

Title:

The Basis of Shared Intentions in Human and Robot Cognition

Authors:

Peter Ford Dominey¹, Felix Warneken²

1. Corresponding Author

Laboratoire d'Etude des Mécanismes Cognitifs (EMC)

Equipe Neurosciences Cognitive et Représentations Multimodales (NCRM)

CNRS - Université Lumière Lyon 2 - Batiment K

5 avenue Pierre Mendès France

69676 Bron cedex, France

+33 (0) 04 78 77 30 53

Peter.Ford.Dominey ('at') univ-lyon2.fr

2. Department of Developmental and Comparative Psychology

Max Planck Institute for Evolutionary Anthropology

Deutscher Platz 6

D-04103 Leipzig, Germany

+49 (0) 341 3550 437

warneken ('at') eva.mpg.de

ABSTRACT:

There is a fundamental difference between robots that are equipped with sensory, motor and cognitive capabilities, vs. simulations or non-embodied cognitive systems. Via their perceptual and motor capabilities, these robotic systems can interact with humans in an increasingly more “natural” way, physically interacting with shared objects in cooperative action settings. Indeed, such cognitive robotic systems provide a unique opportunity to developmental psychologists for implementing their theories and testing their hypotheses on systems that are becoming increasingly “at home” in the sensory motor and social worlds, where such hypotheses are relevant. The current research is the result of interaction between research in computational neuroscience and robotics on the one hand, and developmental psychology on the other. One of the key findings in the developmental psychology context is that with respect to other primates, humans appear to have a unique ability and motivation to share goals and intentions with others. This ability is expressed in cooperative behavior very early in life, and appears to be the basis for subsequent development of social cognition. Here we attempt to identify a set of core functional elements of cooperative behavior and the corresponding shared intentional representations. We then begin to specify how these capabilities can be implemented in a robotic system, the Cooperator, and tested in human-robot interaction experiments. Based on the results of these experiments we discuss the mutual benefit for both fields of the interaction between robotics and developmental psychology.

1. INTRODUCTION:

There is a long history of interaction between theoretical aspects of psychology and the information and computer sciences. The “information processing” model of cognitive psychology developed by Neisser (1967) and Broadbent (1965) borrowed notions such as input, representation, processing and output from computer science and applied them to the analysis of mental processes. Whether or not one holds with specific application of computing metaphors to psychological theories, it

appears clear that the use of such metaphors is useful in that it confronts psychological theory with specific questions to be addressed, related to representations and processes underlying cognitive functions. Today the psychological and computing sciences are entering a new period of interaction that is linked to new technological developments in the domain of robotics. Unlike simulation and traditional artificial intelligence programs that are constrained at best to “live” in simulated artificial worlds, robots are equipped with sensory and motor capabilities that allow them to exist in the physical world of the humans that they can interact with. That is, robots can provide experimental platforms to cognitive scientists for implementing and testing theories about the intricate relation between a developing system and its physical environment. Likewise, from the robot technology perspective, robotics scientists have reasoned that the most complex behavior cannot be exclusively programmed by hand, but rather should result from adaptive and developmental mechanisms that are based on those identified in the development of physiological systems (Brooks 1990, Pfeifer 1999, Pfeifer & Gomez 2005).

One of the most interesting opportunities provided by this interaction between robotics and psychology will be in the domain of developmental psychology. Research in this domain is beginning to focus in on the functional aspects of social cognition that make humans unique in the animal world. It appears that part of the uniquely human aspects concern the ability and motivation to shared intentional states with others (Tomasello et al. 2005). The objective of the current research is to begin to identify some of the core elements of the human ability to share intentions based on experimental and theoretical results from developmental psychology, and to then begin to determine how these elements can be implemented on a corresponding robotic system designed for interacting and cooperating with humans. We believe that this work is important because it motivates psychologists to formalize their hypotheses in sufficient detail that they can lead to implementation and testing in artificial but naturally inspired cognitive systems. Of particular interest are the underlying representations required for these shared

intentions. We also believe that this work is important because it will begin to endow robots with human-like abilities to cooperate.

Tomasello et al (2005) proposed that the human ability to share intentions develops via the interaction of two distinct capabilities. The first concerns the ability to “read” or determine the intentions of other agents through observation of their behavior, and more generally the ability to represent and understand others as intentional goal directed agents. The second capability concerns the motivation to share intentions with others. While non-human and human primates are skilled at the first - reading the intentions of others based on action and gaze direction, only humans seem to possess an additional capability that will make a significant difference. This is the motivation to cooperate: to share mental states, including goal based intentions which form the basis of cooperation.

Perhaps one of the most insightful methods of establishing the properties of human social cognition is the comparison of human and great ape performance in equivalent conditions (see Tomasello & Carpenter 2007). In this context, Warneken, Chen and Tomasello (2006) engaged 18-to-24 month old children and young chimpanzees in goal-oriented tasks and social games which required cooperation. They were interested both in how the cooperation would proceed under optimal conditions, but also how the children and chimps would respond when the adult stopped performing the task. The principal finding was that children enthusiastically participate both in goal directed cooperative tasks and social games, and spontaneously attempt to reengage and help the adult when he stops. In contrast, chimpanzees are uninterested in non-goal directed social games, and appear wholly fixed on attaining food goals, independent of cooperation. Warneken et al. thus observed what appears to be a very early human capacity for (1) actively engaging in cooperative activities just for the sake of cooperation, and (2) for helping or reengaging the perturbed adult (Warneken & Tomasello 2006, Warneken et al. 2006).

In one of the social games, the experiment began with a demonstration where one participant sent

a wooden block sliding down an inclined tube and the other participant caught the block in a tin cup that made a rattling sound. This can be considered more generally as a task in which one participant manipulates an object so that the second participant can then in turn manipulate the object. This represents a minimal case of a coordinated action sequence. After the demonstration, in Trials 1 and 2 the experimenter sent the block down one of the tubes three times, and then switched to the other, and the child was required to choose the same tube as the partner. In Trials 3 and 4 during the game, the experimenter interrupted the behavior for 15 seconds and then resumed.

Behaviorally, children successfully participated in the game in Trials 1 and 2. In the interruption Trials 3 and 4 they displayed two particularly interesting types of response that were (a) to reengage the experimenter with a communicative act (on 38% of the interruption trials for 24 month olds), or less often, (b) to attempt to perform the role of the experimenter themselves (on 22% of interruption trials for 24 month olds). Though (b) was considered a non-cooperative behavior, i.e. as an attempt to solve the task individually, it still indicates that the children had a clear awareness both of their role and that of the adult in the shared coordinated activity. Importantly, after only a few demonstrations of the game (and only one demonstration for the 24 month children) it was apparent that the children had a “bird’s eye view” or third person representation of the interaction, allowing them to subsequently take either role in the game – that of the launcher or of the receiver of the sliding block. This implies a rather clever representation scheme which can keep track of the goal directed actions of multiple agents, and their interaction, allowing the observer to then take the role of either of the observed agents. In a related study, Warneken & Tomasello (2006) demonstrated that 18 and 24 month old children spontaneously help adults in a variety of situations. This is interpreted as evidence for an altruistic motivation to help, and an ability to understand and represent the goals and intentions of others. Indeed, such helping represents a mutual commitment to the shared activity which is one of the defining features of shared cooperative activity (Bratman 1992).

The ability to represent the action from multiple perspectives was examined more directly in a study of role reversal imitation conducted by Carpenter et al. (2005). In one experiment of this study, children observed the experimenter cover a “Big Bird” figurine with a cloth. The experimenter then asked the child “Where is big bird? Can you find him?” and the child (or the experimenter) lifted the cloth to reveal the toy. After three such demonstrations, the experimenter handed the cloth to the child and said “It’s your turn now.” Approximately 70% of the 21 18 month old children tested successfully performed the role reversal. Again, this suggests that the child maintains a representation of the alternating roles of both participants in a third-person perspective that can then be used to allow the child to take on either of the two roles.

In order to begin to think about how such a system has come to be (and could be built), we can look to recent results in human and primate neurophysiology and neuroanatomy. It has now become clearly established that neurons in the parietal and the premotor cortices encode simple actions both for the execution of these actions as well as for the perception of these same actions when they performed by a second agent (di Pellegrino et al. 1992 , Rizzolatti & Craighero 2004). This research corroborates the emphasis from behavioral studies on the importance of the goal (rather than the details of the means) in action perception (Bekkering et al. 2000, Carpenter & Call 2007, Sommerville & Woodward 2005, Tomasello et al. 2005). It has been suggested that these premotor and parietal “mirror” neurons play a crucial role in imitation, as they provide a common representation for the perception and subsequent execution of a given action. Interestingly, however, it has been clearly demonstrated that the imitation ability of non-human primates is severely impoverished when compared to that of humans (see Rizzolatti & Craighero 2004, Tomasello et al. 2005). This indicates that the human ability to imitate novel actions and action sequences in real time (i.e. after only one or two demonstrations) relies on additional neural mechanisms to those found in non-human primates.

In this context, a recent study of human imitation learning (Buchine et al. 2004) implicates

Brodmann's area (BA) 46 as responsible for orchestrating and selecting the appropriate actions in novel imitation tasks. We have recently proposed that BA 46 participates in a dorsal stream mechanism for the manipulation of variables in abstract sequences and language (Dominey et al. 2006). Thus, variable "slots" that can be instantiated by arbitrary motor primitives during the observation of new behavior sequences are controlled in BA 46, and their sequential structure is under the control of corticostriatal systems which have been clearly implicated in sensorimotor sequencing (see Dominey et al. (2006)). This allows us to propose that this evolutionarily more recent cortical area BA 46 may play a crucial role in allowing humans to perform compositional operations (i.e. sequence learning) on more primitive action representations in the ventral premotor and parietal motor cortices. In other words, ventral premotor and parietal cortices instantiate shared perceptual and motor representations of atomic actions, and BA46 provides the capability to compose arbitrary sequences of these atomic actions, while relying on well known corticostriatal neurophysiology for sequence storage and retrieval. The functional result is the human ability to observe and represent novel behavioral action sequences. We further claim that this system can represent behavioral sequences from the "bird's eye view" or third person perspective, as required for the cooperative tasks of Warneken et al. (2006). That is, it can allow one observer to perceive and form an integrated representation of the coordinated actions of two other agents engaged in a cooperative activity. The observer can then use this representation to step in and play the role of either of the two agents. This is a "dialogic cognitive representation," or "we intention" in that it represents the "dialog" of interaction between agents.

Given this overview of some of the core functional elements of cooperative behavior and the corresponding representations (including the "bird's eye view"), we can now take on the task of beginning to specify how these capabilities can be implemented in a robotic system, and tested in human-robot interaction experiments. When making the transition from human behaviour to technological implantation, there is the risk that the implementation will be biased in terms of specific

computational or functionalist solutions. In this context, we are making a concerted effort in “cognitive systems engineering,” a process in which the cognitive robotics systems we build are constrained by (1) functional requirements (i.e. specification of how the system behaves) derived from behaviour from developmental psychology, and (2) architectural constraints from the neurosciences. To as large a degree as possible, we avoid arbitrary constraints from the purely computational aspects of the implementation platform.

2. THE ROBOTIC SYSTEM – THE COOPERATOR

In the current experiments the human and robot cooperate by moving physical objects to different positions in a shared work-space as illustrated in Figures 1 and 2. The cooperative activity will involve interactive tasks that preserve the important aspects of the “block launching” task of Warneken et al., transposed into a domain of objects suitable for our robot system. The 4 moveable objects are pieces of a wooden puzzle, representing a dog, a pig, a duck and a cow. These pieces can be moved by the robot and the user in the context of cooperative activity. Each has fixed to it a vertically protruding metal screw, which provides an easy grasping target both for the robot and for humans. In addition there are 6 images that are fixed to the table and serve as landmarks for placing the moveable objects, and correspond to a light, a turtle, a hammer, a rose, a lock and a lion, as partially illustrated in Figures 1 & 2. In the interactions, human and robot are required to place objects in zones next to the different landmarks, so that the robot can more easily determine where objects are, and where to grasp them. Figure 1 provides an overview of the architecture, and Figure 2, which corresponds to Experiment 6 provides an overview of the actual physical state of affairs during a cooperative interaction.

2.1 Representation

The structure of the internal representations is a central factor determining how the system will function, and how it will generalize to new conditions. Based on the neurophysiology reviewed above, we use a common representation of action for both perception and production. In the context of the current study, actions involve moving objects to different locations, and are identified by the agent, the object, and the target location the object is moved to. As illustrated in Figure 1, by taking the “short loop” from vision, via Current Action Representation, to Motor Command, the system is thus configured for a form of goal-based action imitation. This will be expanded upon below.

In order to allow for more elaborate cooperative activity, the system must be able to store and retrieve actions in a sequential structure, and must be able to associate each action with its agent. We thus propose that the ability to store a sequence of actions, each tagged with its agent, provides an initial capability for dialogic cognitive representation. This form of real time sequence learning for imitation is not observed in non-human primates (see Rizzolatti & Craighero 2004). In this context, an fMRI study (Buchine et al. 2004) which addressed the human ability to observe and program arbitrary actions indicated that a cortical area (BA46) which is of relatively recent phylogenetic origin is involved in such processes. Rizzolatti and Craighero (2004) have thus suggested that the BA 46 in man will orchestrate the real-time capability to store and retrieve recognized actions, and we can further propose that this orchestration will recruit canonical brain circuitry for sequence processing including the cortico-striatal system (see Dominey 2005, and Dominey et al. 2006 for discussion of such sequence processing).

An additional important representational feature of the system is the World Model that represents the physical state of the world, and can be accessed and updated by vision, motor control, and language, similar to the Grounded Situation Model of Mavridis and Roy (2006). The World Model encodes the physical locations of objects and is updated by vision and proprioception (i.e. robot action updates World Model with new object location). Changes observed in the World Model in terms of an object

being moved allows the system to detect actions in terms of these object movements. Actions are represented in terms of the agent, the object and the goal of the action, in the form MOVE(object, goal location, agent). These representations can be used for commanding action, for describing recognized action, and thus for action imitation and narration, as seen below.

In the current study we address behavioral conditions which focus on the observation and immediate re-use of an intentional (goal directed) action plan. However, in the more general case, one should consider that multiple intentional action plans can be observed and stored in a repertory (IntRep or Intentional Plan Repertory in Figure 1). When the system is subsequently observing the behavior of others, it can compare the ongoing behavior to these stored sequences. Detection of a match with the beginning of a stored sequence can be used to retrieve the entire sequence. This can then be used to allow the system to “jump into” the scenario, to anticipate the other agent’s actions, and/or to help that agent if there is a problem.

2.2 Visual perception

Visual perception is a challenging technical problem. To simplify, standard lighting conditions and a small set ($n = 10$) of visual objects to recognize are employed (4 moveable objects and 6 location landmarks). A VGA webcam is positioned at 1.25 meters above the robot workspace. Vision processing is provided by the Spikenet Vision System (<http://www.spikenet-technology.com/>). Three recognition models for each object at different orientations (see Fig. 3) were built with an offline model builder. During real-time vision processing, the models are recognized, and their (x, y) location in camera coordinates are provided. Our vision post-processing eliminates spurious detections and returns the reliable (x, y) coordinates of each moveable object. The nearest of the 6 fixed landmarks is then calculated in order to localize the object.

2.3 Motor Control & Visual-Motor Coordination

While visual-motor coordination is not the focus of the current work, it was necessary to provide some primitive functions (i.e. visually guided object grasping) to allow goal directed action. All of the robot actions, whether generated in a context of imitation, spoken command or cooperative interaction will be of the form *move(x to y)* where *x* is a member of a set of visually perceivable objects, and *y* is a member of the set of 6 fixed landmark locations on the work plan.

Robot motor control for transport and object manipulation with the Cooperator's two finger gripper is provided by the 6 degree of freedom Lynx6 arm (www.lynxmotion.com). The 6 motors of the arm are coordinated by a parallel controller connected to a PC computer that provides transmission of robot commands over the RS232 serial port.

Human users (and the robot Cooperator) are constrained when they move an object, to place it in one of the zones designated next to each of the six landmarks (see Fig 3). This way, when the nearest landmark for an object has been determined, this is sufficient for the robot to grasp that object at the prespecified zone.

In a calibration phase, target points are marked next to each of the 6 fixed landmark locations, such that they are all on an arc that is equidistant to the center of rotation of the robot arm base. For each, the rotation angle of Joint 0 (the rotating shoulder base of the robot arm) necessary to align the arm with that point is then determined. We then determined a common set of joint angles for Joints 1 – 5 that position the gripper to seize an object once the should angle is established. Angles for Joint 6 that controls the closing and opening of the gripper to grasp and release an object were then identified. Finally a neutral position to which the arm could be returned in between movements was defined. The system was thus equipped with a set of action primitives that could be combined to position the robot at any of the 6 grasping locations, grasp the corresponding object, move to a new position, and place the object there.

2.4 Cooperation Control Architecture

The spoken language control architecture illustrated in Fig 4 is implemented with the CSLU Rapid Application Development toolkit (<http://cslu.cse.ogi.edu/toolkit/>). This system provides a state-based dialog management system that allows interaction with the robot (via the serial port controller) and with the vision processing system (via file i/o). Most importantly it also provides the spoken language interface that allows the user to determine what mode of operation he and the robot will work in, and to manage the interaction via spoken words and sentences.

Figure 4 illustrates the flow of control of the interaction management. In the Start state the system first visually observes where all of the objects are currently located. From the start state, the system allows the user to specify if he wants to ask the robot to perform actions via spoken commands (Act), to imitate the user, or to play (Imitate/Play). In the Act state, the user can specify actions of the form “Put the dog next to the rose” and a grammatical construction template (Dominey et al. 2003, Dominey et al. 2005, Dominey & Boucher 2005, Dominey et al. 2004, Dominey et al. 2006) is used to extract the action that the robot then performs, in the form *Move(object, location)*. In the Imitate state, the robot first verifies the current state (Update World) and then invites the user to demonstrate an action (Invite Action). The user shows the robot one action. The robot then begins to visually observe the scene until it detects the action, based on changes in object locations detected (Detect Action). This action is then saved and transmitted (via Play the Plan with Robot as Agent) to execution (Execute action). A predicate(argument) representation of the form *Move(object, landmark)* is used both for action observation and execution. Imitation is thus a minimal case of Playing in which the “game” is a single action executed by the robot.

The more general case corresponds to “games” in which the robot and human will take turns in the execution of a shared plan. In the current implementation of this, the user can demonstrate multiple successive actions, and indicate the agent (by saying “You/I do this”) for each action. Improvements in

the visual processing will allow the more general case in which the system can observe two agents interacting and attribute each action to its respective agent.

The resulting intentional plan specifies what is to be done by whom. When the user specifies that the plan is finished, the system moves to the Save Plan, and then to the Play Plan states. For each action, the system recalls whether that action is to be executed by the robot or the user. Robot execution takes the standard Execute Action pathway. User execution performs a check (based on user response) concerning whether the action was correctly performed or not. Interestingly, the ability of the robot to “help” the user comes quite naturally, based on the shared intentional plan. If the user action is not performed, the robot “knows” the failed action based on its own representation of the plan. The robot can thus communicate with the user, and if the user agrees, the robot can help by performing the action itself. Thus, “helping” was quite naturally implemented by combining an evaluation of the user action, with the existing capability to perform a stored action representation. Still, it is worth noting that one crucial difference between the helping by the robot and what Warneken et al. tested in the helping study (Warneken & Tomasello 2006) was that the children and chimpanzees helped the other *with* their action, not just performing the other’s action completely, but complementing the other’s action.

2.5 “Bird’s Eye View and Role Reversal

In an initial set of experiments (Experiments 1-6 below), the “intentional plan” was represented for the robot as a sequence of actions in the “We Intention” of Figure 1, with the attribution of the agent fixed for each action. We know however from the experimental results of Warneken et al. (2006), and from the role reversal studies of Carpenter et al. (2005) that this representation is flexible, in the sense that the child can take on the role of either of the two represented agents. Once the adult indicates the role he takes, the child then spontaneously adapts and takes the other role. In the current system, we thus introduce a new capability in which, prior to the playing of the game, the roles can be determined

and modified. When control reaches the “Plan Play” node in the controller (Figure 4), i.e. after a new game has been demonstrated, or after the user chooses to play the old game, the robot now asks the user if he wants to go first. If the user responds yes, then the roles of user and robot remain as they were in the demonstration. If the user says no, then the roles are reversed. Reversal corresponds to systematically reassigning the agents (i.e. robot or user) associated with each action. Indeed, technically it would be possible that based upon the first move by the user (or the users insistent waiting for the robot to start), the robot infers who does what (i.e. whether to reverse roles or not) and what role it will take in the cooperative plan, though this has was not implemented in the current version of the system.

3. EXPERIMENTAL RESULTS

For each of the 6 following experiments, equivalent variants were repeated at least ten times to demonstrate the generalized capability and robustness of the system. In less than 5 percent of the trials overall, errors of two types were observed to occur. Speech errors resulted from a failure in the voice recognition, and were recovered from by the command validation check (Robot: “Did you say ...?”). Visual image recognition errors occurred when the objects were rotated beyond 20° from their upright position. These errors were identified when the user detected that an object that should be seen was not reported as visible by the system, and were corrected by the user re-placing the object and asking the system to “look again”. At the beginning of each trial the system first queries the vision system, and updates the World Model with the position of all visible objects. It then informs the user of the locations of the different objects, for example “The dog is next to the lock, the horse is next to the lion.” It then asks the user “Do you want me to act, imitate, play or look again?”, and the user responds with one of the action-related options, or with “look again” if the scene is not described correctly.

3.1 Experiment 1: Validation of Sensorimotor Control

In this experiment, the user says that he wants the “Act” state (Fig 4), and then uses spoken commands such as “Put the horse next to the hammer”. Recall that the horse is among the moveable objects, and hammer is among the fixed landmarks. The robot requests confirmation and then extracts the predicate-argument representation - $Move(X\ to\ Y)$ - of the sentence based on grammatical construction templates. In the Execute Action state, the action $Move(X\ to\ Y)$ is decomposed into two components of $Get(X)$, and $Place-At(Y)$. $Get(X)$ queries the World Model in order to localize X with respect to the different landmarks, and then performs a grasp at the corresponding landmark target location. Likewise, $Place-At(Y)$ simply performs a transport to target location Y and releases the object. Decomposing the *get* and *place* functions allows the composition of all possible combinations in the $Move(X\ to\ Y)$ space. Ten trials were performed moving the four objects to and from different landmark locations. In these ten experimental runs, the system performed correctly. Experiment 1 thus demonstrates that the system has (1) the ability to transform a spoken sentence into a $Move(X\ to\ Y)$ command, (2) the ability to perform visual localization of the target object, and (3) the sensory-motor ability to grasp the object and put it at the specified location. .

3.2 Experiment 2: Imitation

In this experiment the user chooses the “imitate” state. As stated above, imitation is centered on the achieved ends – in terms of observed changes in state – rather than the detailed trajectory or means by which these ends were achieved (Bekkering et al. 2000, Carpenter et al. 2005). Before the user performs the demonstration of the action to be imitated, the robot queries the vision system, and updates the World Model (Update World in Fig 4) and then invites the user to demonstrate an action. The robot pauses, and then again queries the vision system and continues to query until it detects a difference between the currently perceived world state and the previously stored World Model (in State Comparator of Fig 1, and Detect Action in Fig 4), corresponding to an object displacement. Extracting

the identity of the displaced object, and its new location (with respect to the nearest landmark) allows the formation of an *Move(object, location)* action representation. Before imitating, the robot operates on this representation with a meaning-to-sentence construction in order to verify the action to the user, as in “Did you put the dog next to the rose?” It then asks the user to put things back as they were so that it can perform the imitation. At this point, the action is executed (Execute Action in Fig 4). In ten experimental runs the system performed correctly. This demonstrates (1) the ability of the system to detect the final “goal” of user-generated actions as defined by visually perceived state changes, and (2) the utility of a common representation of action for perception, description and execution.

3.3 Experiment 3: A Cooperative Game

The cooperative game is similar to imitation, except that there is a sequence of actions (rather than just one), and the actions can be effected by either the user or the robot in a cooperative, turn taking manner. In this experiment, the user responds to the system request and enters the “play” state. In what corresponds to the demonstration in Warneken et al. (2006) the robot invites the user to start showing how the game works. Note that in these experiments, two experimenters demonstrate the game and the subject is observing this interaction from a third-person-perspective. The experimenters invite the child to see how the game works by showing it to them first and then have them participate afterwards. For technical limitations of the visual system, we currently use the following modification: The user then begins to perform a sequence of actions that are observed by the robot. For each action, the user specifies who does the action, i.e. either “you do this” or “I do this”. The intentional plan is thus stored as a sequence of action-agent pairs, where each action is the movement of an object to a particular target location. Note that because the system can detect actions, if it is capable of identifying distinct users (by some visual cue on their hands for example) then the system could observe two humans perform the task, thus adhering more closely to the protocol of Warneken et al. 2006. In Fig 1, the resulting

interleaved sequence is stored as the “We intention”, i.e. an action sequence in which there are different agents for different actions. When the user is finished he says “play the game”. The robot then begins to execute the stored intentional plan. During the execution, the “We intention” is decomposed into the components for the robot (Me Intention) and the human (You intention).

In one run, during the demonstration, the user said “I do this” and moved the horse from the lock location to the rose location. He then said “you do this” and moved the horse back to the lock location. After each move, the robot asks “Another move, or shall we play the game?” When the user is finished demonstrating the game, he replies “Play the game.” During the playing of this game, the robot announced “Now user puts the horse by the rose”. The user then performed this movement. The robot then asked the user “Is it OK?” to which the user replied “Yes”. The robot then announced “Now robot puts the horse by the lock” and performed the action. In two experimental runs of different demonstrations, and 5 runs each of the two demonstrated games, the system performed correctly. This demonstrates that the system can learn a simple intentional plan as a stored action sequence in which the human and the robot are agents in the respective actions.

3.4 Experiment 4: Interrupting a Cooperative Game

In this experiment, everything proceeds as in experiment 3, except that after one correct repetition of the game, in the next repetition, when the robot announced “Now user puts the horse by the rose” the user did nothing. The robot asked “Is it OK” and during a 15 second delay, the user replied “no”. The robot then said “Let me help you” and executed the move of the horse to the rose. Play then continued for the remaining move of the robot. This illustrates how the robot’s stored representation of the action that was to be performed by the user allowed the robot to “help” the user.

3.5 Experiment 5: A More Complex Game

Experiment 3 represented the simplest behavior that could qualify as a cooperative action sequence. In order to more explicitly test the intentional sequencing capability of the system, this experiment replicates Exp 3 but with a more complex task, illustrated in Figure 2. In this game (Table 1), the user starts by moving the dog, and after each move the robot “chases” the dog with the horse, until they both return to their starting places.

Action	User identifies agent	User Demonstrates Action	Ref in Figure 2
1.	I do this	Move dog from the lock to the rose	B
2.	You do this	Move the horse from the lion to the lock	B
3.	I do this	Move the dog from the rose to the hammer	C
4.	You do this	Move the horse from the lock to the rose	C
5.	You do this	Move the horse from the rose to the lion	D
6.	I do this	Move the dog from the hammer to the lock	D

Table 1. Cooperative “horse chase the dog” game specified by the user in terms of who does the action (indicated by saying) and what the action is (indicated by demonstration). Illustrated in Figure 2.

As in Experiment 3, the successive actions are visually recognized and stored in the shared “We Intention” representation. Once the user says “Play the game”, the final sequence is stored, and then during the execution, the shared sequence is decomposed into the robot and user components based on the agent associated with each action. When the user is the agent, the system invites the user to make the next move, and verifies (by asking) if the move was OK. When the system is the agent, the robot

executes the movement. After each move the World Model is updated. As in Exp 3, two different complex games were learned, and each one “played” successfully 5 times. This illustrates the learning by demonstration (Zöllner et al. 2004) of a complex intentional plan in which the human and the robot are agents in a coordinated and cooperative activity.

3.6 Experiment 6: Interrupting the Complex Game

As in Experiment 4, the objective was to verify that the robot would take over if the human had a problem. In the current experiment this capability is verified in a more complex setting. Thus, when the user is making the final movement of the dog back to the “lock” location, he fails to perform correctly, and indicates this to the robot. When the robot detects failure, it reengages the user with spoken language, and then offers to fill in for the user. This is illustrated in Figure 2H. This demonstrates the generalized ability to help that can occur whenever the robot detects the user is in trouble.

3.7 Experiment 7: Role reversal in the Complex Game

Carpenter et al. (2005) demonstrated that 18 month old children can observe and participate in a cooperative turn-taking task, and then reverse their role, indicating that they develop a third person “bird’s eye view” perspective of the interaction. The current experiment tests the ability of the system to benefit from the “bird’s eye view” representation of the shared intentional plan in order to take either role in the plan. In one test, the same “old game” from experiments 5 and 6 was used, with the modified version of the system that asks, prior to playing the game “do you want to go first”. To test the role reversal, the human responds “no”. In the demonstrated game, the human went first, so the “no” response constitutes a role reversal. The system thus systematically reassigns the You and Me actions of the We intention in Figure 1. Once this reassignment has been made, then the shared plan execution

mechanism proceeds in the standard manner. The system successfully performed this role reversal. Again, it is technically feasible for the robot to infer its own role based upon only what the user does, by detecting whether or not the user initiates the first action in the game, and such an implementation will be pursued in our future work.

4. DISCUSSION

Significant progress has been made in identifying some of the fundamental characteristics of human cognition in the context of cooperative interaction, particularly with respect to social cognition (Fong et al. 2003, Goga & Billard 2005, Kozima & Yano 2001, Lieberman 2007). Breazeal and Scassellati (2001) investigate how perception of socially relevant face stimuli and object motion will both influence the emotional and attentional state of the system and thus the human-robot interaction. Scassellati (2002) further investigates how developmental theories of human social cognition can be implemented in robots. In this context, Kozima and Yano (2001) outline how a robot can attain intentionality – the linking of goal states with intentional actions to achieve those goals – based on innate capabilities including: sensory-motor function and a simple behavior repertoire, drives, an evaluation function, and a learning mechanism.

The abilities to observe an action, determine its goal and attribute this to another agent are all clearly important aspects of the human ability to cooperate with others. The current research demonstrates how these capabilities can contribute to the “social” behavior of learning to play a cooperative game, playing the game, and helping another player who has gotten stuck in the game, as displayed in 18-24 month old children (Warneken et al. 2006, Warneken & Tomasello 2006). While the primitive basis of such behavior is visible in chimpanzees, its full expression is uniquely human (see Warneken et al. 2006 and Warneken & Tomasello 2006). As such, it can be considered a crucial component of human-like behavior for robots.

The current research is part of an ongoing effort to understand aspects of human social cognition

by bridging the gap between cognitive neuroscience, simulation and robotics (Boucher & Dominey 2006, Dominey et al. 2003, Dominey et al. 2005, Dominey & Boucher 2005, Dominey et al. 2004, Dominey et al. 2006). The experiments presented here indicate that functional requirements derived from human child behavior and neurophysiological constraints can be used to define a system that displays some interesting capabilities for cooperative behavior in the context of imitation. Likewise, they indicate that evaluation of another's progress, combined with a representation of his/her failed goal provides the basis for the human characteristic of "helping." This may be of interest to developmental scientists, and the potential collaboration between these two fields of cognitive robotics and human cognitive development is promising. The developmental cognition literature lays out a virtual roadmap for robot cognitive development (Tomasello et al. 2005, Dominey 2005). In this context, we are currently investigating the development of hierarchical means-end action sequences (Sommerville & Woodward 2005).. At each step, the objective will be to identify the characteristic underlying behavior and to implement it in the most economic manner in this continuously developing system for human-robot cooperation.

Here we begin to address the mechanisms that allow agents to make changes in perspective. In the experiments of Warneken et al. the child watched two adults perform a coordinated task (one adult launching the block down the tube, and the other catching the block). At 18-24 months, the child can thus observe the two roles being played out, and then step into either role (Carpenter et al. 2005). This indicates a "bird's eye view" representation of the cooperation, in which rather than assigning "me" and "other" agent roles from the outset, the child represents the two distinct agents A and B, and associates one of these with each action in the cooperative sequence. Then, once the perspective shift is established (by the adult taking one of the roles, or letting the child choose one) the roles A and B are assigned to me and you (or vice versa) as appropriate.

This is consistent with the system illustrated in Figure 1. We could improve the system: rather

than having the user tell the robot “you do this” and “I do this,” the vision system can be modified to recognize different agents who can be identified by saying their name as they act, or via visually identified cues on their acting hands. In the current system we demonstrate that the roles associated with “you” and “me” can be reversed. More generally, they can also be dissociated from “you” and “me” and linked with other agents. The key is that there is a central representation corresponding to the “We intention” in Figure 1, which allows the “bird’s eye view”, and a remapping mechanism that can then assign these component actions to their respective agents (corresponding to the Me and You intentions in Figure 1). Clearly there remains work to be done in this area, but the current results represent a first step in specifying how these intentional representations could be implemented.

Indeed, we take a clear position in terms of internal representational requirements, defined by a hybrid form of representation. At one level, online action and perception are encoded in an “embodied” form in terms of joint angles, and continuous values from the visual system. At a different level, “we intentions” which allow an extension in time, are distinct sequences of predicate-argument propositional elements. Thus there is a continuum of embodiment and representation. In the context of representing a joint activity through observation – the action perception is linked to the sensorimotor system, but the system that stores and replays these sequences can be considered to be independent. Indeed, it is this simulation capability that might well provide the basis for abstract processing (Barsalou 1999) More broadly speaking, though the demands of requiring implementation, robot experiments such as these can help us to shed further light on the nature and necessity of internal representations

An important open issue that has arisen through this research has to do with inferring intentions. The current research addresses one cooperative activity at a time, but nothing prevents the system from storing multiple such intentional plans in a repertory (IntRep in Fig 1). In this case, as the user begins to perform a sequence of actions involving himself and the robot, the robot can compare this ongoing sequence to the initial subsequences of all stored sequences in the IntRep. In case of a match, the robot

can retrieve the matching sequence, and infer that it is this that the user wants to perform. This can be confirmed with the user and thus provides the basis for a potentially useful form of learning for cooperative activity. Indeed, this development in the robotics context provides interesting predictions about how these inferences will be made that can be tested with children.

A potential criticism of this work could hold that while it might demonstrate an interesting and sophisticated simulation, everything of interest seems to be built in rather than emergent or developed, thus of relatively thin relevance to psychologists. We would respond that any implementation must make choices about what is built in and what is emergent. Here we have built in functions that provide the ability to perceive actions, encode action-agent sequences, and to use these sequences in behaviour. What results is the open ended capability to learn arbitrary cooperative behaviors, to help, and to changes perspectives/roles. The relevance to psychologists is twofold, in terms of what the resulting system can do, and in terms of where it fails.

Thus, while we have begun to implement some aspects of these intention representations, we should also stress how the robot's capabilities still differ from what these young children already do, including the following. (1) Children learn intentional plans quickly without direct teaching, but just by observing from the outside how two people interact. (2) They are not told who performs which role, but they themselves are able to parse the interaction into roles. (3) They spontaneously provide help without the experimenter asking them for help and without them asking the experimenter whether he wants help. (4) They not only help the other with his role but they insist on the partner performing his role when he interrupts. In other words, they seem to insist on the joint commitment to perform the respective roles. For the most part, these differences are "peripheral" in that they are related to the perception and action capabilities, rather than to the structure of internal representations. Point (1) will rely on a "salience" system that determines what behavior is interesting and merits learning (perhaps any behavior between multiple agents operating on the same objects). Point (2) will require vision processing that allows

identification of different individuals. For points (3) and (4), the behavior is currently available, i.e. it is wholly feasible for the robot to help and to insist that the other partner participates spontaneously as the situation requires.

In conclusion, the current research has attempted to build and test the Cooperator, a robotic system for cooperative interaction with humans, based on behavioral and neurophysiological requirements derived from the respective literatures. The interaction involves spoken language and the performance and observation of actions in the context of cooperative action. The experimental results demonstrate a rich set of capabilities for robot perception and subsequent use of cooperative action plans in the context of human-robot cooperation. This work thus extends the imitation paradigm into that of sequential behavior, in which the learned intentional action sequences are made up of interlaced action sequences performed in cooperative and flexible alternation by the human and robot. While many technical aspects of robotics (including visuomotor coordination and vision) have been simplified, we believe that this work makes a useful contribution in demonstrating how empirical and theoretical results in developmental psychology can be formalized to the extent that they can be implemented and tested in a robotic system. In doing so, we gain further insight into the core functions required for cooperation, and help to increase the cooperative capabilities of robots in human-robot interaction.

5. ACKNOWLEDGEMENTS

We thank Mike Tomasello, Malinda Carpenter and Elena Lieven for useful discussions during a visit of PFD to the MPI EVA in Leipzig concerning shared intentions; and Giacomo Rizzolatti for insightful discussion concerning the neurophysiology of sequence imitation at the IEEE Humanoids meeting in Genoa 2006. This research is supported in part by the French Minister of Research under grant ACI-TTT, and by the LAFMI.

6. REFERENCES

- Barsalou LW (1999) Perceptual symbol systems, *Behavioral and Brain Sciences*, 22, 577-660
- Bekkering H, Wohlschläger A, Gattis M (2000) Imitation of Gestures in Children is Goal-directed, *The Quarterly Journal of Experimental Psychology: Section A*, 53, 153-164
- Billard A, Schaal (2006) Special Issue: The Brain Mechanisms of Imitation Learning, *Neural Networks*, 19(1) 251-338
- Boucher J-D, Dominey PF (2006) Programming by Cooperation: Perceptual-Motor Sequence Learning via Human-Robot Interaction, *Proc. Simulation of Adaptive Behavior*, Rome 2006.
- Bratman ME (1992) Shared cooperative activity, *The Philosophical Review*, Vol 101, No. 2, 327-341.
- Breazeal C., Scassellati B., (2001) Challenges in building robots that imitate people, in: K. Dautenhahn, C. Nehaniv (Eds.), *Imitation in Animals and Artifacts*, MIT Press, Cambridge, MA,.
- Broadbent D (1965) Information Processing in the Nervous System, *Science*. 1965 Oct 22;150(695):457-62
- Brooks RA (1990) Elephants Don't Play Chess, *Robotics and Autonomous Systems*, 6(3) 3-15.
- Buchine G, Vogt S, Ritzl A, Fink GR, Zilles K, Freund H-J, Rizzolatti G (2004) Neural circuits Underlying Imitation Learning of Hand Actions: An Event-Related fMRI Study. *Neuron*, (42) 323-334.
- Carpenter M, Call J (2007) The question of 'what to imitate': inferring goals and intentions from demonstrations, in Christopher L. Nehaniv and Kerstin Dautenhahn Eds, *Imitation and Social Learning in Robots, Human sand Animals*, Cambridge University Press, Cambridge.

- Carpenter M, Tomasello M, Striano T (2005) Role reversal imitation and language in typically developing infants and children with Autism, *Infancy*, 8(3) 253-287
- Cuijpers RH, van Schie HT, Koppen M, Erhagen W, Bekkering H (2006) Goals and means in action observation: A computational approach, *Neural Networks* 19, 311-322,
- di Pellegrino G, Fadiga L, Fogassi L, Gallese V, Rizzolatti G (1992) Understanding motor events: a neurophysiological study. *Exp Brain Res.*;91(1):176-80.
- Dominey, P.F., (2003) Learning grammatical constructions from narrated video events for human–robot interaction. *Proceedings IEEE Humanoid Robotics Conference*, Karlsruhe, Germany
- Dominey PF (2005) From sensorimotor sequence to grammatical construction: Evidence from Simulation and Neurophysiology, *Adaptive Behavior*, 13, 4 : 347-362
- Dominey PF (2005) Toward a construction-based account of shared intentions in social cognition. Comment on Tomasello et al. 2005, *Beh Brain Sci.* 28:5, p. 696.
- Dominey PF, Alvarez M, Gao B, Jeambrun M, Weitzenfeld A, Medrano A (2005) Robot Command, Interrogation and Teaching via Social Interaction, *Proc. IEEE Conf. On Humanoid Robotics 2005*.
- Dominey PF, Boucher (2005) Learning To Talk About Events From Narrated Video in the Construction Grammar Framework, *Artificial Intelligence*, 167 (2005) 31–61
- Dominey, P. F., Boucher, J. D., & Inui, T. (2004). Building an adaptive spoken language interface for perceptually grounded human–robot interaction. In *Proceedings of the IEEE-RAS/RSJ international conference on humanoid robots*.
- Dominey PF, Hoen M, Inui T. (2006) A neurolinguistic model of grammatical construction processing. *Journal of Cognitive Neuroscience*.18(12):2088-107.
- Ellwood CA (1901) The Theory of Imitation in Social Psychology *The American Journal of Sociology*, Vol. 6, No. 6 (May, 1901), pp. 721-741

- Fong T, Nourbakhsh I, Dautenhahn K (2003) A survey of socially interactive robots. *Robotics and Autonomous Systems*, 42 3-4, 143-166.
- Goga, I., Billard, A. (2005), Development of goal-directed imitation, object manipulation and language in humans and robots. In M. A. Arbib (ed.), *Action to Language via the Mirror Neuron System*, Cambridge University Press (in press).
- Kozima H., Yano H. (2001) A robot that learns to communicate with human caregivers, in: *Proceedings of the International Workshop on Epigenetic Robotics*,.
- Lieberman MD (2007) Social Cognitive neuroscience: A Review of Core Processes, *Annu. Rev. Psychol.* (58) 18.1-18.31
- Lauria S, Buggmann G, Kyriacou T, Klein E (2002) Mobile robot programming using natural language. *Robotics and Autonomous Systems* 38(3-4) 171-181
- Mavridis N, Roy D (2006). Grounded Situation Models for Robots: Where Words and Percepts Meet. *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*
- Nehaniv CL, Dautenhahn K eds. (2002) *Imitation in Animals and Artifacts*; MIT Press, Cambridge MA.
- Nehaniv CL, Dautenhahn K eds. (2007) *Imitation and Social Learning in Robots, Humans and Animals*, Cambridge University Press, Cambridge.
- Neisser, U (1967) *Cognitive psychology* Appleton-Century-Crofts New York
- Oztop E, Kawato M, Arbib M (2006) Mirror neurons and imitation: A computationally guided review. *Neural Networks*, (19) 254-271
- Pfeifer, R. & Gómez, G. (2005), "Interacting with the real world: design principles for intelligent systems", *Artificial Life and Robotics*, **9**(1), pp. 1-6
- Pfeifer, R. & Scheier, C. (1999), *Understanding Intelligence*, Cambridge, MA: The MIT Press
- Rizzolatti G, Craighero L (2004) The Mirror-Neuron system, *Annu. Rev. Neuroscience* (27) 169-192
- Scassellati B (2002) Theory of mind for a humanoid robot, *Autonomous Robots*, 12(1) 13-24

- Sommerville A, Woodward AL (2005) Pulling out the intentional structure of action: the relation between action processing and action production in infancy. *Cognition*, 95, 1-30.
- Tomasello, M., & Carpenter M. (2007). Shared intentionality. *Developmental Science*, 10 (1), 121-125.
- Tomasello M, Carpenter M, Call J, Behne T, Moll HY (2005) Understanding and sharing intentions: The origins of cultural cognition, *Beh. Brain Sc.*; 28; 675-735.
- Warneken F, Tomasello M (2006) Altruistic helping in human infants and young chimpanzees, *Science*, 311, 1301-1303
- Warneken F, Chen F, Tomasello M (2006) Cooperative Activities in Young Children and Chimpanzees, *Child Development*, 77(3) 640-663.
- Zöllner R., Asfour T., Dillman R.: Programming by Demonstration: Dual-Arm Manipulation Tasks for Humanoid Robots. *Proc IEEE/RSJ Intern. Conf on Intelligent Robots and systems (IROS 2004)*.

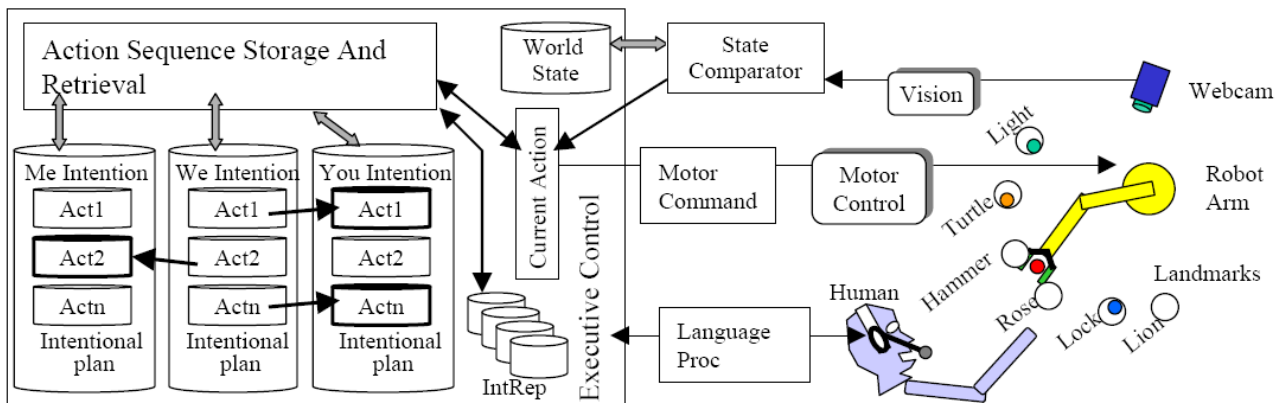


Fig 1. Cooperation System. In a shared work-space, human and robot manipulate objects (green, yellow, red and blue circles corresponding to dog, horse, pig and duck), placing them next to the fixed landmarks (light, turtle, hammer, etc.). *Action*: Spoken commands interpreted as individual words or grammatical constructions, and the command and possible arguments are extracted using grammatical constructions in Language Proc. The resulting Action (Agent, Object, Recipient) representation is the Current Action. This is converted into robot command primitives (Motor Command) and joint angles (Motor Control) for the robot. *Perception*: Vision provides object location input, allowing action to be perceived as changes in World State (State Comparator). Resulting Current Action used for action description, imitation, and cooperative action sequences. *Imitation*: The user performed action is perceived and encoded in Current Action, which is then used to control the robot under the supervision of Executive Control. *Cooperative Games*. During observations, individual actions are perceived, and attributed to the agent or the other player (Me or You). The action sequence is stored in the We Intention structure, that can then be used to separately represent self vs. other actions.

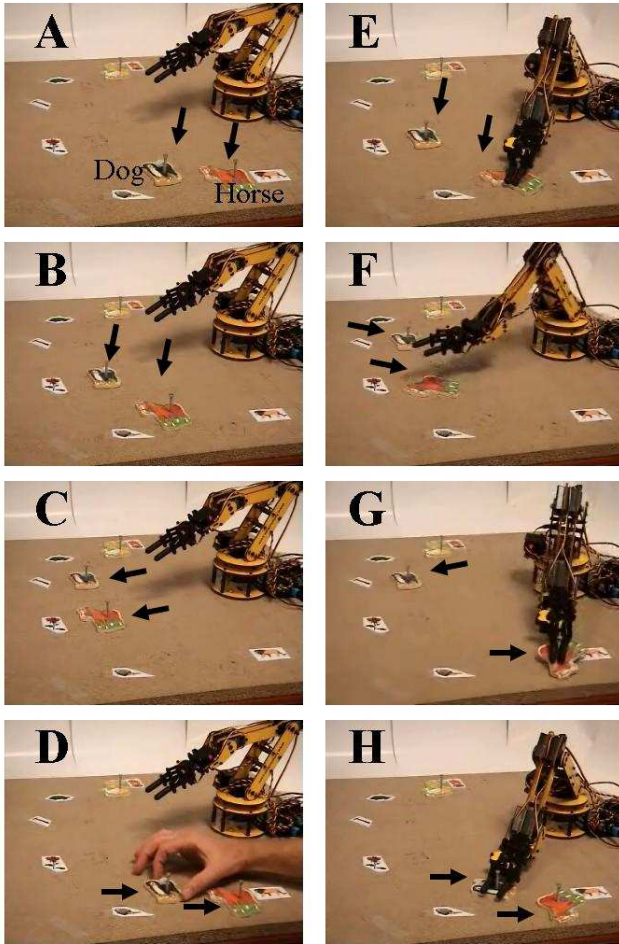


Figure 2. Cooperative task of Exp 5-6. Robot arm Cooperator, with 6 landmarks (Light, turtle, hammer, rose, lock and lion from top to bottom). Moveable objects include Dog and Horse. In A-D, human demonstrates a “horse chase the dog” game, and successively moves the Dog then Horse, indicating that in the game, the user then the robot are agents, respectively. After demonstration, human and robot “play the game”. In each of E – F user moves Dog, and robot follows with Horse. In G robot moves horse, then in H robot detects that the user is having trouble and so “helps” the user with the final move of the dog. See Exp 5 & 6.

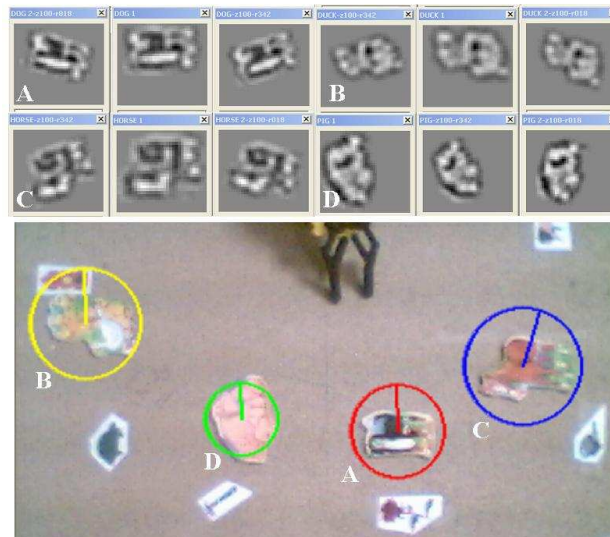


Figure 3. Vision processing. Above: A. – D. Three templates each for the Dog, Duck, Horse and Pig objects at three different orientations. Below, encompassing circles indicate template recognition for the four different objects near different fixed landmarks, as seen from the camera over the robot workspace

