulation, navigation) and decision (e.g., interaction, human-aware planning). In the same time, research on human-human joint action has become a topic of intense research in cognitive psychology and philosophy, bringing elements and even architecture hints to help our understanding of human-human joint action. In this paper, we would like to analyse some findings from these disciplines and connect them to the human-robot joint action case. More precisely, we are trying to address in this paper the following questions:

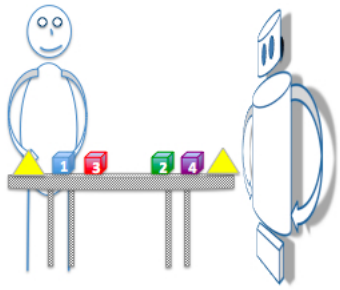
- What a robot needs to understand about the human it interacts with for the interaction to be successful and thus what capacities the robot should be equipped with to ensure it can build this understanding?
- On the other hand, the robot also needs to be understood by its human partner. How this understanding operates and what is needed to enable the robot to behave appropriately and in a way that manifests what it is doing to the human partner

This work is for us a first step toward the definition of an integrative framework needed for the design of an autonomous robot that can engage in interaction with a human partner.[1]

[1] ***This work was conducted within ANR-CONTINT ROBOERGOSUM project (DECISION ANR-***

RELATED WORK AND VISION

Let's illustrate by a simple example, the kind of interaction we envision. A human and a robot have the goal to build a pile with 4 cubes and put a triangle at the top. There are face to face. One after the other, they should stack bricks in the expected order. Each agent has a number of cubes accessible in front of him and would participate to the task by placing its cubes on the pile. At the end, one of the agents should place a triangle at the top of the pile.



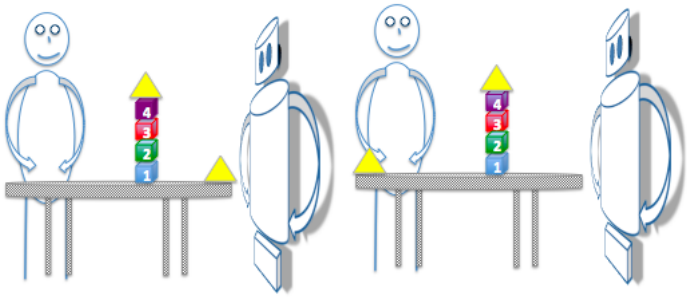
The picture illustrates the initial state. Actions available for each agent are the following (with object = cube or triangle):

- take an object on the table
- take an object from the pile
- put an object on the pile
- give an object to the other agent
- support the pile

Each agent is able to infer the state of the world so it knows:

- where each object is
- if an object is reachable for itself
- if an object is reachable for the other one

Moreover, we assume each agent is able to observe the activity of the other. The expected final state could be one of the following:



Possible deviations could be for example that an agent drops a brick on its side / in the opposite side (e.g. if the brick falls down on the opposite side so that it becomes unreachable for the intended agent to put it on the pile, consider whether the other agent should put the brick directly on the pile or give it to the intended agent) or that the pile collapse. Moreover, during the execution of the task, a number of behaviours can arise, among all: Proactive behaviour (one agent could [be lead to] help the other one by supporting the pile while the other places a brick on it), "Inactive" behaviour (one agent does not act at all) or "Incorrect" behaviour (one agent does not pile bricks in the correct order or one agent removes a correctly placed brick from the pile).

A number of robotics systems, that could be more or less considered as frameworks or architectures dedicated to human-robot interaction have been built, among all [2, 3, 5, 7, 9,

10, 11, 13, 14, 22, 23, 25, 26, 28, 29]. These works have made the robotics community move a step toward understanding human-robot interaction specificities. We want now to have a look on the needed elements to build a frame around all that contributions and this paper is a first step toward this search.

Our aim here is to link human-robot interaction needs to human-human joint action research and to see how it can help to frame such architecture effort.

ACTING AUTONOMOUSLY

Before entering the joint action domain, we feel necessary to situate the context of autonomous (or individual) action. According to Pacherie [18], today dominant position in philosophical action theory is that "behaviour qualifies as action just in case it has a certain cause or involves a certain sort of psychological process". In the same stream, Tomasello [30] proposes that an "intention is a plan of action the organism chooses and commits itself to in pursuit of a goal. An intention thus includes both a means (action plan) as well as a goal" and that "choosing an intended course of action (decision making), the organism consults both its stored knowledge/skills and its mental model of current reality". We are now equipped, with a definition of an action, an intention, a goal and a plan, elements that should be handled to enable acting.

In the 90's the robotics community tackled the problem of robot control architecture and gave to it several solutions. One of them was the three-layered architecture [12, 1, 16, 17, 25, 29], which defines:

A functional level which includes all the basic built-in robot action and perception capacities. These processing functions and control loops (image processing, obstacle avoidance, motion control, etc.) are encapsulated into controllable communicating modules. In order to make this level as hardware independent as possible, and hence portable from a robot to another, it is interfaced with the sensors and effectors through a logical robot level. In order to accomplish a task, the next level activates the modules.

An execution control level, or executive, which controls and coordinates the execution of the functions distributed in the modules according to the task requirements. It is at this level that context-based action refinement is performed.

A decision level which includes the capacities of producing the task plan and supervising its execution, while being at the same time reactive to events from the previous level. This level may be decomposed into two or more layers, based on the same conceptual design, but using different representation abstractions or different algorithmic tools, and having different temporal properties.

This architecture relies on representation of action, goal, plan as well as robot's knowledge and skills. However, robot knowledge representation and management is still an open problem.

Interestingly, Pacherie [18][19] proposes an action theory that also distinguishes three main stages in the process of action specification:

- distal intentions level (D-intentions) in charge of the dynamics of decision making, temporal flexibility and high level rational guidance and monitoring of action;
- proximal intentions level (P-intentions) which inherits a plan from the previous level and which role is to anchor this plan in the situation of action, this anchoring has to be performed at two levels: temporal anchoring and situational anchoring;
- motor intentions level (M-intentions), which encodes what neuroscientists call motor representations; with two levels of dynamics: local (specific to each level of intention), global (transition from one level of intention to the next).

This nicely shows a convergence between a philosophical theory of action and a robot control architecture dedicated to action. It seems relevant to have a look if we can build a similar convergence with joint action theory.

ACTING JOINTLY

As stated by Knoblich [15] "What distinguishes joint actions from individual actions is that the joint ones involve a shared intention and shared intentions are essential for understanding coordinate joint action". Tomasello [30] says nothing else when he assumes that "Understanding the intentional actions and perception of others is not by itself sufficient to produce humanlike social or cultural activities. Something additional is required. Our hypothesis for this "something additional" is shared intentionality" and more precisely [30] "shared intentionality refers to collaborative interactions in which participants have a shared goal (shared commitment) and coordinated action roles for pursuing that shared goal".

Pacherie proposes a theory of joint action, which also considers three levels of action [20, 21, 22]. If we try to map this theory to robot architecture, we can describe these three levels as the following:

SHARED DISTAL/DECISIONAL LEVEL

At this level, acting lonely, the robot handles its goal, plan and decision-making; all elements that it represents would be realized by itself. Acting jointly, the robot must be able to handle joint goal, plan and action representation and possibly cooperative decision-making (including e.g. joint planning abilities). It will represent not only what would be achieved by itself but also by the other (with potentially different levels of granularity and completeness). Moreover, high level monitoring would include not only its monitoring but also more generally monitoring of the joint goal and consequently monitoring of the other actions too.

Then, forming a plan, even a collaborative one, questions hold for the robot: how does it share the plan with the human? Could the robot assume that the plan is shared? How can we handle "jointly" the plan negotiation? The human-robot joint goal can be the result of an explicit interaction (a request from the human for instance) or implicitly if the robot pro-actively decides to engage into a joint action.

SHARED PROXIMAL/EXECUTION LEVEL

It is at that level that will arise situational and temporal anchoring of the action, which means parameterization of functional level and functions launching and monitoring. At that level, the robot and the human need to be able to share representations (in the best case jointly) and to coordinate their perceptions (to achieve joint attention) in order to coordinate their

actions and possibly realize adjustment (dyadic, triadic and collaborative) in the current context.

Coupled Motor/Functional level

This level will correspond to robot sensory-motor behaviour that would allow to achieve high-bandwidth interaction with its human partner. An example could be exchanging an object with a human and the associated force-feedback processes. In such tight situation, involving precise coordination between the actors, the parameterization of the functional level needs to be coupled with the one of the other actor. That means, e.g. that the robot control loop would be directly parameterized by the other actor move or action.

We see that this three layers division seems meaningful not only for the human-human case but also for the human-robot case. Having that in mind, we will now explain which elements are needed to setup a framework based on this.

What is needed for joint action?

We want now to identify and localize the main ingredients and process involved in joint action and how they can make sense in a robotics context. To do so, we will inspire from joint action theory [20], shared intentions theory [30] and other works in psychology of joint action [15] or language [6]; some of those works derive from joint intention theory [8] and shared cooperative activity [4].

According to Knoblich [15]: "a joint action is a social interaction whereby two or more individuals coordinate their actions in space and time to bring about a change in the environment". In [20], Pacherie proposes several dimensions of joint action, in our case, we will consider what she calls: small scale, egalitarian, involving face-to-face interaction. We will first study intentional action understanding as a first step to joint action, then analyse joint attention and elements that need to be shared to end up with a proposal of joint representation definition.

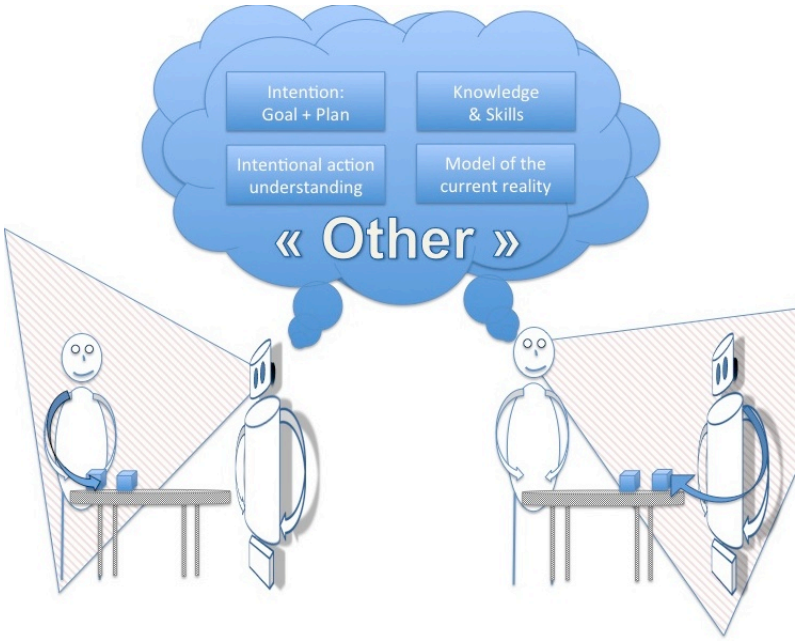
How to SHARE?

It is obvious that dialog and negotiation is a way to share [8], but there we will focus on lower level means to achieve joint action.

Intentional action understanding

A very interesting prerequisite to joint action established by Tomasello in [30] is understanding of intentional action. We mean that each actor should be able to read its interactor actions. For an observer to understand an intentional action, he must, viewing an actor's action and more precisely actor's course of actions, be able to: represent the actor's intention (i.e. its goal and plan, and possibly to understand that a choice has been made between several plans) and to understand what the actor is attending to in its perceptive field. This kind of "reverse engineering" process is possible under the assumption that the viewer owns

representations about/of the other: its knowledge and skills (and possibly its lack of knowledge) and its model of the current reality as illustrated in the following figure



that represents the intentional action understanding. At left, the robot represents the other (in this case the human) and infers that what it is doing. At right, the human represents the other (in this case the robot) and infer what it is doing.

What does this say for a robot to understand a human intentional action? That means, we must equip the robot abilities to represent "the other". To this end, the question has to be answered if it could use its proper representations adapted to the "other" or if any other form of representations is needed. This capability should be, of course, limited to the context and the tasks the robot will be involved into: for instance navigation and associated activities, or simple object manipulation in domestic environment.

On the other side, for a human to understand a robot intentional action, he must have access to robot knowledge and skills, this means the robot should be (and behave so that to be) understandable to the human. That means too, that the human must be able to infer the robot model of the current reality and it is not so simple since the robot sensing abilities are not fully readable by the human.

JOINT ATTENTION

One key means to share perceptual representation in face-to-face interaction is joint attention. Attention is the cognitive process of selectively concentrating on one aspect of the environment while ignoring other things (from http://en.wikipedia.org/wiki/Attention). Pacherie defines joint attention as "two people attending to the same object or event + actual attention sharing (there must be some causal connection between the two subjects' acts of attending) + mutual manifestness (the fact that both are attending to the same object or event should be open or mutually manifest)"([20] page 355). This concept, that we could find too in [30] or [15] for example, is key because it states that if joint attention is established, whatever information I can get, I can consider my interactor would have it too if it occurs in the joint attention space. It includes what both interactors perceive, but also what

only one interactor perceives (e.g. if one part of the table is hidden to the robot, the robot can establish that it cannot see a part of the environment, whereas the human is able to see this part - and vice-versa, the robot can assume the human knows that a part of the table is hidden to the robot and that the human can see this part.). This raises a number of questions: How can a robot know that the human it interacts with joint attended with him to the joint task? What are the cues that should be collect to infer joint attention? Symmetrically, how can a robot exhibit joint attention? What cues the robot should exhibit to let the human infer that joint attention is met? Moreover, once joint attention is achieved (or at least a given level of joint attention if we consider it is not a 0/1 option), how should it be managed during the overall course of joint actions? Tomasello [30] explains that actors need to handle cooperative perception while joint goal unfolds. How can we handle cooperative perception? Does this need to be taken into account at planning level or at anchoring level?

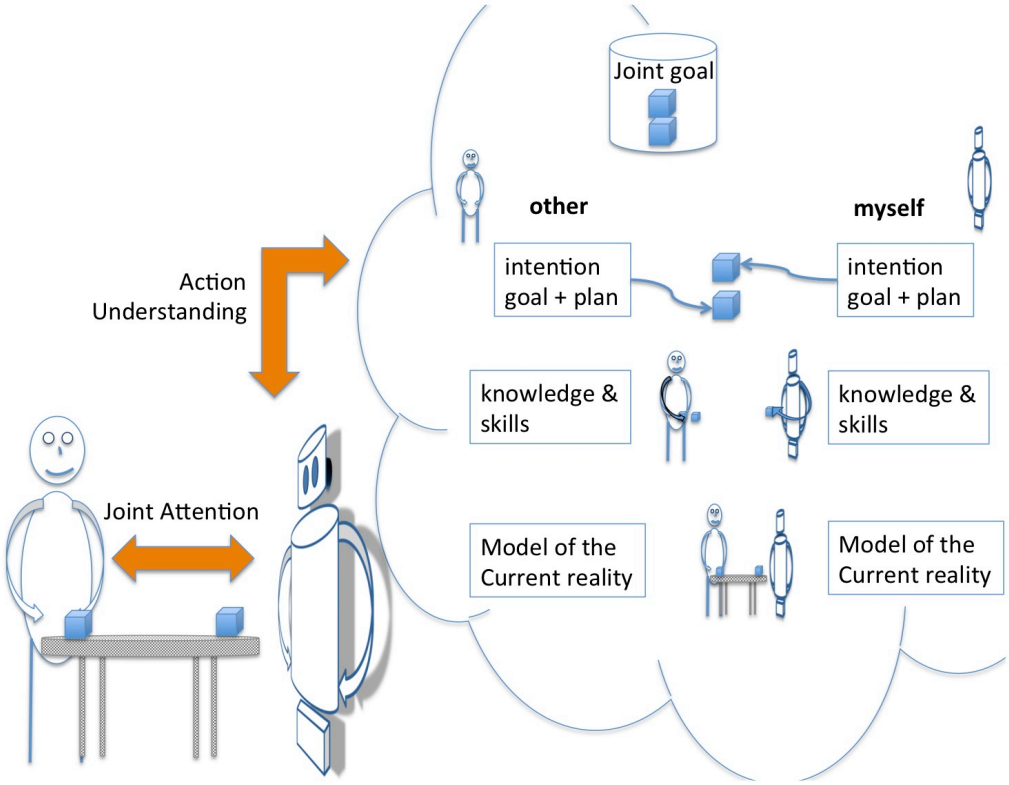
Under joint attention assumption, in the joint attention space, all events that happen are supposed to be shared between the interactors. It has to be noticed that this information needs to be filtered by perspective taking abilities ([15]): if the robot is in face of the human and perceives that brick 1 is at its left side, it should infer that brick 1 is at the right side of human. Moreover, that does not say anything about what both the interactors perceive means for both of them. It is there shared action/task/goal representation is helpful.

WHAT TO SHARE?

In [20], Pacherie establishes that a number of elements must be handled by each agent to drive a joint action:

- self-predictions: agents each represent their own actions and their predicted consequences in the situation at hand.
- other-predictions: agents each represent the actions, goal, motor and proximal intentions of their coagents and their consequences.
- dyadic adjustment: agents each represent how what they are doing affects what others are doing and vice-versa and adjust their actions accordingly.
- joint action plan: agents each have a representation (which may be only partial) of the hierarchy of situated goals and desired states culminating in the overall joint goal
- joint predictions: agents each predict the joint effects of their own and other's actions
- triadic adjustment: agents each use joint predictions to monitor progress toward the joint goal and decide on their next moves, including moves that may involve helping others achieve their contributions to the joint goal

In our context that could be illustrated by the following figure



That means the robot needs to be able to handle: its world representation, a world representation of the human it interacts with (again potentially limited to the task to perform), the possible effect of its actions on the human actions (and vice-versa), their joint goal and action plan representation, a prediction of their actions, a mean to monitor progress toward the joint goal (and possibly mean to revise the on-going joint plan). A triadic adjustment means that the robot and the human can adapt their behaviour toward the joint goal. That means for example, that if the human brings down its brick in the robot space, the robot will place the brick on the stack. If it had done a dyadic adjustment it had make accessible the brick to the human to let him finish its action, a dyadic adjustment means that the robot and the human can adapt their behaviour to the other actions (not toward the joint goal). It has to be noticed that Tomasello [30] does not use exactly the same nomenclature and adds another adjustment (engagement in its vocabulary) level: the collaborative level where he considers the two must plan together toward the joint goal (he does not consider that it is done at previous levels of dyadic and triadic) that could handle behaviour where the human can hold the stack while the robot places its last brick (from [30] page 682). To be able to deal with such elements, the robot must share representations with the human it interacts with: perceptual representation, (joint) action/task/goal representation. This idea of sharing representation drifts from shared intentionality [30], shared intention [15] or interdependence of the individual intentions [21]. However, it has to be noticed that shared representation does not mean common representation. Representations could differ, the important thing is that we are aware of.

WHERE TO SHARE?

Representation sharing could be helped by mechanism such as affordance. From [15], object affordances are the action opportunities that an object provides for an agent with a particular action repertoire whereas common affordance states that when two agents have similar action repertoires and perceive the same object, they are likely to engage in similar actions because the object affords the same action for both of them. That means we must give to the robot access to object and common affordances model to help its human understanding.

Representation sharing would help the robot to achieve perception-action matching [20] in 2 directions: action-to-goal prediction (goal attribution to observed action execution) and goal-to-action prediction (anticipate the observed actor's next actions). It goes in the same direction as [15] common predictive models: "action simulation can lead to emergent coordination because it induces the same expectations about the unfolding of actions in different actors and thus induces similar action tendencies for future actions". Consequently that would help actions prediction and monitoring and also enable dyadic, triadic and collaborative adjustment.

That means that:

- perceiving an object, the robot and the human it interacts with must share information such as:
    - I perceive the object, you perceive the object, I know you know what I perceive;
    - I know what it is (or not), I know you know what it is (or not), I know you know what I know;
    - I know what is its purpose (or not), I know you know its purpose (or not), I know you know what I know;
    - I know how to handle the object (or not), I know you know how to handle the object (or not), I know you know what I know;

  e.g. for a robot seeing a telephone on the table next to the human, what would be inferred/shared?

- perceiving an action, the robot and the human it interacts with must share information such as:
    - I perceive the action, you perceive the action, I know you know what I perceive;
    - I know what it means (or not), I know you know what it means (or not), I know you know what I know;

  e.g. for a robot, viewing the human scratching its head, what would be inferred/shared?

The needed information, its various levels and the way to represent it, even it has been studied in different ways remains an open question that need to be tackle. Moreover, in case of loss or lack of information sharing, the robot needs to be able to inform or facilitate the information acquisition of the human. This has to do with expressive multi-modal behaviour and what is called mutual manifestness (see below).

In addition, we have the intuition that shared representation seems not enough in the context

of human-robot interaction. We need what we can call joint representation, i.e. if we refer to Pacherie joint attention definition, we could define joint representation as:"two people sharing a representation + actual representation sharing (there must be some causal connection between the two subjects' acts of sharing a representation) + mutual manifestness (the fact that both are sharing the representation should be open or mutually manifest)". This means that we know (or not) a representation should be shared mutually. In human-human interaction, assumption can be easily made from both sides on what the other knows or not, this is far more difficult in human-robot interaction. On the robot side, this indicates that we need to integrate into the robot means to share representations explicitly with the human but also means to recognize and understand them (and to learn them if needed). On the other side, a human interacting with a robot is often disconcerted because it is difficult for him to have intuitions about robot capabilities and inabilities. What is missing here is what Tomasello [30] named cultural creation/learning, or Clark [6] common ground, and that is something we need to come up with.

CONCLUSION

In this paper we have proposed an analysis of some findings in Psychology and Philosophy in the domain of human-human joint action in order to come up with needs in terms of knowledge and abilities that a robot, interacting with a human, need to handle.

We propose that intentional action understanding, joint attention and joint representation management are key elements to better human-robot interaction unfolding. Then, we've seen that framework proposed by Philosophy such as [22] could be inspiring in the search to frame an architecture dedicated to human-robot interaction.

This work is a first step toward the objective to identify and incrementally give an accurate description of the different needed abilities and how they are involved in the overall process of collaborative human-robot task achievement.

Future steps would be to continue to analyse inputs from Philosophy and Psychology and to analyse if our requirements have been already implemented in a robotics architecture (even part of) and how. From this basis, we will continue to try to formalize (when possible) and to devise the pertinent human and task related models and the associated decision-making, planning and situation assessment processes.

References

[1] Alami, R., Chatila, R., Fleury, S., Ghallab, M., Ingrand, F.: An architecture for autonomy. International Journal of Robotic Research (1998)

[2] Breazeal, C.: Towards sociable robots. Robotics and Autonomous Syst. (2003)

[3] Breazeal, C., Berlin, M., Brooks, A., Gray, J., Thomaz, A.: Using perspective tak- ing to learn from ambiguous demonstrations. Robotics and Autonomous Systems (2006)

[4] Bratman, M.E.: Shared cooperative activity. The Philosophical Review (1992)

[5] Cangelosi A. (2010). Grounding language in action and perception: From cognitive agents to humanoid robots. *Physics of Life Reviews*, *7*(2), 139-151

[6] Clark, H.H.: Using Language. Cambridge University Press (1996)

[7] Clodic, A., Cao, H., S, A., Montreuil, V., Alami, R., Chatila, R.: Shary: A supervision system adapted to human-robot interaction. Springer Tracts in Advanced Robotics (2009)

[8] Cohen, P.R., Levesque, H.J.: Teamwork. Nous (1991)

[9] Demiris Y., "Prediction of intent in robotics and multi-agent systems", Cognitive Processing, 8: 151-158, 2007.

[10] Ferland F., Létourneau D., Aumon A., Frémy J., Legault M.A., Lauria M., Michaud F. : Natural interaction design of a humanoid robot, Journal of Human-Robot Interaction, 2012

[11] Fong, T.W., Kunz, C., Hiatt, L., Bugajska, M.: The human-robot interaction operating system, Proc. Conference on Human-Robot Interaction (HRI), (2006)

[12] E. Gat, Integrating planning and reacting in a heterogeneous asynchronous architecture for controlling real-world mobile robots, in Proceedings of the tenth national conference on Artificial intelligence, AAAI'92. AAAI Press, 1992, pp. 809–815.

[13] Grosz, B.J., Kraus, S.: Collaborative plans for complex group action. Artificial Intelligence (1996)

[14] Johnson, M., Demiris, Y.: Perceptual perspective taking and action recognition. Advanced Robotic Systems (2005)

[15] Knoblich G., Butterfill S., .S.N.: Psychological research on joint action: theory and data. In B. Ross (Ed.), The Psychology of Learning and Motivation (2011)

[16] N. Muscettola, P. P. Nayak, B. Pell, and B. C. Williams, "Remote agent: To boldly go where no ai system has gone before," 1998.

[17] I. Nesnas, A. Wright, M. Bajracharya, R. Simmons, T. Estlin: CLARAty and challenges of developing interoperable robotic software IEEE/RSJ Internat. Conf. on Intell. Robots and Systems (IROS) (2003)

[18] Pacherie, E.: The phenomenology of action: A conceptual framework. Cognition (2008)

[19] Pacherie,E.:Action.InK.Frankish&w.Ramsey(eds.),TheCambridgehandbook of cognitive science. Cambridge University Press (2012)

[20] Pacherie, E.: The phenomenology of joint action: Self-agency vs joint-agency. In Axel Seemann (ed.), Joint Attention: New Developments, Cambridge MA: MIT Press (2012)

[21] Pacherie, E.: Is collective intentionality really primitive? In M. Beaney, C. Penco & M. Vignolo (Eds.), Mental processes: representing and infering, Cambridge, Cambridge Scholars press (2007)

[22] Pacherie, E.: Framing joint action. Review of Philosophy and Psychology (2011)

[23] Pandey, A.K., Alami, R.: Towards effect-based autonomous understanding of task semantics for human-robot interaction. International Journal of Social Robotics (IJSR) (2013)

[24] M. Petit, S. Lallée, J. Boucher, G. Pointeau, P. Cheminade, D. Ognibene, E. Chinellato, U. Pattacini, I. Gori, U. Martinez-Hernandez, H. Barron-Gonzalez, M. Inderbitzin, A. Luvizotto, V. Vouloutsi, Y. Demiris, G. Metta and P. Dominey, "The Coordinating Role of Language in Real-Time Multi-Modal Learning of Cooperative Tasks", IEEE Transactions on Autonomous Mental Development, 5:1, pp 3-17, 2013

[25] G. N. Saridis, "Architectures for intelligent controls. in: Intelligent control systems: Theory and applications." IEEE Press, 1995.

[26] Scheutz M.: Computational Mechanisms for Mental Models in Human-Robot Interaction, HCI International, 2013, 304-312

[27] Sidner, C.L., Lee, C., Kidd, C., Lesh, N., Rich, C.: Explorations in engagement for humans and robots. Artificial Intelligence (2005)

[28] Sisbot, E.A., Clodic, A., Alami, R., Ransan, M.: Supervision and motion planning for a mobile manipulator interacting with humans. (2008)

[29] Tambe, M.: Towards flexible teamwork. JAIR (1997)

[30] Tomasello, M., Carpenter, M., Call, J., Behne, T., H., M.: Understanding and sharing intentions: The origins of cultural cognition. Behavioral and Brain Sciences (2005)

[31] Trafton, J., Cassimatis, N., Bugajska, M., Brock, D., Mintz, F., Schultz, A.: En- abling effective human-robot interaction using perspective-taking in robots. IEEE Transactions on Systems, Man, and Cybernetics (2005)