Marie Curie Individual Fellowship
FP7 Program

Report 1:

Concepts and Research Methodology

Produced by:	Raquel Ros Espinoza

Grant agreement Number:	220368
Project title:	Adapting Robot Behavior based on Interaction
Project acronym:	ARBI
Project starting date:	01 August 2008
Project ending date:	31 July 2008
Project coordinator:	Dr. Rachid Alami

Toulouse, August 2010.

1 Introduction

My work during the first year has been basically devoted to learn about the Human-Robot Interaction (HRI) field –which was new for me– trying to identify those areas that were more relevant to my project (ARBI), to the research lines of the host institution, and of course, those which draw my special attention in some way. One of this areas is the application of cognitive sciences to robotics. I will next try to explain the reasons that led me to study this area, and the therefore, to reconsider a different approach to achieve the goals of my research during this period of my career.

ARBI project

From the ARBI project, I remark the following text which refers to the global goal of the project: "We also need to provide the robots with other abilities such as (*i*) being proactive, (*ii*) perform understandable actions, (*iii*) achieve social acceptance and (*iv*) learn." I refer to the first three abilites as follows:

> *Being proactive*, ability to propose tasks (actions) to human partner based on the understanding of the other's intentions.
>
> *Perform understandable actions*, the actions performed by the robot should be legible for the human partner. In other words, the intention of the robot while executing a task should be clear.
>
> *Achieve social acceptance*, the human should feel comfortable with respect to her robot partner.

In all three cases the common idea behind them is the one of trying to understand the world from the other's point of view in order to improve my own behavior (let's assume that I am the robot). In the first case, placing myself in my partner's position will help me understand what she is doing and why she is doing so, i.e. it will allow me to understand her goals and intentions. In the next two cases, observing my own behavior from her point of view will allow me to evaluate my own performance to determine whether it is understandable and socially acceptable or not.

Understanding the other's point of view corresponds to the concept of *Perspective Taking*. Perspective taking (PT) has been widely studied in psychology and has many senses. I summarize the main concepts and types in next section. Researchers agree that it is fundamental for enhancing social interaction:

> "Effective social interaction depends on perspective-taking. Social interactions entail responding to the actions of others, whether those actions are verbal or physical. Anticipating how to behave in social situations may be promoted by perspective-taking, by considering the actions of others from their point of view. Speaking to others, understanding others, and reacting to others all require some comprehension of what the world looks like to them. Perspective-taking is undoubtedly an effective strategy in social situations and might

> occur spontaneously in anticipation of social interaction. Consequently, seeing another person in a scene might prompt thinking about the world from the others perspective."
>
> *Extracted from Tversky and Hard* [25].

> "Actions are ambiguous, so people evaluate other peoples beliefs, goals, and intentions in order to interpret their actions. Ones theory of mind provides the ability to infer other individuals mental states, to consider their perspective, and thereby to interpret and predict their actions."
>
> *Extracted from Wu and Keysar* [27].

Therefore, we believed that going deeper into this notion and related topics was indispensable for achieving the global goal of the project.

Host Institution Research Line

Breazeal [3] defines different types of social robots. Two of them are the following:

- social receptive: interactions with people affect the robot's internal structure at deeper levels, such as organizing the motor system to perform new gestures, or associating symbolic labels to incoming perceptions, learning through demonstration. Cognitive models are usual in these robots.
- sociable: they not only perceive human social cues, but also model people in social and cognitive terms in order to interact with them. They also behave proactively, not only to benefit the person (aiding in a task), but also to benefit itself (learning and improving its own skills).

Currently the host institution works with social receptive robots and is moving towards sociable robots requiring the inclusion of cognitive skills in the design of the robots. They are involved in a European Project, the CHRIS[1] project: "Cooperative Human Interaction Systems". The main goal of the project is:

> "Specifically this project addresses the problem of a human and a robot performing co-operative tasks in a co-located space, [...]. These issues include communication of a shared goal (verbally and through gesture), perception and understanding of intention (from dextrous and gross movements), cognition necessary for interaction, and active and passive compliance."

[1]The CHRIS project is funded by the European Commission's Seventh Framework Programme (FP7) under grant agreement number 215805 and runs from 2008-2012.

The project involves interdisciplinary research areas. One of them is Psychology, leaded by the Max Planck Institute for Evolutionary Anthropology. Their main contribution to the project is to provide the basic cognitive building blocks of initial interactions, those of young children, to understand human cooperative interaction. Based on the notions coming from psychology, our aim is to find models that we can integrate in a robotic system to enhance its interaction with humans partners in collaborative tasks.

During the duration of the ARBI project I have been largely involved in the CHRIS project, mainly discussing and learning with the Max Planck partners on cognitive concepts related to theory of mind, perspective taking, altruistic and cooperative behavior.

2 Cognitive Concepts

Social robots interact with people, people who have mental states defined by goals, beliefs, desires and intentions. If we aim at designing a robot who is able to interact with people it is essential that the robot understands others in terms of mental states [5]. Moreover, robots should also posses mental states and related cognitive skills in order to interact in the same manner with humans. Therefore, we believe that implementing robots based on cognitive skills is crucial. In this section I will review some basic concepts used through the report, while going through research focused on the design of robots based on cognitive abilities. We will refer to either humans or robots as agents, making the difference between them by using the word of "robotic agent" when needed.

2.1 Theory of Mind

In psychology, the ability to attribute to one self and others mental states, i.e. goals, desires, intentions, beliefs, feelings, etc. is knows as *Theory of Mind*, ToM, (also know as *mindreading*, first introduced by Premack and Woodruff in [18]). We can find a first attempt to provide a robot with a theory of mind in the work presented by Scassellati [19]. In his work, he discusses the viability of applying these notions and presents first steps towards basic cognitive skills based on visual perception. Breazeal et al.[4] proposes an embodied cognition approach for mindreading abilities implemented in the Leonardo robot. They mainly focus on two aspects: perspective taking and false beliefs attribution. I will come back to these two notions in next sections.

Autism[2] is a failure to develop social abilities, language and other communication skills to the usual level, "...causing a disorder in understanding and coping with the social environment regardless of IQ." [1]. Baron-Cohen et al. [1] prove through experimentation that autistic children are not able of attributing beliefs to others.

A well know test to evaluate if someone is able to attribute beliefs to others, even if they differ from the self-believes, is the *false beliefs* task. A classical

[2]Definition from the Cambridge Dictionary.

example is the story of two characters, Sally and Anne, who are playing with a toy. They place the toy in a box and then Sally leaves. In the meantime, Anne moves the toy to another box. Next, Sally enters the scene again looking for the toy. The question is: Where should Sally look for it, in the first box or in the second one? If the ability of false believe has evolved, the right answer is the first box, because she believes that the toy is where she last saw it. A very similar task is used in the experiments performed by Baron-Cohen et al. with autistic children. They show that autistic children are not able to give the right answer when asked about where the character should look for the object. In other words, they are aware of the current position of the object and they attribute this same belief to the character (when in fact, the character has a wrong belief about the object location). Hence, to some extent, a robotic agent that is not able to attribute beliefs to other agents could be considered as an autistic robot who will have difficulties to interact with people.

Hogrefe and Wimmer prove an intermediate level to false belief attribution in [9]. In their studies they found out that 3-4 year-old children are able to attribute *absence of knowledge* (or *ignorance*) but not false beliefs (which occurs around 4-5 years old). Thus, in the Anne-Sally story, a 3-year-old child would be able to indicate that Sally does not know where the toy is now, but at the same time, she would say that Sally believes that the toy is in the second box. Based on the results, they suggest that children are not able to represent incompatible propositions about the true state of affairs (A is true, I belief A is true, she beliefs A is false), but that they do represent that the other does not share the representation of the true state of affairs. Hence, knowing that someone does not know something is an important level to achieve attribution of false beliefs.

Researchers agree that a direct consequence of knowing others' mental states is that it allows us to anticipate and predict their behavior.

> Baron-Cohen et al. [1]: "The ability to make inferences about what other people believe to be the case in a given situation allows one to predict what they will do. This is clearly a crucial component of social skills".
>
> Scassellati [19]: "A robotic system that possessed a theory of mind would allow for social interactions between the robot and the humans... [A robot] can learn to anticipate the reactions of the observer, and can modify its own behavior accordingly. "

Let's go back to the false belief example, but this time a robot will take part of the scene. By observing Sally's action, i.e. reaching the first box, and knowing her beliefs, i.e. the toy is within the box, a robot with cognitive skills may infer that Sally's goal is to obtain the toy and thus, predict her behavior: Sally is reaching the box to get the toy. Based on this inference, the robot can then help her indicating where the toy really is, either explicitly, through language, or through gestures by referring to the other box. Gray et al. propose a model for solving the false belief task in [8, 4].

2.2 Perspective Taking

There is no clear definition of "Perspective Taking". As Traversky indicates "Perspective is one of those extraordinarily useful words that, as a consequence, has many senses" [26]. However, it seems that there are three main categories of perspective taking:

- perceptual perspective taking: ability to perceive the environment from another's point of view through senses (visual, spatial, auditory).
- conceptual perspective taking: Marvin et. al [15] refer to it as "an inference about less tangible aspects of another's internal experience such as thoughts, desires, attitudes, plans." This definition could be compared to the definition of theory of mind, as Baron-Cohen et. al also mention in [1].
- emotional perspective taking or empathy: Perner defines it as "the ability to predict what a person would feel in certain circumstances" [17].

We will focus on the first type of perspective taking, i.e. perceptual perspective taking, and more precisely, on visual and spatial perspective taking.

2.2.1 Visual Perspective Taking

Flavell [7] identifies two levels of visual perspective taking. The first one, Level 1, refers to "what the person sees", while the second one, Level 2, refers to "how the person sees". Level 1 deals with the ability of knowing than an object might be visible for a subject, but occluded for another one. One of the applications of this ability is that it allows us to clarify some ambiguous situations based on the speaker's visual perception of the environment. Psychological studies have proven that humans develop this ability at 24-month-old (Moll and Tomasello, [16]). Suppose a situation where there are two books on a desk, one is visible for both, speaker and listener, while the other one is occluded to the speaker by some other bigger object. When the speaker asks for the book to the listener, the latter should be able to reason about the fact that the speaker can only refer to one of the books (the visible one from her point of view), since the other one is occluded.

2.2.2 Spatial Perspective Taking

Spatial perspective taking refers to the the ability to compute qualitative spatial location of objects (or agents) with respect to a frame (eg. the keys on my left). Based on the frame of reference, the description of an object varies. Humans mix perspectives frequently during interaction. This is more effective than maintaining a consistent one, either because the (cognitive) cost of switching is lower than remaining with the same perspective, or if the cost is about the same, because the spatial situation may be easily described from one perspective rather than another [26]. Ambiguities arise when one speaker refers to an object within a reference system (or changes the reference system, i.e., switches

perspective) without informing her partner about it. For example, the speaker could ask for the "keys on the left". Since no reference system has been given, the listener would not know where exactly to look. However, asking for "the keys on your left" gives enough information to the listener to understand where the speaker is referring to. The reference system has to be defined properly because the terms of reference (left, right, above,...) may be identical in different systems [20]. On the contrary, when using an exact, unambiguous term of reference to describe a location (eg. "go north") no ambiguity arises.

Researchers in spatial cognition propose different ways for classifying frames of references([12, 13]). However, for simplicity, we will use the classification proposed by Trafton ([22]): egocentric (self-based), addressee centered (other-based), object centered (object-based), exocentric (world-based) and deictic. Taking into account spatial perspective taking is essential in human interaction. People refer to objects in the environment not only using their identifiers (names), but also by positioning them with respect to a frame reference. For instance, "Can you give me the book on your left?" A robotic system must be able to compute these type of locations in order to understand which object the human is talking about even if description is based on a different representation to the one the internal robot uses to model the environment.

Researchers in robotics have considered these two notions of perspectives when modeling robotic systems for human interaction. Trafton et al. [24] use both visual and spatial perspective taking for finding out the referent indicated by a human partner. In another work [23], they also design a robot that is able to play hide and seek a child does. The strategy they model in the robot is to find those places that are not visible for the human partner. Berlin et al. [2, 4] present their work focused on using visual perspective taking skills for learning from a human teacher. A teacher classifies objects in a given way. The robot then learns the classification function based on the teacher's visual perception of the world. Johnson and Demiris [10] apply visual perspective taking for action recognition. In their work, a robot who has complete visual access of the environment observes another robot with partial access performing a task. The first robot can recognize the task performed by the second robot because it is able to reason about its partial perception.

2.3 Shared Attention

Shared attention (or *joint attention*) occurs when there is an interest in sharing something among the participants. Tomasello and Carpenter indicate that shared attention takes places not only when two people look at something at the same time (which corresponds to gaze following), but also when the participants are aware about the other's focus of attention on that same thing [21]. Kaplan and Hafner [11] present a significant survey on joint attention and the underlying skills to achieve shared attention either between a robot and a human, or between two robots. They claim that for reaching joint attention, there are four prerequisites:

- attention detection: tracking the attention of another.
- attention manipulation: influence another's attention through gestures or words.
- social coordination: ability to engage in coordinated interaction (e.g. turn-taking, role-switching) with others.
- intentional understanding: view others as intentional agents (agents with ToM).

We concentrate on the first two, i.e. attention detection or *gaze following* and attention manipulation or *focusing attention*, where the latter makes use of the former.

Gaze following consists on observing what our partner is looking at. Scassellati [19] summarizes the four steps children develop to efficiently achieve this ability based on Butterworth [6]: (*i*) detecting which side where the observed person is looking at, i.e. to her left or to her right (6 months-old); (*ii*)detecting the first salient object in the observed person line of gaze (9 months-old); (*iii*)detecting any salient object in the observed person line of gaze(12 months-old); and (*iv*) following gaze on regions outside the current field of view of the observer (18 months-old). According to Kaplan, focusing attention along with gestures and words takes place in (*iii*) and (*iv*), respectively.

Focusing attention is used to direct someone's attention to bring that information into common ground. To achieve efficient interaction, common ground is essential so the involved partners are aware of the context where the interaction is taking place. It is thus another key skill to implement into a robotic system if we aim at improving human-robot interaction.

3 Research Methodology

Based on the ideas and concepts introduced above, we believe that the appropriate approach for achieving the goals and the project is to: first, model and implement some of the basic skills already mentioned; second, design the algorithms and strategies for combining them at a decisional reasoning level; and third, evaluate the approach through experimentation. I next describe each stage of the approach.

3.1 Basic Skills

Within the project, we have identified the following basic skills in terms of priority for achieving the goals and also, the feasibility for implementing them in our current cognitive architecture. Next, I will go through them, as well as indicate the current state of the work.

3.1.1 Gaze Following and Focusing Attention (towards Shared Attention)

First steps towards shared attention have been introduced in a recent thesis at the host institution [14]. Although gaze following should be modeled based on the gaze direction of the partner, in our approach we consider the head direction (using motion capture) instead of actual gaze assuming that the human is looking what it is in front of her. We were forced to adopt this simplification due to lack of the required equipment to obtain precise gaze information. However, as mentioned in the preamble, we collaborate within in the CHRIS project, where one of the partners is actually dealing with gaze following. We hope that we can use at some point, their equipment to know exactly where the human partner is looking at.

Focusing attention should be achieved through recognition of at least one of the following gestures: pointing, placing or showing. We have identified the prerequisites for implementing the generation of these three gestures, mainly visibility and "reachability" of the object of the interacting agents.

3.1.2 Perceptual Perspective Taking (visual, spatial and affordances)

Used for referring, this ability allows a robot to better interact with humans when an ambiguous situations take place, reducing unnecessary flow of information (i.e. asking the human which object she refers to exactly). We have implemented visual perspective level 1 and spatial perspective taking based on two frames of references (egocentric and addressee-centered).

3.1.3 False Belief task and Ignorance (towards Theory of Mind)

So far we have started developing the models and mechanisms that will allow the robot to have a theory of mind. We are using an ontology system to store symbolic information about the world. We also include here the robot's own beliefs and its beliefs about the human partner.

3.1.4 Dialog

Finally, using dialog to communicate is fundamental. Researchers have addressed the problem of understanding natural language long ago. Efficient models capable of completely understanding a dialog do not exist yet. However, we believe that a simplification of it combined with alternative communication and reasoning skills could be enough as a starting point for the work we address in this project. For this reason, we have designed a dialog system capable of processing written natural language in two levels:

1. syntactic: analyses the structure of the utterance, and
2. semantic: understand the meaning of the utterance fitting it in one of the following categories:

- giving information, e.g. *The bottle is yellow.*
- asking information, e.g. *Where is the bottle?*
- establishing a goal (or task), e.g. *Can you clean the table?*
- commanding, e.g. *Stop!*

3.2 Decisional Reasoning

Ambiguity rises in terms of communication and action [27]. In the former case, humans refer to objects or describe the environment without noticing that the information provided is ambiguous to her listener. In the latter case, a person executes an action thinking that is understandable to her partner, though it might not be the case. In both situations the ambiguity has to be somehow solved. Either internally, i.e. the perceiver will try to clarify the ambiguity by herself, or externally, i.e. explicitly asking the partner. We have tackled the first problem, i.e. ambiguities in communication.

3.3 Experimentation

Validations tests have been performed in order to evaluate the individual skills separately. Eventually, different combinations of these skills have led us to define a set of game scenarios. The evaluations have been performed in simulation and in two alternative robotic platforms at the host institution: HRP-2 and Jido. The aim was to evaluate the performance of the algorithms implemented and their accuracy, as well as the improvement of the overall human-interaction.

References

[1] S. Baron-Cohen, A. M. Leslie, and U. Frith. Does the autistic child have a "theory fo mind"? *Cognition*, 21:37–46, 1985. TOM, cognition.

[2] M. Berlin, J. Gray, A. L. Thomaz, and C. Breazeal. Perspective taking: An organizing principle for learning in human-robot interaction. 2006. hri, psp.

[3] C. Breazeal. Toward sociable robots. *Robotics and Autonomous Systems*, 42:167–175, 2003. hri, social robots.

[4] C. Breazeal, J. Gray, and M. Berlin. An embodied cognition approach to mindreading skills for socially intelligent robots. *The International Journal of Robotics Research*, 28(5):656–680, May 2009. hri, tom, false beliefs, psp.

[5] C. Breazeal, A. Takanishi, and T. Kobayashi. *Springer Handbook of Robotics*, chapter Social Robots that Interact with People, pages 1349–1369. Springer Berlin Heidelberg, 2008.

[6] G. Butterworth and N. Jarrett. What minds have in common in space: Spatial mechanisms serving joint visual attention in infancy. *British Journal of Developmental Psychology*, 9:55–72, 1991.

[7] J. H. Flavell. *Perspectives on Perspective Taking*, chapter 5, pages 107–139. 1992. cognitive, psp.

[8] J. Gray, M. Berlin, and C. Breazeal. Intention recognition with divergent beliefs for collaborative robots. In *Proceedings of the 2007 Artificial Intelligence and Simulation of Behavior (AISB) workshop on Mindfuil Environments*, 2007. false beliefs.

[9] G.-J. Hogrefe, H. Wimmer, and J. Perner. Ignorance versus false belief: A developmental lag in attribution of epistemic states. *Child Development*, 57(3):567–582, June 1986. cognition, false belief, ignorance.

[10] M. Johnson and Y. Demiris. Perceptual perspective taking and action recognition. *International Journal of Advanced Robotic Systems*, 4(4):301–308, December 2005. psp, action recognition.

[11] F. Kaplan and V. V. Hafner. The challenges of joint attention. *Interaction Studies*, 7(2):135–169, 2006. joint attention.

[12] W. J. M. Levelt. Some perceptual limitations on talking about space. *Limits in Perception*, pages 323–358, 1984.

[13] S. C. Levinson. Frames of reference and molyneux's question: Crosslinguistic evidence. *Language and Space*, pages 109–170, 1996.

[14] L. F. Marin-Urias. *Reasoning About Space for Human Robot Interaction*. PhD thesis, LAAS-CNRS & Université de Toulouse, 2009.

[15] R. S. Marvin, M. T. Greenberg, and D. G. Mossler. The early development of conceptual perspective taking: Distinguishing among multiple perspectives. *Child Development*, 47:511–514, 1976. psp, cognition.

[16] H. Moll and M. Tomasello. Level 1 perspective-taking at 24 months of age. *British Journal of Developmental Psychology*, 24:603–613, 2006. cognition.

[17] J. Perner, J. L. Brandl, and A. Granham. What is a perspective problem? developmental issues in belief ascription and dual identity. *International Journal for Contemporary Philosophy*, 5(2), 2003. psp, cognition.

[18] D. Premack and G. Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526, December 1978. cognition, TOM.

[19] B. Scassellati. Theory of mind for a humanoid robot. *Autonomous Robots*, 12:13–24, 2002. hri, cognition.

[20] H. A. Taylor, S. J. Naylor, R. R. Faust, and P. J. Holcomb. "could you hand me those keys on the right?" disentangling spatial reference frames using different methodologies. *Spatial Cognition and Computation*, 1(4):381–397, 1999. spatial perspective taking, feature based description, disambiguation.

[21] M. Tomasello and M. Carpenter. Shared intentionality. *Developmental Science*, 10(1):121–125, 2007. cognition.

[22] J. G. Trafton, N. L. Cassimatis, M. D. Bugajska, D. P. Brock, F. E. Mintz, and A. C. Schultz. Enabling effective human-robot interaction using perspective-taking in robots. *IEEE Transactions on systems, man and cybernetics - Part A: Systems and Humans*, 35(4):460–470, July 2005. Cognitive modeling, human–robot-interaction, perspective-taking.

[23] J. G. Trafton, A. C. Schultz, , D. Perznowski, M. D. Bugajska, W. Adams, N. L. Cassimatis, and D. P. Brock. Children and robots learning to play hide and seek. In *Proceedings of the 1st Conference on Human-Robot Interaction*, pages 242–249. ACM, 2006. hri, behavior modelling.

[24] J. G. Trafton, A. C. Schultz, M. Bugajska, and F. Mintz. Perspective-taking with robots: Experiments and models. In *IEEE International Workshop on Robots and Human Interactive Communication*, pages 580–584, 2005. hri, psp.

[25] B. Tversky and B. M. Hard. Embodied and disembodied cognition: Spatial perspective-taking. *Cognition*, 110:124–129, January 2009. spatial psp.

[26] B. Tversky, P. Lee, and S. Mainwaring. Why do speakers mix perspectives? *Spatial Cognition and Computation*, 1(4):399–412, 1999. psp, psychology.

[27] S. Wu and B. Keysar. The effect of culture on perspective taking. *Psychological Science*, 18(7):600–606, 2007. cultural influence, psp.