Key Elements for Human Robot Joint Action

Aurélie Clodic, Elisabeth Pacherie, Rachid Alami, Raja Chatila

Key Elements for Human-Robot Joint Action

Authors Aurélie Clodic, Elisabeth Pacherie, Rachid Alami, Raja Chatila

published in :

Sociality and Normativity for Robots
Philosophical Inquiries into Human-Robot Interactions
Series Studies in the Philosophy of Sociality (Springer)
Volume 9
Published 2017
Editors Raul Hakli, Johanna Seibt
ISBN 978-3-319-53131-1 (print) | 978-3-319-85071-9 (online)

Key Elements for Human-Robot Joint Action

Aurélie Clodic, Elisabeth Pacherie, Rachid Alami, and Raja Chatila

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 and addresses issues in perception, decision, action, communication and learning, as well as their integration. At the same time, research on human-human joint action has become a topic of intense research in cognitive psychology and philosophy, providing elements and even offering architecture hints to help our understanding of human-human joint action. In this paper, we analyse some findings from these disciplines and connect them to the human-robot joint action case. This work is a first step toward the development of a framework for human-robot interaction grounded in human-human interaction.

Key words: Action, Joint action, Architecture for Social Robotics, Human Robot Interaction

Aurélie Clodic
LAAS-CNRS, Université de Toulouse, CNRS, Toulouse, France, e-mail: aurelie.clodic@laas.fr

Elisabeth Pacherie
Institut Jean Nicod, CNRS UMR 8129, Institut d'Etude de la Cognition, Ecole Normale Supérieure & PSL Research University, Paris, France e-mail: elisabeth.pacherie@ens.fr

Rachid Alami
LAAS-CNRS, Université de Toulouse, CNRS, Toulouse, France, e-mail: rachid.alami@laas.fr

Raja Chatila
Sorbonne Universités, UPMC, Univ Paris 06, UMR 7222, Institut des Systèmes Intelligents et de Robotique, F-75005, Paris, France & CNRS, UMR 7222, ISIR, F-75005, Paris, France, e-mail: raja.chatila@isir.upmc.fr

1 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, addressing perception (e.g., tactile or visual), decision (e.g., human-aware planning, supervision) and action (e.g., manipulation, navigation). At the same time, research on human-human joint action has become a topic of intense research in cognitive psychology and philosophy, providing elements and even offering control architecture hints to help our understanding of human-human joint action. We analyse some findings from these disciplines and connect them to the human-robot joint action case.

The work presented in this paper is a first necessary step toward the definition of an integrative framework needed for the design of autonomous robots that can engage in interaction with human partners. More precisely, we address the following questions:

- What knowledge does a robot need to have about the human it interacts with, and which processes does it need to handle to manage a successful interaction?
- Conversely, what information should the human possess to understand what the robot is doing and how the robot should make this information available to its human partner?

2 A simple scenario

We introduce a simple human-robot interaction scenario to illustrate the issues we address: a human and a robot have the common goal to build a stack with four blocks and to put a pyramid on the top of the stack. They are face to face. They should stack the blocks in a specific order (1, 2, 3, 4). Each agent participates to the task by placing his/its blocks on the stack. At the end, one of the agents should place a pyramid on the top of the stack. The actions available to each agent are the following (with "object" = block or pyramid): take an object on the table, put an object on the stack, remove an object from the stack, place an object on the table, give an object to the other agent, support the stack (see next).

Fig. 1 illustrates the initial state. Each agent can initially access only a subset of blocks and one of the two pyramids.

Each agent is able to perceive the state of the world and so knows where each object is, whether they can reach a given object, and can infer whether their partner can reach a given object. Moreover, we assume that each agent is able to observe the activity of the other. Fig. 2 depicts the two possible final states.

A number of deviations from a nominal course are possible. For example, the stack might collapse or an agent might drop a block on their side of the table or on the opposite side. If the block falls on the opposite side, the question arises whether the other agent should put it directly on the stack or give it to the initial agent.

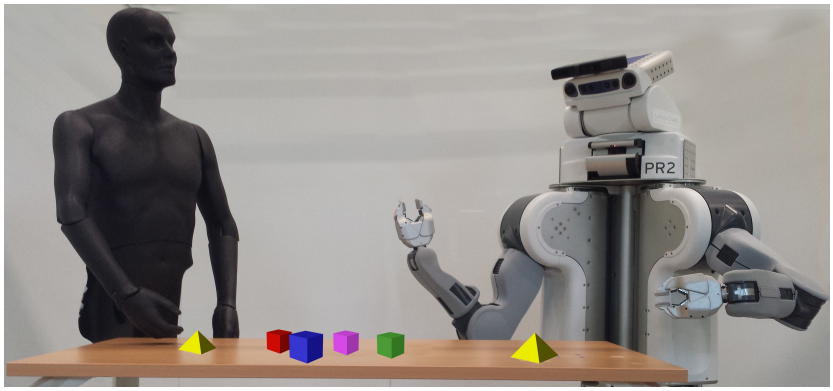

Fig. 1 Initial state.

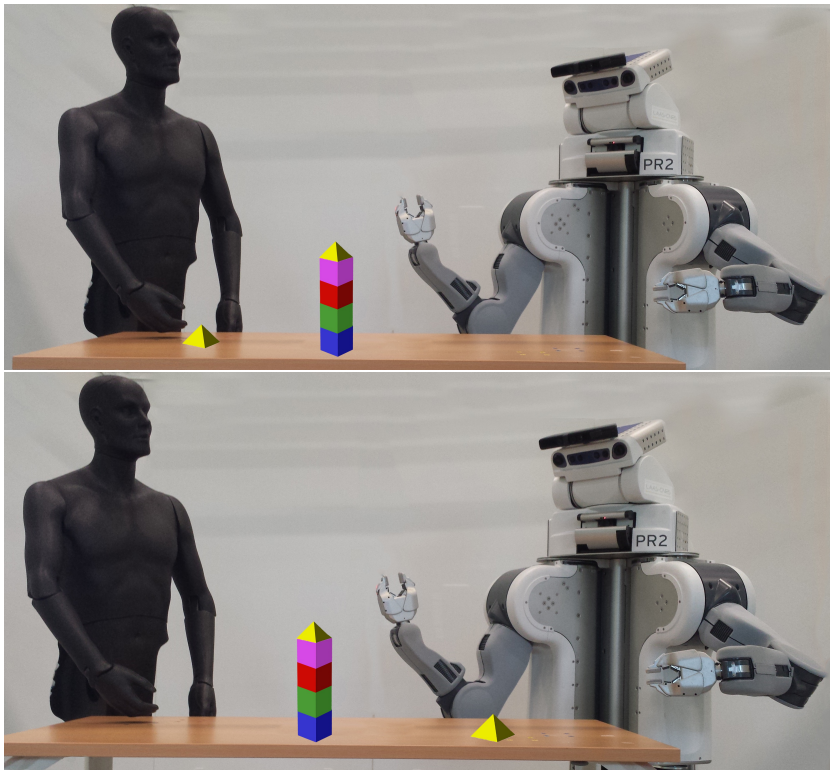

Fig. 2 Possible final states.

Moreover, during the execution of the task, different types of behaviours are possible, including proactive behaviour (one agent could help the other one by supporting the stack while the other places a block on it), passive behaviour (one agent does not act at all) or incorrect behaviour (one agent does not stack blocks in the correct order or removes a correctly placed block from the stack).

The task needs to be set up at the beginning, for example through a dedicated command sent by the human to the robot. The way this is achieved is beyond the scope of this paper.

3 Acting Autonomously

Both philosophical and robotics approaches to joint action typically build from models of individual human or autonomous agents. Interestingly, existing approaches from both areas share a number of insights regarding the architecture needed to support individual action.

According to classical philosophical accounts of individual action, behaviour qualifies as action only if it has a certain type of mental antecedent or involves certain types of psychological processes (e.g. Davidson, 1980; Mele, 1992; Searle, 1983). Typically, this mental antecedent is identified as an intention. Intentions are often characterized as plans of action the organism chooses and commits itself to in pursuit of a goal (e.g. Bratman, 1987).

According to this view, intentions are executive attitudes whose functions include terminating practical reasoning about ends, prompting practical reasoning about means and plans, helping to coordinate the agent's behaviour over time and with the behaviour of other agents, initiating and sustaining intentional action, and monitoring and guiding it until completion. Intentions thus include representations of both goals and means towards achieving these goals, that is, action plans that can range from simple representations of basic actions to complex strategies for achieving distant goals. In addition, as pointed out by Bratman (1987), action plans are subject to rationality constraints. The various elements that form the building blocks of an action plan must be mutually consistent (internal consistency). The plan as a whole should be consistent with the agent's beliefs about the world and about current reality, including her beliefs about her own capacities and skills (external consistency). Finally the plan must take into account the wider framework of activities and projects in which the agent is also involved and be coordinated with them in a more global plan (global consistency). This brief overview of references from philosophy provides working definitions of an action, an intention, a goal and a plan, elements that should be handled to enable acting.

In parallel to this work, research in Artificial Intelligence and Robotics has defined concepts for autonomous agent actions, such as in STRIPS (Fikes and Nilsson, 1971), also based on means-ends analysis. At the same time, the robotics community has addressed the problem of *robot control architectures*, with the objective of building consistent and efficient robot system structures integrating perception,

decision and action capacities, and providing for both deliberation and reactivity. Several solutions were proposed. One of them, which is commonly used today, is the three-layered architecture (Gat, 1992; Alami et al., 1998; Muscettola et al., 1998; Nesnas et al., 2003; Saridis, 1995; Tambe, 1997), which defines:

- A *decision level*, which includes the capacities for producing a plan to accomplish a task and for supervising its execution, while being at the same time reactive to events from the next level below. The coexistence of these two features, a time-consuming planning process, and a time-bounded reactive supervisory process raises the key problem of their interaction and their integration to balance deliberation and reaction at the decisional level. Basically, the supervisory component uses the planner which may include temporal reasoning as a resource when needed and feeds the next level with the sequence of actions to be executed.
- An *execution control level*, or executive, which controls and coordinates the execution of functions distributed in operational modules (next level) according to the task requirements to achieve the plan. It is at this level that context-based action refinement is performed.
- A *functional level* which includes all the basic built-in robot action and perception capacities. These functions are encapsulated into controllable communicating modules that enable the implementation of data processing and motor control loops (image processing, obstacle avoidance, motion control, etc.). In order to make this level as hardware independent as possible, and hence portable from a robot to another, it is connected with the sensors and effectors through a logical robot interface, i.e., an abstraction of these physical devices.

This architecture relies on representations of actions, goals, plans as well as robot's knowledge and skills. Building these representations remains an active research issue.

Interestingly, Pacherie (2008, 2012) proposes a dynamic model of intentions that also distinguishes three main stages in the process of action specification:

- A *distal intentions* level (D-intentions) in charge of the dynamics of decision making, temporal flexibility and high level rational guidance and monitoring of action;
- A *proximal intentions* level (P-intentions) that inherits a plan from the previous level and whose 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;
- A *motor intentions* level (M-intentions), which encodes the fine-grained details of the action (corresponding to what neuroscientists call motor representations), is responsible for the precision and smoothness of action execution, and operates at a finer time scale than either D-intentions or P-intentions.

This suggests an interesting convergence between a philosophical account of the structure and dynamics of human action and a robot control architecture dedicated to action. From this, it appears relevant to consider whether a similar convergence could be established with regard to joint action.

4 Coordination requirements in joint action

Successful joint action depends on the efficient coordination of participant agents' goals, intentions, plans, and actions. In other words, it is not enough that agents have a common goal and that then each set their own sub-goals, devise their own individual action plan and execute this plan. They must also *coordinate* their own sub-plans with those of their co-agents so as to have a coherent joint action plan and they must coordinate their actions during the *execution* phase to insure the successful completion of the joint action. For that they must monitor their partner's intentions and actions, predict their consequences and use these predictions to adjust their sub-plans, or in the execution phase, what they are doing to what their partners are doing. These processes, however, also play an important role in competitive contexts. In a fight, for instance, being able to anticipate the opponent's moves and to act accordingly is also crucial. A further requirement in the case of joint action is that co-agents share a goal and understand the combined impact of their respective intentions and actions on their joint goal and adjust them accordingly. In other words, agents should be able to align their representations of what they themselves and their partners are doing, and of how these actions together contribute to the shared goal.

As Michael and Pacherie (2015) point out, various forms of uncertainty can undermine mutual predictability, the alignment of representations and hence coordination. They include:

Motivational uncertainty: we might be unsure how convergent a potential partner's interests are with our own interests and thus unsure whether there are goals we share and can promote together. Additionally, even if we know what their current preferences are and that they match ours, we might be unsure how stable these preferences are.

Instrumental uncertainty: even assuming that we share a goal, we might be unsure what plan to follow to achieve that goal, or, if we have a plan, we might be unsure how roles should be distributed among us, or, even if the plan and the distribution of roles are settled, we might be uncertain when and where we should act.

Common ground uncertainty: we might be unsure how much of what is relevant to our deciding on a joint goal, planning for that goal and executing our plan is common ground, or mutually manifest to us. In other words, it is not sufficient, to ensure coordination, that we are actually motivated to pursue the same goals and have sufficiently similar instrumental beliefs and plans regarding how these goals should be achieved. We must also know or believe that this is the case.

These coordination constraints apply both to human-human joint action and to human-robot joint action and they can undermine both of them. However, they do not apply with the same strength. In human-human joint action, a human faces another human. The fact that they are both humans brings lots of shared background knowledge and assumptions can be made from both sides on what the other knows or not. This is far from easy to assess in the human-robot case. In this latter case, alignment processes need to be considered carefully to ensure an acceptable level of mutual predictability. 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 correct intuitions about robot capabilities or inabilities and perception abilities or weaknesses. To deal with this issue, some propose to train the human to use the robot, as we do for other technological devices (Cakmak and Takayama, 2014).

In what follows, we will first consider what resources humans can exploit in order to reduce uncertainty and achieve coordination at the level of intentions and action planning as well as at the level of action execution. To do that we will draw on recent conceptual and empirical work investigating the cognitive processes by which coordination in joint action is achieved. We will then consider human-robot joint action and the specific challenges it raises in addition to the challenges common with human-human joint action.

5 Coordination processes in human-human joint action

Successful joint action requires agents to coordinate both their intentions and their actions. There has been a great deal of work in recent years, both conceptually and empirically, investigating the cognitive processes by which uncertainty is reduced and coordination achieved. Philosophical accounts of joint action have tended to concentrate on the conceptual requirements for shared intentions and to emphasize high-level action planning prior to acting. They are thus essentially concerned with the characterization of shared distal intentions. In contrast, cognitive psychology studies of joint action have explored the perceptual, cognitive, and motor processes that enable individuals to flexibly coordinate their actions with others online. The processes they describe are thus essentially processes involved in the formation and operation of shared proximal intentions and coordinated motor intentions. Because philosophers and psychologists focus on processes of uncertainty reduction that operate at different levels of action specification, it is important to bring together their complementary perspectives to shed light on the whole range of processes involved in acting together.

Philosophical accounts of shared intentions are attempts to cash out what it takes for agents to act in a jointly intentional manner. These accounts typically agree that shared (distal) intentions are more than mere summations of individual intentions. They agree therefore that something more is needed, although they tend to disagree on what more is needed. Rather than trying to adjudicate between different accounts, we take here the plurality of accounts as evidence that shared distal intentions may take different forms and be arrived at in a variety of ways.

According to Michael Bratman's very influential account (Bratman, 2014), shared intentions are characterized by a form of mutual responsiveness of each to each in their relevant intentions and plans. Responsiveness in intention means that each will adjust his subsidiary intentions concerning means and preliminary steps to the subsidiary intentions of others in a way that keeps track of the intended end of the joint

action. It is thus essentially a matter of responsiveness in planning. Bratman describes negotiation, bargaining, shared reasoning and shared deliberation as some of the central processes through which mutual responsiveness in intentions is achieved.

Other philosophers have emphasized the essential role of joint commitments in joint actions. Thus, according to Margaret Gilbert (2009, 2014), joint commitments constitute the core of shared intentions: agents share an intention to do A if and only if they are jointly committed to intend as a body to do A. In the basic case, a joint commitment is created when each of two or more people openly expresses his personal readiness jointly with the other to commit themselves in a certain way, and it is common knowledge between them that all have expressed their readiness. According to Gilbert, these commitments have social normative force: participants in a joint activity have obligations towards each other to act in conformity with their shared intentions and correlative entitlements or rights to others so acting.

Finally, Raimo Tuomela (2007) points out that when agents act jointly as members of a group, what he calls we-mode joint action, they are often committed not just to a particular joint goal but also to a set of values, standards, beliefs, practices, social coordination conventions, pre-established scripts and routines and so on, that form the ethos of the group. The group ethos may thus serve to minimize uncertainty in joint actions.

While these philosophers have divergent views regarding the nature of the social glue that binds together the intentions of individuals in joint action (practical rationality for Bratman, the social normativity of joint commitments for Gilbert and collective acceptance of a group ethos for Tuomela), their accounts tend to be cognitively demanding: the coordination processes involved in forming and maintaining a shared intention rest on advanced representational, conceptual and communicational skills and sophisticated forms of reasoning about the complex interplay between each other's individual beliefs and intentions and the shared goal, about the mutual obligations and entitlements the shared intention generates, or about its relations to the group ethos.

In contrast to philosophical approaches, cognitive psychology studies of joint action have tended to focus not on the conceptual requirements for shared intentions but rather on the perceptual, cognitive, and motor processes that enable individuals to flexibly coordinate their actions with others online. Following Knoblich and colleagues (Knoblich et al., 2011), we can distinguish between two broad categories of online coordination processes: emergent and intentional.

In intentional coordination, agents plan their own motor actions in relation to the joint goal and also to some extent to their partners' actions. As emphasized by Knoblich et al. (2011), shared task representations play an important role in goal-directed coordination. Shared task representations do not only specify in advance what the respective tasks of each of the co-agents are, they also provide control structures that allow agents to monitor and predict what their partners are doing, thus enabling interpersonal coordination in real time. Empirical evidence shows that having shared task representations influences perceptual information processing, action monitoring, control and prediction during the ensuing interaction (Heed et al., 2010; Schuch and Tipper, 2007; Sebanz et al., 2006). Thus, for instance, people tend

to predict the sensory consequences not only of their own but also of other participants' actions (Wilson and Knoblich, 2005) and to automatically monitor their own and others' errors (van Schie et al., 2004). Furthermore, several studies have shown that actors may form shared representations of tasks quasi-automatically, even when it is more effective to ignore one another (Atmaca et al., 2008; Sebanz et al., 2005; Tsai et al., 2008).

An important complement to the co-representation of tasks and actions is the co-representation of perception. In particular, joint attention provides a basic mechanism for sharing representations of objects and events and thus for creating a perceptual common ground in joint action (Tomasello and Carpenter, 2007; Tollefsen, 2005). Joint attention can also allow agents to perform joint actions more efficiently. For instance, a study by Brennan and colleagues (Brennan et al., 2007) demonstrated that co-agents in a joint visual search task were able to distribute a common space between them by directing their attention depending on where the other was looking and that their joint search performance was thus much more efficient than their performance in an individual version of the search task.

Another type of process that may contribute to better online coordination can be captured with the term 'coordination smoother', i.e. any kind of modulation of one's movements that 'reliably has the effect of simplifying coordination' (Vesper et al., 2010, p. 2). For example, one may exaggerate one's movements or reduce variability of one's movements to make them easier for the other participant to interpret (Pezzulo, 2011). Although coordination smoothers may in some cases be produced automatically, the term may also be applied to processes, such as nods, winks and gestures, which are produced intentionally. And of course, there are a myriad other ways in which intentional alignment processes can reduce uncertainty, linguistic communication during the action being the paradigmatic case.

In emergent coordination, coordinated behaviour occurs due to perception-action couplings that make multiple individuals act in similar ways. One source of emergent coordination involves interpersonal entrainment mechanisms. For instance, people sitting in adjacent rocking chairs will tend to synchronize their rocking behaviour, even if the chairs have different natural rocking tempos (Richardson et al., 2007). The perception of common or joint affordances can also lead to emergent coordination. A joint affordance is a case where an object affords action to two people that is may not afford to each of them individually. Thus, a seesaw may afford action to two kids, but not to a single child. A third source of emergent coordination is perception-action matching, whereby observed actions are matched onto the observer's own action repertoire and can induce the same action tendencies in different agents who observe one another's actions (Jeannerod, 1999; Prinz, 1997; Rizzolatti and Sinigaglia, 2010). It is likely that such processes make partners in a joint action more similar and thus more easily predictable, and thereby facilitate mutual responsiveness in action. Importantly, however, emergent forms of coordination can operate independently of any joint plans or common knowledge, which may be altogether absent, and do not ensure by themselves that the agents' actions track a joint goal.

Humans thus have at their disposal a vast array of coordination tools and processes, ranging from advanced representational, conceptual and communicational skills and sophisticated forms of reasoning to intentional and automatic online alignment processes, that they can use to reduce motivational, instrumental and common ground uncertainty and to promote interpersonal coordination. To enable efficient joint action, these processes must work together, as there are complementary limits on what each can do.

We must now examine whether, and under what conditions, these processes could play a similar role in human-robot interactions. It is important to note that some redundancy is present in the human case, as several combinations of these processes can be used to achieve the coordination required for successful joint action. Given that humans might have different expectations regarding a robot's capacities and given that the specificities of robotic cognitive architectures compared to human cognitive architectures may induce different cost/efficiency ratios in the use of these processes, the question also arises whether these processes should be deployed in different ways in human-robot interactions.

6 A (tentative) translation of coordination processes in human-robot joint action

We've seen that joint action presupposes the sharing of information at different levels, from object representations to task, action, intention and goal representations through the use of several processes. We will analyse now how such processes can make sense in a human-robot case and what kind of capacities they presuppose either on the robot or on the human side.

A pre-requisite of these processes is self-other distinction. As raised by Pacherie (2012, p. 359), it is important that agents be able to keep apart representations of their own and of others' actions and intentions.

On the robot side, this means that it should be able to handle a representation of itself and a representation of the human it interacts with, i.e., it must maintain a "mental" model of itself and a "mental" model of the human it interacts with. It should also be capable of updating these "mental" models as the action unfolds and the representations of the agents evolve. This in turn requires perspective-taking abilities, since the representations of the agents may evolve differently depending on their respective points of view.

On the human side, we can assume that the agent is able to handle several "mental" models. However, questions can be asked: does the human need to handle representations of robots' actions and intentions, in fact does he create such representations when interacting with a robot? Do we have to make it explicit at the beginning of a human-robot interaction that a robot makes use of actions and intentions representations (and which ones) to encourage its human partner to infer them?

Equipped with self-other distinction ability, the robot and its human partner need to understand what the other perceives (or does not perceive). More precisely, they

must share knowledge about their interaction space. It is necessary that both the robot and the human identify objects to be acted upon, their location as well as location of possible obstacles. Thus they track the same objects and features of the situation and are mutually aware that they both do so. Here, *joint attention* is key because if joint attention is established, whatever information I can get, I can consider my partner would have it too if it occurs in the joint attention space. The interaction space includes what both partners perceive, but also what only one partner 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 but that the human can see it. Conversely, the robot can assume the human knows that a part of the table he can see is not visible to the robot).

This means that each agent must be equipped with situation assessment abilities that will enable them to anchor the situation of action (and this is in itself a complicated matter for the robot). Then, when acting jointly, each must ensure that they track the same objects and features of the situation as their partner. On the robot side, this means that the robot must have (necessarily partial) access to the human model of the real situation. On the human side, this means that the robot perception abilities should be readable by the human to enable him or her to draw inferences about what the robot perceives or not (noting that robot sensing abilities are not always easy to decode). Finally, both the robot and the human must be aware of that, so they both understand what are their perception capacities (and limitations).

This raises a number of questions: how can a robot know that the human it interacts with attended with him to the joint task? What are the cues that should be collected to infer joint attention? Symmetrically, how can a robot exhibit joint attention? What cues should the robot exhibit to let the human infer that joint attention is achieved? Moreover, once joint attention is achieved (or at least a given level of joint attention if we consider it is not a 0/1 question), how should it be managed during the overall course of joint action? How can we handle cooperative perception between a robot and a human and thus create perceptual common ground? Is there a need to negotiate about what should be jointly attended (or not)?

Another capacity, emphasized, among others, by Tomasello et al. (2005) as a prerequisite to joint action, is *understanding intentional action*. Each agent should be able to read its partner's actions. To understand an intentional action, an agent should be able, when observing a partner's action or course of actions, to infer their partner's intention (i.e. their goal and plan). They should be able to exploit cues exchanged and to understand what their partner is attending to in their perceptive field.

That means that the robot needs to be able to understand what the human is currently doing and to be able to predict the outcomes of the human's actions. To do so, it must be equipped with action recognition abilities (again potentially constrained to the current situation) and predictive action models enabling it to predict the outcomes of both its and the human partner's actions.

Complementarily, the human should be able to understand what the robot is currently doing and to predict the outcomes of robots' actions. To do so, he must be

able to infer what is the underlying action when observing the robot's movement and to predict its outcome.

This process could be helped by the use of coordination smoothers. We can imagine that coordination smoothers could be added to already existing movement to facilitate this understanding. A human interacting with a robot would perhaps exaggerate her/his movement amplitude or do her/his movement exactly in front of the robot's dedicated perception sensor to ensure a good perception and understanding of his move. On the other side, the development of human-aware robot motion planning, that takes into account not only safety and efficiency but also legibility and social norms at planning level, could be considered as a software instance of coordination smoothers.

Equipped with *self-other distinction*, *joint attention* and *intentional action understanding* abilities, our agents should be able to understand actions in their perceptual context but this context should be enlarged to include the task and the joint goal to get the overall picture and allow coordination. This is where *shared task representations* come on stage. Equipped with such representations, our agents would be able not only to understand what the other is doing but also to predict what he/it will do next, e.g. by the use of action-to-goal or goal-to-action predictions. These predictions would help to make the entire interaction space more foreseeable. It enables also each agent to adapt his/its behaviour by taking into account this knowledge.

If we paraphrase the definition of Knoblich et al. (2011), this means that we must equip the robot with a model of the respective tasks of each of the co-agents and also with control structures that will allow it to monitor and predict what its partners are doing. On the other side, the human must be aware of the respective tasks of each of the co-agents and how to monitor them. Doing that can be considered as putting in perspective all the processes already described. For example, joint attention allows to know that both agents track the same object, intentional action understanding allows to infer that the robot is currently moving this object to a goal position, and the existence of a shared task representation enables to interpret the action as a contribution to the common goal.

It is important to point out what it means to share information in terms of information alignment. The robot and the human need to understand, to interpret the world in the same way, their understandings/interpretations need to be aligned at some point and this is a component of the ability to share. It is crucial for enabling coordination and communication among the agents.

For example, it is not sufficient for the robot to perceive that a blue object stands at position (x, y, z); it must know what object it is in order to be able to share this information with the human. It is not sufficient for the robot to interpret an arm movement as "something moves in front of me"; rather, it must interpret it as "the human hands an object to me" so it can react accordingly to this action.

This concern needs also to be taken into account on the human side. The human should be aware of the limitations of the representational resources of the robots to avoid over-interpretations by humans of what the robot knows and understands about the scene.

This alignment issue can also be considered from a broader viewpoint. Tuomela (2007) states that the involved agents should share what he calls group ethos, Tomasello et al. (2005) speak about cultural creation/learning, and Clark (1996) about common ground. We have already considered the set of information that needs to be shared to handle a joint action, but here the spectrum is larger. It concerns the set of values, standards, beliefs, practices, social coordination conventions, pre-established scripts and routines. How is it possible to model such concepts in a robot and to what extent could we consider that they are shared by the human and the robot?

Another philosophical account concerns joint commitment. According to Gilbert (2014) a shared intention to perform a joint action essentially involves a joint commitment on the part of the co-agents, where a joint commitment creates a set of mutual obligations and entitlements for agents to perform their part in the joint endeavour. How can a robot and a human express and share this kind of engagement? How can they express their readiness to be jointly committed in the first place? How do they monitor whether or not they stay committed (or not) as the action unfolds? How is the joint commitment terminated once the common goal is achieved? How are such commitments represented and updated?

Finally, Bratman (2014) proposes that shared intentions are characterized by mutual responsiveness. Engaged in a joint action, agents need to be able to share not just representations, but also reasoning processes toward the joint goal, to be able to deliberate together, to negotiate. Such processes required high level reasoning abilities. How can a robot and a human reason together toward their joint goal? What are the reasoning abilities that need to operate? Which media could be used to enable such reasoning? How to model it? We propose in Fig. 3, a representation of knowledge and processes a robot need to handle to operate a joint action with a human.

7 A framework for joint action

As stated by Knoblich et al. (2011), philosophers generally agree that "what distinguishes joint actions from individual actions is that the joint ones involve a shared intention and shared intentions are essential for understanding coordination in joint action" (p. 60). Tomasello et al. (2005) say nothing else when they claim 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 they add: "shared intentionality refers to collaborative interactions in which participants have a shared goal (shared commitment) and coordinated action roles for pursuing that shared goal".

Successful joint action depends on the efficient coordination of participant agents' goals, intentions, plans, and actions. In other words, in joint action, it is not enough that agents control their own actions, i.e., correctly predict their effects, monitor

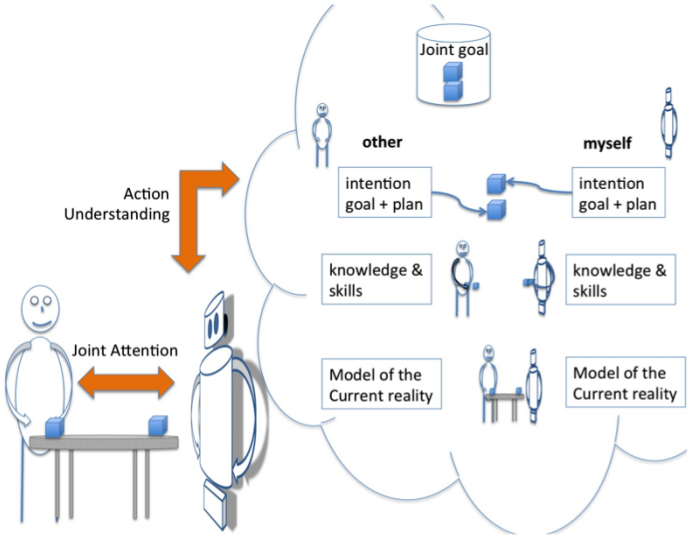

Fig. 3 Knowledge and processes for joint action. The robot builds and maintains a distinct mental model of itself and of its human partner concerning the state of the world. It also reasons and builds its own behaviour based on its estimation of its human partner intentions, knowledge and skills.

their execution and make adjustments if needed. In addition, they must also coordinate their intentions and actions with those of their co-agents so as to achieve their joint goal.

It has to be noticed that AI community has proposed seminal work on teamwork such as Cohen and Levesque (1991) and Grosz and Kraus (1996). What we propose here is to analyse Pacherie's (2007; 2011; 2012) theory of joint action, which also considers three levels of action. If we try to map this theory to robot architecture, we can describe these three levels as the following: a shared distal/decisional level, a shared proximal/execution level and a coupled motor/functional level.

7.1 Shared Distal / Decisional Level

At this level, acting alone, 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 in-

clude not only the robot's monitoring of its own actions and goals but also more generally monitoring of the joint goal and consequently monitoring of the other actions too.

Pacherie (2012) explains that at this level, the participating agents (1) represent the overall goal yet need not represent the whole plan but only their own sub-plans and the meshing parts of the sub-plans of others and (2) some of what they represent is to be performed by others. Agents need to be able to handle triadic and dyadic adjustment at that level.

7.2 Shared Proximal / Execution Level

It is at this level that situational and temporal anchoring of the action take place, which means that the action plan inherited from the distal/decision level must be further refined and adjusted to the situation at hand in order for the action to be launched and its unfolding monitored and controlled. 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.

Pacherie (2012) explains that for agents to share a proximal intention, the following should hold: (1) agents each represent their own actions and their predicted consequences in the situation at hand *(self-predictions)*, (2) agents each represent the actions, goals, motor and proximal intentions of their co-agents and their consequences *(other-predictions)*, (3) agents each represent how what they are doing affects what others are doing and vice-versa and adjust their actions accordingly *(dyadic adjustment)*, (4) 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 action plan)*, (5) agents each predict the joint effects of their own and others' actions *(joint predictions)*, and (6) 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 *(triadic adjustment)*.

That means the robot needs to be able to handle: its world representation, a world representation of the human it interacts with (potentially limited to the task to be performed), 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 means to monitor progress toward the joint goal (and possibly a means to revise the ongoing joint plan). A triadic adjustment means that the robot and the human can adapt their behaviour toward the joint goal. This implies, for example, that if the human drops his object in the robot space, the robot will place the object on the stack. If it had done a dyadic adjustment it would have made the object accessible to the human to let him finish the action. A dyadic adjustment means that the robot and the human can adapt their behaviour to the other's actions (not toward

the joint goal) Interestingly, Tomasello et al. (2005) have proposed that the capacity for triadic engagement presents two phases in the course of human development. At around 9 to 12 months of age, infants begin to interact together with a goal-directed agent toward some shared goal. In doing this, both perceptually monitor the behaviour and perceptions of their partners with respect to that shared goal. However, it is only at around 12 to 15 months of age, that they begin to engage in significant amounts of coordinated joint engagement, understanding not just the shared goal but also beginning to understand the complementarity between their own and their partner's specific action plans. As Tomasello and colleagues point out: "This means, for instance, that the child understands that in pursuing the shared goal of building a block tower the adult holds the edifice steady while she, the child, places blocks. Infants of this age not only share goals but also coordinate roles" (Tomasello et al., 2005, p. 682). This understanding thus makes possible more flexible triadic adjustment processes, such as reversing roles with a partner or helping the partner play his role. These adjustment mechanisms exploit shared task representations that not only specify in advance what the respective tasks of each of the co-agents are but also provide control structures that allow for flexible coordination.

7.3 Coupled Motor / Functional Level

This level corresponds to robot sensory-motor behaviour that would allow to achieve high-bandwidth interaction with the human partner. An example could be exchanging an object with a human and the associated force-feedback processes. In such tight situations involving precise coordination between the actors, the parameterization of the functional level needs to be coupled with the one of the other actor. This means that the robot control loops would be directly parameterized by the other actor's motions or actions.

8 Conclusion: Toward a framework for joint action

In this paper we proposed an analysis of some findings in psychology and philosophy in the domain of human-human joint action. Our aim was to identify knowledge, representations and processes that a robot, interacting with a human, needs to possess and exploit. Complementarily, we analysed what information needs to be shared with the human to enable a consistent interaction. We have seen that, as already pointed out by cognitive psychology and philosophy, self-other distinction, joint attention, intentional action understanding and shared task representations as well as common ground, joint commitment and mutual responsiveness make sense in our context. We came up with a set of questions about their management in our context.

We then tried to apply the framework proposed by Pacherie (2012) to a human-robot case and show that it could fit the development of an architecture dedicated to human-robot interaction. It is inspiring in the search to frame an architecture dedicated to human-robot interaction. We show that this three-layer division seems meaningful not only for the robot, human and the human-human cases but also for the human-robot case.

This paper is a first step toward the objective of identifying and describing precisely the different robot abilities and how they are involved in the overall process of collaborative human-robot task achievement. To this end, we placed ourselves purposefully at a conceptual level. This analysis is obviously sustained by the work of the human-robot interaction community to which we contribute. We refer the interested reader to Lemaignan et al. (2016) and Kruse et al. (2013) where we discuss a number of such abilities studied and implemented by robotics researchers.

Acknowledgements This work has been funded by the French Agence Nationale de la Recherche ROBOERGOSUM project ANR-12-CORD-0030.

References

Alami, R., Chatila, R., Fleury, S., Ghallab, M., and Ingrand, F. (1998). An architecture for autonomy. *The International Journal of Robotics Research*, 17(4):315–337.

Atmaca, S., Sebanz, N., Prinz, W., and Knoblich, G. (2008). Action co-representation: the joint snarc effect. *Social Neuroscience*, 3(3-4):410–420.

Bratman, M. (1987). *Intention, Plans, and Practical Reason*. Harvard University Press, Cambridge, MA.

Bratman, M. (2014). *Shared Agency*. Oxford University Press, Oxford.

Brennan, S. E., Chen, X., Dickinson, C., Neider, M., and Zelinsky, G. (2007). Coordinating cognition: The costs and benefits of shared gaze during collaborative search. *Cognition*, 106:1465–1477.

Cakmak, M. and Takayama, L. (2014). Teaching people how to teach robots: The effect of instructional materials and dialog design. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction*, pages 431–438. ACM.

Clark, H. H. (1996). *Using Language*. Cambridge University Press, Cambridge.

Cohen, P. R. and Levesque, H. J. (1991). Teamwork. *Nous*, 25(4):487–512.

Davidson, D. (1980). *Essays on Actions and Events*. Oxford University Press, Oxford.

Fikes, R. E. and Nilsson, N. J. (1971). STRIPS: A new approach to the application of theorem proving to problem solving. *Artificial intelligence*, 2(3-4):189–208.

Gat, E. (1992). Integrating planning and reacting in a heterogeneous asynchronous architecture for controlling real-world mobile robots. In *AAAi*, volume 1992, pages 809–815.

Gilbert, M. (2009). Shared intention and personal intentions. *Philosophical Studies*, 144:167–187.

Gilbert, M. (2014). *Joint Commitment: How We Make the Social World*. Oxford University Press, Oxford.

Grosz, B. J. and Kraus, S. (1996). Collaborative plans for complex group action. *Artificial Intelligence*, 86(2):269–357.

Heed, T., Habets, B., Sebanz, N., and Knoblich, G. (2010). Others' actions reduce crossmodal integration in peripersonal space. *Current Biology*, 20:1345–1349.

Jeannerod, M. (1999). The 25th bartlett lecture. to act or not to act: Perspectives on the representation of actions. *Quarterly Journal of Experimental Psychology*, 52A:1–29.

Knoblich, G., Butterfill, S., and Sebanz, N. (2011). Psychological research on joint action: Theory and data. *Psychology of Learning and Motivation-Advances in Research and Theory*, 54:59–101.

Kruse, T., Pandey, A. K., Alami, R., and Kirsch, A. (2013). Human-aware robot navigation: A survey. *Robotics and Autonomous Systems*, 61(12):1726–1743.

Lemaignan, S., Warnier, M., Sisbot, E. a. C. A., and R., A. (2016). Artificial cognition for social human-robot interaction: An implementation (in press). *Artificial Intelligence*.

Mele, A. R. (1992). *Springs of Action*. Oxford University Press, Oxford.

Michael, J. and Pacherie, E. (2015). On commitments and other uncertainty reduction tools in joint action. *Journal of Social Ontology*, 1(1):89–120.

Muscettola, N., Nayak, P. P., Pell, B., and Williams, B. C. (1998). Remote agent: To boldly go where no ai system has gone before. *Artificial Intelligence*, 103(1):5–47.

Nesnas, I. A., Wright, A., Bajracharya, M., Simmons, R., and Estlin, T. . (2003). CLARAty and challenges of developing interoperable robotic software. In *Intelligent Robots and Systems, 2003. (IROS 2003). Proceedings. 2003 IEEE/RSJ International Conference on*, volume 3, pages 2428–2435. IEEE.

Pacherie, E. (2007). Is collective intentionality really primitive? In Beaney, M., Penco, C., and Vignolo, M., editors, *Mental Processes: Representing and Inferring*, pages 153–175. Cambridge Scholars Press, Cambridge.

Pacherie, E. (2008). The phenomenology of action: A conceptual framework. *Cognition*, 107(1):179–217.

Pacherie, E. (2011). Framing joint action. *Review of Philosophy and Psychology*, 2(2):173–192.

Pacherie, E. (2012). The phenomenology of joint action: Self-agency vs. joint-agency. In Seemann, A., editor, *Joint Attention: New Developments*, pages 343–389. MIT Press, Cambridge.

Pezzulo, G. (2011). Shared representations as coordination tools for interaction. *Review of Philosophy and Psychology*, 2(2):303–333.

Prinz, W. (1997). Perception and action planning. *European Journal of Cognitive Psychology*, 9:129–154.

Richardson, M. J., Marsh, K. L., Isenhower, R. W., Goodman, J. R. L., and Schmidt, R. C. (2007). Rocking together: Dynamics of unintentional and intentional interpersonal coordination. *Human Movement Science*, 26:867–891.

Rizzolatti, G. and Sinigaglia, C. (2010). The functional role of the parieto-frontal mirror circuit: Interpretations and misinterpretations. *Nature Reviews Neuroscience*, 11:264–274.

Saridis, G. N. (1995). Architectures for intelligent controls. In Gupta, M. M. and Sinha, N. K., editors, *Intelligent Control Systems: Theory and Applications*, pages 127–148. IEEE Press, Piscataway, NJ.

Schuch, S. and Tipper, S. P. (2007). On observing another person's actions: Influences of observed inhibition and errors. *Perception & Psychophysics*, 69:828–837.

Searle, J. (1983). *Intentionality*. Cambridge University Press, Cambridge.

Sebanz, N., Knoblich, G., and Prinz, W. (2005). How two share a task: Corepresenting stimulus–response mappings. *Journal of Experimental Psychology: Human Perception and Performance*, 31:1234–1246.

Sebanz, N., Knoblich, G., Prinz, W., and Wascher, E. (2006). Twin peaks: An erp study of action planning and control in co-acting individuals. *Journal of Cognitive Neuroscience*, 18:859–870.

Tambe, M. (1997). Towards flexible teamwork. *Journal of Artificial Intelligence Research*, 7:83–124.

Tollefsen, D. (2005). Let's pretend: Children and joint action. *Philosophy of the Social Sciences*, 35(75):74–97.

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

Tomasello, M., Carpenter, M., Call, J., Behne, T., and Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, 28:05.

Tsai, C. C., Kuo, W. J., Hung, D. L., and Tzeng, O. J. (2008). Action corepresentation is tuned to other humans. *Journal of Cognitive Neuroscience*, 20(11):2015–2024.

Tuomela, R. (2007). *The Philosophy of Sociality*. Oxford University Press, Oxford.

van Schie, H. T., Mars, R. B., Coles, M. G., and Bekkering, H. (2004). Modulation of activity in medial frontal and motor cortices during error observation. *Nature Neuroscience*, 7(5):549–554.

Vesper, C., Butterfill, S., Knoblich, G., and Sebanz, N. (2010). A minimal architecture for joint action. *Neural Networks*, 23(8-9):998–1003.

Wilson, M. and Knoblich, G. (2005). The case for motor involvement in perceiving conspecifics. *Psychological Bulletin*, 131(3):460.