Cooperation with Robots? A Two-Dimensional Approach

Anika Fiebich*, Nhung Nguyen*, and Sarah Schwarzkopf *

*equal contribution, alphabetical order

Abstract In this treatise, we aim to characterize cooperation in human-robot interaction. Therefore we provide a two-dimensional approach to cooperation that allows (1) determining where precisely a specific phenomenon that is called ‘cooperation’ lies on the axis of a ‘behavioral dimension’ and the axis of a ‘cognitive dimension’ and (2) showing what this implies for the robustness of the cooperation. This approach not only enables scientists from different disciplines and traditions to locate themselves in the debate when investigating what they call ‘cooperation,’ it also provides a framework to spell out the cognitive preconditions that being engaged in cooperation on either dimension involves. Identifying such preconditions serves as a fruitful means to address the leading question of the present treatise. The analysis shows that robots are capable of being engaged in human-robot cooperation on either dimension. However, the implications of having a shared intention with respect to joint commitments being involved are only partly implemented in the robotic systems so far.

Keywords cooperation, shared intention, joint action, joint commitments, human-robot interaction

A. Fiebich,
Ruhr-University Bochum, Institute of Philosophie II,
Universitätsstraße 150, 44780 Bochum, Germany.
Email: anifiebich@gmail.com

N. Nguyen,
Artificial Intelligence Group, Faculty of Technology, Bielefeld University,
33594 Bielefeld, Germany.
Email: nhung.nguyen@mail.de

S. Schwarzkopf,
Center for Cognitive Science, Institut für Informatik und Gesellschaft, University of Freiburg,
Friedrichstr. 50, 79098 Freiburg, Germany.
Email: s.schwarzkopf@gmail.com

1. The Starting Point

A couple tackles a box. A group of employees forges out a plan how to avoid bankruptcy of their company. An orchestra plays Beethoven's Ninth Symphony. Phenomena that are called 'cooperation' are frequent in the scientific literature. The list is endless. The main aim of this treatise is to characterize cooperation in robots. Therefore we provide a two-dimensional approach to human-human and human-robot cooperation that allows determining where precisely a specific phenomenon that is called 'cooperation' lies on the axis of a 'behavioral dimension' and the axis of a 'cognitive dimension.' The dimensions are orthogonal to each other. Moreover, we discuss the cognitive preconditions that come along with being engaged in cooperation on these two dimensions. This methodological distinction serves as a fruitful means to analyze whether and to what extent cooperation on a 'behavioral dimension' and on a 'cognitive dimension' is implementable in robots.[1]

The term 'cooperation' is used inflationarily. That is, one and the same phenomenon may be called 'cooperation' under one definition but not under another. Consider two examples: Imagine a group of graduate students who are taught about Adam Smith's 'Hidden Hand' in a business school. The students form a pact to help humanity by way of each pursuing his or her selfish interests. To an outside observer, their behavior may not look like cooperation at a first glance. However, if the observer knows about the pact that the graduate school students have formed and their shared intention to pursue that pact, he may well be in a position to call their behavior cooperation (see Searle 1990 for a discussion of this example).

Now imagine Natalie is the organizer of a conference and arrives at the venue long before the conference starts. She aims to prepare the conference hall and hence starts to carry the many boxes that she finds there out of the hall. Steve is a participant of the conference. He also arrives far too early and wants to relax at the venue before the conference starts. Steve feels disturbed by the many boxes. Hence he starts to carry the boxes out of the hall. Imagine that Natalie observes Steve, and Steve observes Natalie carrying the boxes out of the room. They start to coordinate their behavior by way of Natalie carrying the big boxes and Steve taking the small ones. But neither of them realizes that the other one has recognized him- or herself as partaking in the action as well. Hence no mutual common knowledge about having the same goal is involved. It is not even necessary that they have knowledge about having the same goal. All that is required is that each of them is aware of his or her own goal (i.e., getting rid of the boxes in the room) and to coordinate the own behavior to that of the other in a way that is fruitful to achieve that goal (i.e. carrying the big boxes when the other is

---

[1] We refer to embodied artificial agents as robots.

carrying the small ones, and vice versa). Now imagine Ipke is sitting in the conference hall. Ipke may see that Natalie and Steve are cooperating. But are they doing so in the same manner as the graduate students?

2. A Two-Dimensional Approach to Cooperation

We suggest that the answer is: No. These examples point to two main dimensions of how to describe cooperation that are predominantly used in the contemporary scientific literature: (i) A *behavioral dimension* that focuses on the behavior patterns of the cooperating human or robotic agents; and (ii) a *cognitive dimension* that focuses on the cognitive states of the agents who are engaged in cooperative actions.[2] In the following, we exemplify these two dimensions by dipping into the philosophical and empirical literature on cooperation and discussing various examples that lie on different stages of the two dimensions. Moreover, we spell out the cognitive preconditions that being engaged on either dimension presuppose and discuss which of these preconditions are implementable in robots. We end with a summary and appeal to future research to address some questions that are still open.

2.1. The Cognitive Dimension of Cooperation

The cognitive dimension of cooperation is endorsed not only by philosophers (e.g., Searle 1990; Bratman 1993; Tuomela 2010) but also by psychologists (e.g., Carpenter 2009; Tomasello et al. 2005) who emphasize the role that a particular kind of intentionality plays for cooperation. In individual actions, i.e. actions performed by one agent, the conception of 'intentionality' refers to the individual's intention to act (see e.g., Searle 1983) whereas in collective actions, i.e. actions performed by a group of agents, the conception of 'collective intentionality' refers to the intentions of the group. Various accounts have been proposed to determine what the intentions of a group of agents are. Pragmatically, we discuss only those that we consider the dominant theoretical accounts that have been proposed, including Searle's (1990) conception of 'we-intention,' Tuomela's (2010) conception of 'intention in the we-mode,' and Bratman's (1993) conception of 'shared intention.'

---

[2] In general, a third dimension could be added that discusses the phenomenological experiences of agents whilst cooperating. For the present purposes, however, the two dimensions mentioned here are sufficient.

Searle (1990) highlights that "the notion of we-intention, of collective intentionality, implies the notion of cooperation" (p. 406; italics in original). In Searle's account, a 'we-intention' is not reducible to a set of I-intentions that is, the sum of the personal intentions of single individuals to aim for the same goal even if aiming for the same goal is mutual common knowledge among the individuals. A 'we-intention' of a single individual represents the contribution of the individual to the joint action; e.g., "we are making the sauce by means of me stirring." Searle defines a 'we-intention' as a special attitude of a single individual involving a 'we-activity.' Notably, in Searle's account, an agent may have a we-intention even in cases where there is no other agent; hence a "brain in a vat" (Searle 1990, p. 407) may have nothing but the illusion of having a body that is engaged in cooperation with another agent.

Searle's conception of 'we-intention' is similar to Tuomela's (2010) conception of 'intention in the we-mode' insofar as both philosophers emphasize a special kind of attitude as being distinctive for having such an intention. However, Tuomela not only discusses the intentional content of this intention but also the reasons for and the implications of having such an intention with respect to commitments and the specific phenomenal experiences in group activities. Tuomela (2010) distinguishes between group activities that agents perform in the 'we-mode' opposed to the 'pro-group I-mode.' For example, two drivers may happen to arrive from different directions at a tree trunk lying on the road and blocking their way. To continue their drive, they jointly remove the trunk. This is an example of a pro-group I-mode joint action. Agents who have an intention in the pro-group I-mode to be engaged in a group activity accept a shared goal of the group (e.g., to remove a trunk) due to private reasons (e.g., you want to continue your drive to be in time for your grandma's coffee party and I want to escape the police who are chasing me after I robbed the city bank). In joint actions performed in the we-mode, in contrast, agents pursue a shared goal on the basis of group reasons (maybe determined by a group ethos) and they are collectively committed to each other to pursue that goal. That is, none of the agents is in a position to stop pursuing the goal without the agreement of the other agent(s). Tuomela's explorations of joint commitments are compatible with and (partly) draw on Gilbert's (2009) account (see Sect. 3 for a discussion of Gilbert's account).

Bratman (1993; 1992) accounts for what he calls a 'shared intention' not as an attitude or intention in the mind of a single agent (such as proposed by Searle, see above) but as a "state of affairs that consists primarily in attitudes (none of which are themselves the shared intention) of the participants and interrelations between those attitudes" (p. 107-8). That is, at least two individuals are required for there to be a shared intention. Mutual common knowledge of aiming for the same goal[3], intending to participate in a cooperative action to pursue that goal in accordance with and because of the other's intention, being willed to compromise and to mesh individual sub-plans are distinctive features of how the individual attitudes of the participants are interrelated in the case of a shared intention. Imagine two people having the shared intention of painting a house together: One has the sub-plan of painting it red whereas the other has the sub-plan of painting it blue. If both of them do not compromise and mesh the individual sub-plans, they won't succeed in pursuing their shared intention. Notably, Bratman (2009) also emphasizes that the intentions of the single individuals involve a we-activity in the Searlean sense; "a Searlean we-intention is, then, a candidate for the intentions of individual participants that together help to constitute a shared intention, though Searle himself does not say how the we-intentions of different participants need to be inter-related for there to be a shared intention" (p. 41). Having a shared intention is, by Bratman's (1992) account, essential for 'shared cooperative activities' that are characterized by mutual responsiveness, commitment to the joint activity and commitment to mutual support of the agents.

For the present purposes, we make use of a hybrid account of the intention of a group. That is, we postulate that such an intention needs to involve a we-activity (as proposed by Searle); is characterized by the joint commitment of the group members to pursue the intention until its end (as emphasized by Gilbert and Tuomela); and presupposes mutual common knowledge of aiming for the same goal, intending to participate in a cooperative action to pursue that goal in accordance with and because of the other's intention as well as being willed to compromise and to mesh individual sub-plans (as highlighted by Bratman). Finally, we follow Bratman in considering at least two agents necessary for there to be a shared intention since we strive for an analogy to the behavioral dimension of cooperation that presupposes the participation of at least two agents (see below).

In our account, being engaged in cooperation on the cognitive dimension presupposes minimally that

(a_i) two (or more) agents perform actions to pursue the same goal, and

[3] One remark needs to be made with respect to this definition. We deliberatively do not adopt Bratman's expression of 'mutual common knowledge' but rather prefer to use a cognitively less demanding definition to stress that mutual common knowledge of aiming for the same goal neither requires an infinite number of overlapping embedded mental states nor do these mental states need to be linguistic in content (Wilby 2010) and hence can be involved in basic cooperative phenomena such as joint attention (Fiebich and Gallagher 2013) that are already present in preverbal infants (Tollefsen 2005).

(a_{ii}) the agents know that they have the same goal.

Cooperation on the cognitive dimension occurs in various degrees. For example, when the agents have mutual common knowledge of having the same goal (i.e., each agent knows about having the same goal like the other agent and also knows that the other agent knows about that), the phenomenon lies on a higher point on the axis than if the agents only have knowledge themselves of having the same goal like the other agent. Pursuing the same goal on the basis of having a shared intention marks the endpoint (i.e. the highest stage) on the cognitive dimension of cooperation; we call this point 'cognitive cooperation supreme.' These are the necessary criteria that are together sufficient for a phenomenon to be 'cognitive cooperation supreme':

(ai) Two (or more) agents perform actions to pursue the same goal, and

(aii') the agents have a shared intention, i.e., a common goal involving we-intentions.

In general, a phenomenon may be described as cooperation on the cognitive dimension to a more moderate degree, ranging from the agents having knowledge of having the same goal over having mutual common knowledge of having the same goal (i.e., having a 'common goal') to pursuing the same goal on the basis of having a shared intention.

2.2. The Behavioral Dimension of Cooperation

Cooperation described on the behavioral dimension is typically not explicitly spelled out but implicitly often underlies the experimental paradigms that investigate various aspects of cooperative behavior in social interactions. In our account, 'cooperation' on the behavioral dimension is characterized by the coordinated behavior of the interacting agents. The more complex the coordinated behavior is to an outside observer, the higher is the degree of the described phenomenon on the continuum of the behavioral dimension.

Sebanz et al. (2006) define joint action as ''any form of social interaction whereby two or more individuals coordinate their actions in space and time to bring about a change in the environment'' (p. 70; italics added). By this definition, the goal that both agents have is to change the environment. A variety of cognitive mechanisms (such as joint attention, action observation, task-sharing and action coordination) may play a role in the success of joint actions, however, no shared intention or mutual common knowledge of having the same goal need be involved. This view is implicitly shared among a number of psychologists (e.g., Isenhower et al. 2010; Chaminade et al. 2012) as well as computer scientists (e.g., Sofge et al. 2005; Lenz et al. 2008).

In psychology, for example, studies with physically coupled dyads (Reed et al. 2006; Reed et al. 2006) interacting independently of each other with only haptic feedback suggest that the participants develop cooperative strategies without knowing that they are cooperating. Here the term cooperation is used without presupposing that the agents have a shared intention to cooperate. Likewise, in other studies, participants are asked to play sequential turn-taking games (McCabe et al. 2001; Decety et al. 2004) in which they have only few possibilities to interact. In Decety et al.'s (2004) study, for example, participants play a game with the experimenter. Both strive towards the goal of the game, i.e. the creation of a specific pattern on a computer screen, in alternating turns. In most of these board-game-like operationalizations of cooperation, the participant interacts with a confederate or with a computer script.

Of course, cooperation that is described on the behavioral dimension may also involve specific cognitive factors. For example, in other studies, mutual common knowledge of having the same goal is present among the participants and cooperation is defined as the coordination of actions to achieve a common goal. Chaminade et al. (2012) manipulate the 'amount of cooperation' by a changing number of individual action possibilities in a joint motor task. Shockley et al. (2003) use a similar principle: Each participant needs to coordinate his or her behavior to that of the other participant by helping each other to achieve a particular goal. Here, as in those studies in which two real participants interact with each other and do so simultaneously (Isenhower et al. 2010; Newman-Norlund et al. 2008), mutual common knowledge of having the same goal is involved. However, it is unclear whether a shared intention is involved or the participant's interaction with the other person is nothing but using the other person as a 'social tool' (Moll and Tomasello 2007), i.e., a means to an end. Also, the motivation to cooperate by promising extra money for accurate and quick joint performance (e.g., Böckler et al. 2011) or by simulating an in-group and letting team members work together against an out-group (e.g., Vonk 1998) does not control for whether participants are engaged in a 'joint action in a we-mode' on the basis of a group ethos or whether they just participate for their own benefit (i.e., a 'joint action in a pro-group I-mode,' see Sect. 2.1). In Isenhower et al.'s (2010) study, participants are not explicitly asked to cooperate. In the beginning they move wooden planks independently of each other. Then, the participants start to carry the planks together at the latest when they encounter difficulties carrying a plank alone. This situation may, but does not need to, evoke the formation of a shared intention.

In our view, Sebanz et al.'s (2006) notion of joint action serves as a fruitful means to capture the three necessary criteria that together are sufficient for there to be 'cooperation' on the behavioral dimension, which are:

(A_1) Two (or more) agents coordinate their behavior in space and time,

(A_2) which is observable from the outside,

(A_3) to bring about a change in the environment.

Some remarks need to be added with respect to these criteria: in (A3), 'to bring about a change in the environment' amounts to bring about a change in the physical or in a virtual environment by either achieving a particular end state (e.g., to locate the table together in front of the TV), or an end product (e.g., to build a tower together). Notably, (A1) does not necessarily require that the agents coordinate their behavior in a real-interactive setting. As the conference example illustrated (see Sect. 1), one agent A may coordinate his or her behavior to the behavior of the other agent B with a time delay. To coordinate his or her behavior to that of B it is not even required that A perceives B's behavior. It is sufficient that A coordinates his or her behavior to what A detects to be the result of B's behavior. For example, Steve does not need to observe Natalie carrying all the big boxes out of the conference hall. It is sufficient that Steve realizes that Natalie has carried the big boxes so that he takes the small ones. For there to be cooperation, however, Steve needs to perceive the disappeared big boxes in the conference hall as the result of Natalie's action.

Moreover, rather than referring to particular cognitive (e.g., motivational) states that guide the behavior of the agents, the term 'to' in 'to bring about a change in the environment' in (A_3) refers to the efficiency of the more or less complex coordinated behaviors of two agents as a means to an end that is detected by an outside observer (independent of whether or not the goal recognized by the observer is the same goal as pursued by the agents). Such complexity can be analyzed solely on the behavioral dimension without the need to refer to cognitive notions such as shared goal (i.e., having mutual common knowledge of aiming for the same goal) or shared intention. Complexity is defined by the amount of observable coordinative action sequences within the cooperation. For example, randomly taking any box out of the room is less complex than taking the boxes out of the room in a specific order. Finally, we emphasized that the coordinated behavior of the agents needs to be observable from the outside (A_2) since we draw here on a conception of cooperation as it is used in experimental paradigms.

We propose that there is no end point of the continuum of the behavioral dimension of cooperation analog to the end point of the cognitive dimension of cooperation. There is no 'behavioral cooperation supreme.' Rather, the axis of the behavioral dimension has an open end which is characterized by highly complex coordinated behavior of the agents to achieve a particular goal as it is recognized from the outside.

This requires the agents to have the same goal with neither knowledge about having the same goal nor a shared intention being necessarily involved.

2.3. Two Dimensions

As illustrated in Fig. 1, ‘cognitive cooperation supreme’ is the end of the axis of the cognitive dimension. At this stage, two agents perform particular actions to pursue a shared intention. In the business school example, the graduate students behave in a selfish manner to pursue their shared intention of helping humanity. This is a case of ‘cognitive cooperation supreme.’ Moreover, the coordinated behavior patterns of the graduate students are not at all observable from the outside. Hence the phenomenon lies on a zero point of the behavioral axis. We propose calling such a phenomenon that lies on the zero point of the behavioral axis and not on the zero point of the cognitive axis ‘pure cognitive cooperation.’

In contrast, given that Steve and Natalie pursue their goal of emptying the room (see example in Sect. 1) by carrying the boxes out of the room, their behavior is coordinated in order to bring about a change in the environment but no knowledge of having the same goal has to be involved. Hence the phenomenon lies on the zero point of the cognitive axis. We propose calling a phenomenon that lies on the zero point of the cognitive axis and not on the zero point of the behavioral axis ‘pure behavioral cooperation.’

Of course, a variety of options exist in between the extreme cases of 'pure cognitive cooperation' and 'pure behavioral cooperation.' Two agents may pursue a shared intention in a highly coordinated fashion. Furthermore, the behavior of the agents may be more or less coordinately structured. That is, coordination can be regarded as a matter of degree. The cognitive states of the agents can also be regarded as a matter of degree, ranging from having a shared intention to having mutual common knowledge of having the same goal to having knowledge of having the same goal (see Fig. 1).

A phenomenon that lies in the middle of both axes could be a modified version of the conference example in which Natalie and Steve have mutual common knowledge of having the same goal of emptying the conference hall by carrying the boxes out of the room but without having a shared intention and hence without joint commitments to pursue that goal until its end. Moreover they may coordinate their behavior in a moderately complex way as it is observable from the outside. Moderate complex coordination applies to behavior in which agents are not consistently mutually responsive over time whereas little complex coordination means that the agents coordinate their behavior only occasionally.

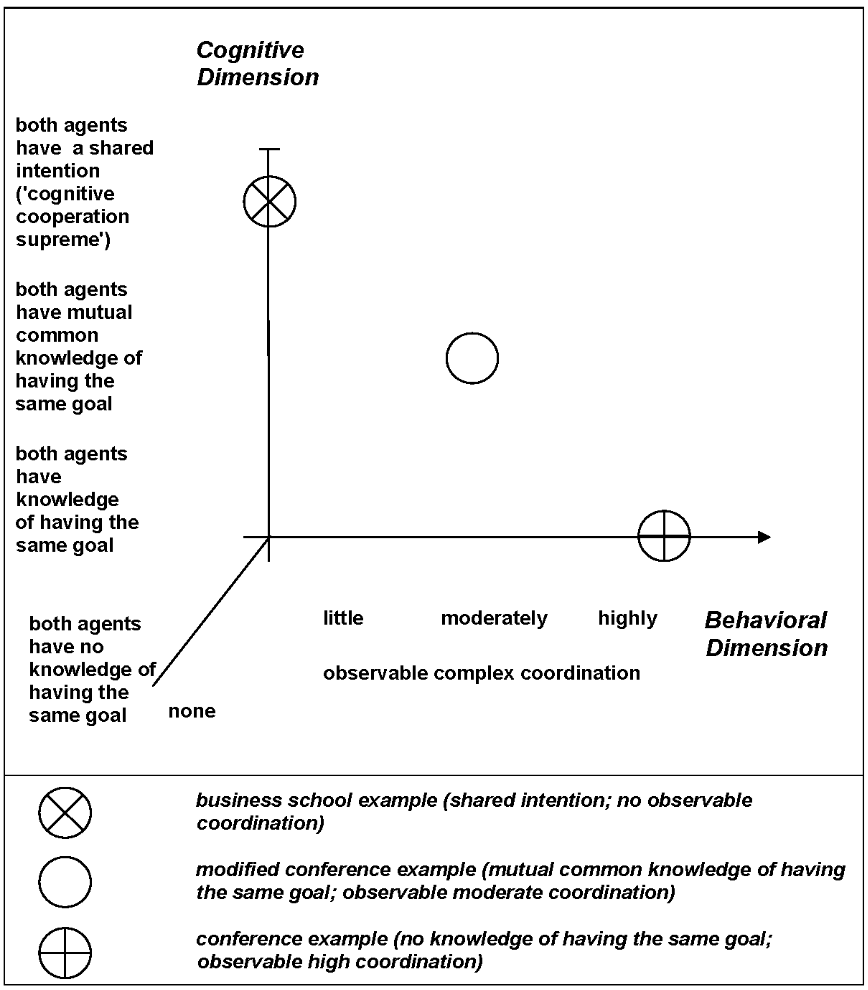

Fig. 1 Two Dimensions of Cooperation

These two dimensions shall serve as a fruitful means for scientists not only to locate themselves in the debate but also to facilitate the dialogue on cooperation among different disciplines and traditions. Moreover, the two-dimensional approach to cooperation proposed here has implications concerning the robustness of a particular cooperative phenomenon (see Sect. 3), as well as for determining whether and to what extent cooperation on either dimension is implementable in artificial systems such as robots (see Sect. 4).

3. Implications of the Two-Dimensional Approach for the Robustness of Cooperation

Cooperation is robust when it is resistant against various confounding factors that may inhibit the agents from pursuing their goal until its end. The role of the robustness of cooperation has been discussed by various philosophers with respect to commitments; all are in agreement that commitments play a central role for cooperation. Bratman (1992), for example, emphasizes that commitment to the joint activity (i.e., pursuing the joint activity until its end) as well as commitment to mutual support (i.e., helping each other in situations that demand for that) are central features of what he calls a 'shared cooperative activity.' He also adds mutual responsiveness to the list. That is, he proposes that each agent is obliged to be responsive to the other's actions by coordinating his or her own behavior to the actions of the other; "each seeks to guide his behavior with an eye to the behavior of the other, knowing that the other seeks to do likewise" (p. 328). As indicated above, observable coordinated behavior patterns are only necessarily involved when describing cooperation on the axis of the behavioral dimension but not of the cognitive dimension. Moreover, 'pure behavioral cooperation' requires observable coordinated behaviors of the agents (to various degrees) but does not presuppose knowledge about aiming for the same goal.

Following a number of philosophers (Bratman 1992; Gilbert 2009; Tuomela 2010), we propose that joint commitments are only involved when the cooperating agents have a shared intention.[4] That is, joint commitments are only involved in what we have called 'cognitive cooperation supreme.' In Gilbert's (2009) account, agents who

[4] Contrary to Bratman (1992), Gilbert proposes that (2009) "people share an intention when and only when they are jointly committed to intend as a body to do such-and-such in the future" (p. 167). We keep neutral to this. Hence we do not discuss Gilbert's 'disjunction criterion' according to which the shared intention of a plural subject to pursue a goal continues to exist even when the agents who constitute the plural subject no longer have personal contributory intentions. Whether or not cooperating agents act as a 'plural subject' or not is not relevant for the present purposes.

have a shared intention are obliged to pursue that intention due to the following criteria: (i) the *concurrence criterion* according to which "absent special background understandings, the concurrence of all parties is required in order that a given shared intention be changed or rescinded, or that a given party be released from participating in it"; and (ii) the *obligation criterion* according to which "each party to a shared intention is obligated to each to act as appropriate to the shared intention in conjunction with the rest" (p.175).

The concurrence criterion ensures that the agents are not in the position to rescind from the joint commitment of pursuing a shared goal unless the concurrence of all parties is given. To illustrate this recall the conference example. Natalie and Steve have the same goal to empty the room without mutual common knowledge being involved. They structure their carrying of the boxes out of the room in a coordinated fashion. But if Steve were to change his mind spontaneously and stop carrying the boxes because he prefers to have a coffee with Ipke, Natalie would not be in a position to complain. She would also not be entitled to complain if having the same goal of carrying the boxes out of the room is mutual common knowledge between them. However, if Natalie and Steve have the shared intention of carrying the boxes out of the room in order to prepare the venue for the conference and Steve were to change his mind under these conditions, Natalie would be in a position to complain. As the obligation criterion states, Steve would 'owe' Natalie the future action of carrying the boxes. Of course, Natalie may kindly allow Steve to go for a coffee with Ipke if he asks. But according to the concurrence criterion, concurrence of Natalie and Steve is required to release him from the joint commitment to carry the boxes. These two criteria spelled out by Gilbert (2009) refer to what Bratman (1992) has called 'commitment to joint activity.' Bratman emphasizes that 'commitment to mutual support' plays a crucial role in shared cooperative activities. For example, if somebody shuts the door in Natalie's face when she wants to go through, Steve is obliged to open the door for her since Natalie is carrying a box and hence unable to do so.

All these examples illustrate that 'cognitive cooperation supreme' is more robust than any cooperation that is described on a lower stage of the cognitive dimension since the joint commitment to the joint activity and the joint commitment to mutual support that naturally comes along with having a shared intention enhances the chances that the agents' goal will be pursued till its end. In all other cooperative phenomena, the agents are only privately committed to achieve their individual goal, which is the same goal as that of the other agent. Hence, no obligation to pursue the goal until its end is involved.

Joint commitments are involved in long-term joint actions as well as in short-term joint actions. However, they are more likely to be called into action to reduce the risk of the shared intention failing in long-term joint actions; the longer the joint action lasts, the more likely it is that a disturbance will emerge (e.g., an agent is not able to contribute his or her part to the joint action in a given situation and requires help from the cooperator).

Joint commitments may also render meshing sub-plans redundant. As pointed out by Bratman (1993), when agents pursue a shared intention, meshing individual sub-plans may be required. For example, when Natalie and Steve have the shared intention to carry the boxes out of the conference hall, Natalie may have the sub-plan of carrying the bigger boxes while Steve carries the smaller ones. A problem arises if Steve has the same sub-plan and is only willed to carry the big boxes. If Natalie and Steve find no agreement, their shared intention runs the risk of failing. However, if Natalie and Steve have the joint commitment from the beginning that one of them has to carry the big boxes, no meshing of sub-plans is required anymore.

'Cognitive cooperation supreme' may be even more robust when the shared intention the agents have involves a joint commitment to pursue particular rules and regularities to achieve the goal. Of course, the very same effect may be yielded when the agents feel only privately committed to the very same rules and regularities. Notably, however, rules and regularities do not need to make a cooperation more robust. Agreed upon rules and regularities may reduce the cognitive workload insofar as (some) decision making processes on how to pursue the shared intention lapse. In some cases, however, where the rules are very complex and novel to the agents, rules and regularities may make the achievement of the shared goal more difficult.

Notably, we do not propose that joint commitments are the only factor that determines the robustness of a cooperative phenomenon. Joint commitments ensure that agents pursue the shared intention until its end. Highly complex coordinated behavior, in turn, may (but does not need to) enhance the success of the cooperation to lead to the aimed goal. Moreover, rules and regularities that may make cooperation more robust could also be involved in a cooperative phenomenon that is described as 'pure behavioral cooperation.' Indeed, we can imagine a situation in which two agents do not know they have the same goal but in which each agent coordinates his or her behavior to that of the other agent by pursuing the very same rules as the other agent, and vice versa, which makes the cooperation more robust due to the efficiency of those rules for the given goal – but again without either agent knowing the other is pursuing these very rules.

4. Applications to Human-Robot Cooperation

There exists a wide range of intelligent agents[5] (Goodrich and Schultz 2007; Fong et al. 2003), such as robots, that are capable of interacting with humans. In order to build *cooperative robots*, technical models are typically designed on the basis of empirical findings and theoretical considerations from philosophy or psychology. Our aim is to discuss whether and if so, to what extent cooperation described on the behavioral and cognitive dimension that we postulated in Sect. 2 is implementable in robots. In order to address this question, we now spell out the cognitive preconditions that come along with cooperation on either dimension to determine which preconditions robots satisfy to be engaged in cooperation on various stages of the axes of the cognitive or behavioral dimension. Therefore we translate the cognitive preconditions into technical preconditions, which describe the abilities that robots need to possess in order to satisfy the criteria of our two dimensions of cooperation. We conclude with a discussion of the implications of human-robot cooperation compared to human-human cooperation with respect to joint commitments and rules and regularities.

4.1. Preconditions of Human-Robot Cooperation

We start by discussing the cognitive preconditions of the agents in cooperation on the behavioral dimension in which minimally (A_1) two (or more) agents coordinate their behavior in space and time, (A_2) which is observable from the outside, (A_3) to bring about a change in the environment.

There are robotic systems that satisfy the minimal criteria of cooperation on the behavioral dimension. In order to coordinate their behavior, each agent needs to perceive the (result of the) other agent's behavior, and to react to the other's behavior in a way that is fruitful to achieve the own goal (which is the same goal as that of the other agent, without necessarily knowledge of this fact being involved). Therefore the agent has to represent a goal and an action plan of how to achieve it efficiently.

Lenz et al. (2008) and Grigore et al. (2012) presented work on human-robot coordination of planning and acting within a joint physical task. In these studies, Sebanz et al.'s (2006) notion of joint action was explicitly adopted. The cognitive

[5] We refer to the definition of intelligent agents as introduced by Wooldridge and Jennings (1995). An intelligent agent in their definition has one of the following capabilities to satisfy predefined objectives (p. 116): reactivity (i.e. perceiving the environment, responding to changes), proactiveness (i.e. taking initiative in goal-directed behavior) and social ability (i.e. interacting with other agents, e.g. humans).

system, in this case the robot, is able to represent its own goal and an action plan. Moreover, the robot can perceive the movements of the other agent in order to coordinate its own actions to those of the other. The work of Lenz et al. (2008) involved a joint construction task with toy blocks between a robot and a human. With the camera sensor system, which acts as artificial 'eyes,' the robot is able to detect the bodily movements of the human partner. By pattern recognition methods, the robot is able to match the movement patterns with a database of predefined body actions, e.g., hand-grasping an object. The work of Grigore et al. (2012) included a physical task of preparing a drink and handing it over to a human. In both cases, the robot is able to perceive and to recognize the actions of the human partner as well as to react on them.

In more complex behavioral cooperation, the agents coordinate their behavior to bring about a change in the environment in a highly complex way. This requires a large amount of attention and reactivity towards the other's movements. For example, the system of Gray et al. (2005) recognizes the human's action of pushing a button by detecting the human's body motion through hardware sensors attached to the human body. The robot uses this information to track the human's success and to evaluate how to best help the human. These studies show that robots are capable of cooperation on a behavioral dimension.

Now we go further to specify the preconditions that are needed to describe cooperation on the cognitive dimension. Let us consider the minimal criteria: (a_i) two (or more) agents perform actions to pursue the same goal, and (a_{ii}) the agents know that they have the same goal.

This requires not only representing the own goal but also the goal of the other agent as well as recognizing that these goals are one and the same. Knowledge of having the same goal is modeled by establishing a goal for the robot (e.g., grasping an object, handing it over to a human partner), detecting the human's actions (e.g., that the human opens his or her hand and approaches object), recognizing the partner's goal (e.g., the partner is ready to receive the object), and matching the robot's goal with that of the human (i.e., both goals are congruent). Once the same goal is established, the robot starts to pursue its goal (i.e. grasping the object, handing it over to human partner). In the work of Foster and Matheson (2008) the robot JAST is capable of representing a joint assembly goal[6] that is to be achieved by the robotic system and the human. The human interaction with JAST is based around jointly assembling wooden construction toys. JAST fulfills the minimal criteria of a cooperation on the cognitive dimension.

Recall that in a 'cognitive cooperation supreme,' the second criterion is that ($a_{ii'}$) the agents have a shared intention, i.e., a common goal involving we-intentions. The

[6] A goal within an assembling task (e.g. a building task with toy blocks) that is to be reached by the involved partners.

presupposition of having a common goal is having mutual common knowledge of having the same goal, which makes it a cognitive precondition of 'cognitive cooperation supreme.' Another cognitive precondition is representing the goal involving we-intentions. These two preconditions are required for having a shared intention.

Mutual common knowledge of having the same goal can be very basic, e.g., achieving joint attention together (Fiebich and Gallagher 2013). Joint attention is implementable in robots. One example of signaling joint attention to the human partner was presented by Pfeiffer-Lessmann and Wachsmuth (2009). In this study, the robot is equipped with the ability to establish joint attention using gaze. The robot detects when both the human and the robot are looking at the same object.

The crucial question is whether robots are capable of having a shared intention or whether cooperation under a cognitive description is system-specific and dedicated only to humans. Having a shared intention goes beyond having mutual common knowledge of having the same goal: It also involves we-intentions of both agents. The challenge is to translate a we-intention into technical terms. Following Searle's (1990) conception of 'we-intention,' Hoffman and Breazeal (2004) made an attempt to model a we-intention in the robot Leonardo. In their technical setup, Leonardo performs the task of pressing three buttons. The task is represented by the agent and divided into subtasks with specific actions that are to be performed. These actions are only performed when the predefined conditions of the action are satisfied. All actions are performed until the end of the task, i.e. until the conditions of the goal (in this case of switching all buttons on) are met. Leonardo is able to engage in performing this task jointly with a human and is able to have a common goal involving what the authors have modeled as we-intention. The idea is that in cooperation, a task is not only performed by Leonardo himself but together with a human partner. Recall that according to Searle (1990), the content of a 'we-intention' represents the agent's contribution to the joint action. The task should be divided between the robot and the human. That is, Leonardo needs to take the human's actions into account, when he decides what to do next. For each action that is to be performed in order to reach the common goal, Leonardo negotiates who is taking the turn for the next action. If he is able to do it, he will suggest taking the turn by using his hand and showing a pointing gesture to himself. He then waits for the human's reaction. The human can acknowledge or refuse the turn-suggestion. The task is then assigned as being performed by Leonardo himself or the human partner. Leonardo represents the common goal not only as his individual goal, but also as the goal of the partner. If an action is assigned to the human, Leonardo monitors and evaluates the execution of the action with respect to the common goal as if he were performing the action himself. This is understood as Leonardo's commitment to the partner's action.

According to Bratman (1993), a 'shared intention' involves intending to participate in a cooperative action to pursue a common goal in accordance with and because of the other's intention. In the technical system underlying Leonardo's behavior, this is realized during the whole interaction. Leonardo continually tracks a joint plan (i.e. the currently executed task) which is to be performed in order to reach that common goal. He derives his own intentions based on the common plan. He also derives the human partner's actions and changes made in the world that apply to the sub-plans.

In a more recent study, Dominey and Warneken (2011) created a similar implementation of a 'we-intention' in a robotic system. In this study, the robot stored a particular action sequence that the experimenter demonstrated. This action sequence included different agents (i.e., the robot and the experimenter) for different actions in a cooperative game (e.g., 'chasing the dog with the horse'). The robot was not only capable of storing this 'we-intention' but also of switching flexibly between the roles of the agents in the action sequence, dependent on whether the experimenter made the first move or not.

In sum, Hoffman and Braezael's (2004) as well as Dominey and Warneken's (2011) studies can be regarded as successful attempts of implementing a shared intention and therefore a 'cognitive cooperation supreme' in an artificial system. Pfeiffer-Lessmann (2011) presented similar work that will be discussed below in more detail. Whether the implications of having a shared intention are also addressed is an open question.

4.2. Implications of Human-Robot Cooperation

In Sect. 3, we have discussed two factors that may make cooperation more robust: rules and joint commitments. We will now discuss to what degree these two factors have already been successfully implemented in robots.

For a robot it is crucial to determine appropriate coordinated behavior patterns in order to achieve a common goal. Thus the main question is how to achieve these patterns. One method of acquiring knowledge about how to structure and select coordinated behavior patterns is by first collecting empirical data from human-human interaction where people cooperate to solve tasks in a specific domain. Characteristic features of human behavior are identified and adopted for implementation in the robotic systems. In such systems, these empirical data are used to model rules, in order to coordinate and plan interaction structures that resemble human cooperative behavior in task-solving (Allen et al. 2001). That means empirical data helps engineers to program rules in advance which enable robots to exhibit what seem to be appropriate behavioral patterns in cooperation. The goal of this research area is to build robots that act in a way that, from a human's point of view, is judged to be like human behavior. The more rules are predefined, the more human-like the behavior might appear. But situations may occur that were not considered in the programming. The technical consequences of having cooperation with rules and regularities that are *not* defined in advance are that unpredicted and new situations might occur when the human partner does not react in the presupposed way. In dynamically changing environments, rule-based systems are prone to errors, as there are no predefined rules for dealing with the new situation. Then the robot chooses the wrong action to respond or is not even able to continue the interaction at all. That means the less rules and regularities are defined in advance, the more ambiguities are there to choose the appropriate (re-)actions – and the more difficult it is for artificial systems to engage in cooperation. Therefore, the more rules are defined in advance, the more robust the human-robot cooperation. In order for robots to deal with situations in which rules are not explicitly defined, they need sophisticated learning mechanisms. But these mechanisms are still deficient and currently a huge amount of engineering work (Goodfellow et al. 2013) is still to be done to reach human abilities.

The other implication concerns joint commitments. It remains an open question whether Hoffman and Breazeal (2004) implemented having a joint commitment in the robot. Leonardo is privately committed to the joint goal of pressing all the buttons because he continues to pursue the goal once the other agent signals him to join the task. This communicative signal could be regarded as an indicator of a joint commitment between the agents. However, Hoffman and Breazeal (2004) do not discuss Leonardo's reaction when the human partner suddenly stops contributing to the joint action task. Does he complain, or try to reengage the partner? This case does not seem to be implemented in the robotic system Leonardo. In Dominey and Warneken's (2011) study, in contrast, if the experimenter does not take over his action part himself and denies the robot's question "Is it OK?" after a 15 second delay, the robot starts to help the experimenter. Pfeiffer-Lessmann (2011, 235-236), in turn, presented cMax, a robot that is able to deal with a situation in which a human partner stops reacting during an ongoing cooperation. In the presented example cMax tries to give a conversational turn to the human partner, but the human does not react and turns away. cMax is able to recognize that the human is not taking the current turn according to discourse rules. In this case cMax expresses another giving-turn gesture with his hand in order to provoke a reaction from the human. If the human still does not react, cMax makes another attempt with a more articulated action. When the human does not react, cMax then assumes that the human wants to stop the interaction. cMax takes over and closes the conversation by saying: "Let's leave it at that." cMax builds expectations during the cooperation according to contact management rules. As these rules were broken by the human, cMax's simulated emotional state is affected[7]. cMax becomes more and more annoyed. Since joint commitments play a crucial role for the robustness of cooperation, future research in computer science should put more effort in implementing full-fledged joint commitments in robots.

5. General Discussion

In this treatise, we proposed a two-dimensional approach to cooperation according to which a phenomenon can be described as lying on the axes of two dimensions; an axis of a behavioral dimension that is defined by the complexity of coordinated behaviors of the agents as observable from the outside, and an axis of a cognitive dimension that is defined by the cognitive states of the agents. This approach not only enables scientists from different disciplines and traditions to locate themselves in the debate when investigating what they call 'cooperation,' it also provides a framework to spell out the cognitive preconditions that being engaged in cooperation on either dimension involves. Identifying such preconditions served as a fruitful means to address the leading question of the present treatise, which is to determine whether and to what extent behavioral and cognitive cooperation are implementable in robots.

We would like to draw scientists' attention to the methodological distinction of a behavioral and a cognitive dimension of cooperation. The object of investigation differs depending on the dimension used to characterize cooperation. Whereas the latter dimension is defined by the cognitive states of the agents, the former is defined by the observability of the agents' behavior patterns as being cooperative or not. This differentiation enables cognitive scientists and behavioral researchers to discuss research phenomena in a more specific way. Moreover, the distinction between a behavioral and a cognitive dimension builds a fruitful framework for accurate operationalizations in empirical studies and for defining exact phenomena in order to implement them in robots. While cognitive psychological accounts as well as philosophy of mind accounts are more interested in the cognitive dimension, e.g., there are two paradigms of research in computer science: The first one aims at simulating human-like cognitive processes in robots interacting with humans. This again helps further understanding of the functioning of the human mind. Computer scientists in this field are more interested in the cognitive dimension of cooperation. The second paradigm of research in computer science tries to model robots that are perceived as acting human-like. This method intends to increase the human's willingness to cooperate with the robot in the same way as with a human cooperation partner.

[7] The underlying simulated emotional system was presented by Becker et al. (2004).

Computer scientists working in this field are more interested in the behavioral dimension of cooperation.

The analysis has shown that robots are capable of being engaged in human-robot cooperation on either dimension. However, the implications of a shared intention with respect to joint commitments being involved are only partly implemented in the robotic systems that we have discussed above. Recall that when one cooperator does not fulfill his or her part in the planned action sequence, the other cooperator is in the position to rebuke (Gilbert 2009, see Sect. 3). This has been successfully implemented in the robot cMax by Pfeiffer-Lessmann (2011). Recall also that joint commitments involve a commitment to mutual support (Bratman 1993). That is, if one cooperator is not capable of performing his or her action part, the other is obliged to help out. Dominey and Warneken (2011) have accounted for this case. They have modeled a robotic system that helps the human partner by performing his or her part in a situation in which he or she does not do anything when it is his or her turn and denies the question 'Is it OK?.' Note, however, that the robot is taking over the whole action of the partner. That is, Dominey and Warneken (2011) do not account for the case that part of the action has already been performed by the experimenter and the robot has to *supplement* the action to its end. Accounting for such a case is a challenge for future research. Moreover, future research may account for the robot's sensitivity to whether the human partner is not able or not willed to perform his or her action part in situations where this is not explicitly stated. In addition, future research may address the question whether and to what extent knowing that one is engaged in cooperation with a robot rather than with another human has an impact on the experienced feeling or awareness of being a fellow of this group of agents (so-called 'fellow-feeling') and how this, in turn, may influence other factors such as e.g., standing to the joint commitment or willingness to help. By combining a robot's ability to cooperate with an artificial emotion system, Pfeiffer-Lessmann's work on the robot cMax may be a starting point for future research on the implementation of a fellow-feeling in robots.

Acknowledgements We would like to thank Olle Blomberg, Cédric Paternotte, and Nadine Pfeiffer-Lessmann for helpful comments on a previous version of this treatise. Moreover, we thank Stephanie Nicole Schwenke for proofreading this piece in grammar and spelling. This research has been supported by the VW-project "The Social Mind".

References

Allen, James F., Byron, Donna K, Dzikovska, Myroslava, Ferguson, George, Galescu, Lucian, and Amanda Stent. 2001. Towards Conversational Human-Computer Interaction. *AI Magazine,* 22(4): 27-37.

Becker, Christian, Kopp, Stefan, and Ipkes Wachsmuth. 2004. Simulating the emotion dynamics of a multimodal conversational agent. In *Proceedings of affective dialogue systems: Tutorial and research workshop (ads 2004)*. 154-165. Berlin Heidelberg: Springer.

Böckler, Anne, Knoblich, Günther, and Natalie Sebanz. 2011. Giving a helping hand: effects of joint attention on mental rotation of body parts, *Experimental Brain Research,* 211: 531–545.

Bratman, Michael E. 1992. Shared Cooperative Activity. *The Philosophical Review* 101(2): 327-341.

Bratman, Michael E. 1993. Shared Intention. *Ethics,* 104(1): 97-113.

Bratman, Michael E. 2009. Shared Agency. In *Philosophy of the Social Sciences: Philosophical Theory and Scientific Practice*, ed. C. Mantzavinos, 41-59. Cambridge: Cambridge University Press.

Carpenter, Malinda 2009. Just how joint is joint action in infancy? *Topics in Cognitive Science* 1: 380-392.

Chaminade, Thierry, Marchant, Jennifer L., Kilner, James, and Christopher D. Frith. 2012. An fMRI study of joint action – varying levels of cooperation correlates with activity in control networks, *Frontiers in Human Neuroscience 6*: 1-11 (Art. 179).

Decety, Jean, Jackson, Philip L., Sommerville, Jessica A., Chaminade, Thierry, and Andrew N. Meltzoff. 2004. The neural bases of cooperation and competition: an fMRI investigation, *NeuroImage* 23: 744– 751.

Dominey, Peter, and Felix Warneken. 2011. The basis of shared intentions in human and robot cognition, *New Ideas in Psychology* 29: 260-274.

Fiebich, Anika, and Shaun Gallagher. 2013. Joint attention in joint action. *Philosophical Psychology*, 26(4): 571-587.

Fong, Terrence, Nourbakhsh, Illah, and Kerstin Dautenhahn. 2003. A Survey of Socially Interactive Robots, *Robotics and Autonomous Systems* 42(3-4): 143-166.

Foster, Mary E., and Colin Matheson. 2008. Following Assembly Plans in Cooperative, Task-Based Human-Robot Dialogue, In *Proceedings of the 12th Workshop on the Semantics and Pragmatics of Dialogue (Londial 2008)*.

Gilbert, Margaret 2009. Shared intention and personal intentions, *Philosophical Studies* 144: 167-187.

Goodfellow, Ian J., Erhan, Dumitru, Carrier, Pierre L., Courville, Aaron, Mirza, Mehdi, Hamner, Ben, and Yoshua Bengio. 2013. Challenges in Representation Learning: A report on three machine learning contests. *In Neural Information Processing*, ed. M. Lee, A. Hirose, Z.-G. Hou, R.M. Kil, 117-124. Springer Berlin Heidelberg.

Goodrich, Michael A., and Alan C. Schultz. 2007. Human-robot interaction: a survey. *Found. Trends Hum.-Comput. Interact.* 1(3) *(January 2007)*: 203-275.

Gray, Jesse, Breazeal, Cynthia, Berlin, Matt, Brooks, Andrew, and Jeff Lieberman. 2005. Action parsing and goal inference using self as simulator, *Robot and Human Interactive Communication, ROMAN 2005:* 202-209.

Grigore, Elena C., Eder, Kerstin, Pipe, Anthony G., Melhuish, Chris, and Ute Leonards. 2012. Joint action understanding improves robot-to-human object handover. *IROS 2013*: 4622-4629.

Hoffman, Guy, and Cynthia Breazeal. 2004. Collaboration in Human-Robot Teams. In *Proceedings of AIAA First Intelligent Systems Technical Conference, Chicago, IL. (AIAA)*: 1-18.

Isenhower, Robert W., Richardson, Michael J., Carello, Claudia, Baron, Reuben M., and Kerry L. Marsh. 2010. Affording cooperation: Embodied constraints, dynamics, and action-scaled invariance in joint lifting, *Psychonomic Bulletin & Review* 17(3): 342-347.

Lenz, Claus, Nair, Suraj, Rickert, Markus, Knoll, Alois, Rösel, Wolfgang, Gast, Jürgen, Bannat, Alexander, and Frank Wallhoff. 2008. Jointaction for humans and industrial robots for assembly tasks. In *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication*, 130-135.

McCabe, Kevin, Houser, Daniel, Ryan, Lee, Smith, Vernon, and Theodore Trouard. 2001. A functional imaging study of cooperation in two-person reciprocal exchange, *PNAS* 98 (20): 11832-11835.

Moll, Henrike, and Michael Tomasello. 2007. Cooperation and human cognition: the Vygotskian intelligence hypothesis, *Philosophical Transactions of the Royal Society B*: 1-10.

Newman-Norlund, Roger D., Bosga, Jurjen, Meulenbroek, Ruud G.J., and Harold Beckering. 2008. Anatomical substrates of cooperative joint-action in a continuous motor task: Virtual lifting and balancing, *NeuroImage* 41: 169–177.

Pfeiffer-Lessmann, Nadine 2011. *Kognitive Modellierung von Kooperationsfähigkeiten für einen künstlichen Agenten.* Bielefeld Universitätsbibliothek.

Pfeiffer-Lessmann, Nadine, and Ipke Wachsmuth. 2009. Formalizing joint attention in cooperative interaction with a virtual human, In B. Mertsching, M. Hund, & Z. Aziz (Eds.), *KI 2009: Advances in Artificial Intelligence*, 540-547. Berlin: Springer (LNAI 5803).

Reed, Kyle, Peshkin, Michael, Hartmann, Mitra J., Grabowecky, Marcia, Patton, James, and Peter M. Vishton. 2006. Haptically Linked Dyads. Are Two Motor-Control Systems Better Than One? *Psychological Science,* 17(5): 365-366.

Reed, Kyle B., Peshkin, Michael, Hartmann, Mitra J., Patton, James, Vishton, Peter M., and Marcia Grabowecky. 2006. Haptic cooperation between people, and between people and machines, Proceedings of the 2006 IEEE/RSJ (International Conference on Intelligent Robots and Systems), October 2006, Bejing (China): 2109-2114.

Searle, John 1983. *Intentionality. An Essay in the Philosophy of Mind*. Cambridge: University Press.

Searle, John 1990. Collective Intentions and Actions. In Cohen, P., Morgan, J. and Pollack, M. (Eds.), *Intentions in Communication*, 401-415. Cambridge, MA: MIT Press.

Sebanz, Natalie, Bekkering, Harold, and Günther Knoblich. 2006. Joint action: bodies and minds moving together. *Trends in Cognitive Sciences,* 10(2): 70-76.

Shockley, Kevin, Santana, Marie, and Carol A. Fowler. 2003. Mutual Interpersonal Postural Constraints Are Involved in Cooperative Conversation, *Journal of Experimental Psychology: Human Perception and Performance,* 29(2): 326-332.

Sofge, Donald, Bugajska, Magdalena D., Trafton, J. Gregory, Perzanowski, Dennis, Thomas, Scott, Skubic, Marjorie, Blisard, Samuel, Cassimatis, Nicholas L., Brock, Derek P., Adams, William, and Alan C. Schultz. 2005. Collaborating with Humanoid Robots in Space. *I. J. Humanoid Robotics 01/2005;* 2: 181-201.

Tollefsen, Deborah 2005. Let's Pretend! Joint Action and Young Children, *Philosophy of the Social Sciences,* 35: 75-97.

Tomasello, Michael, Carpenter, Malinda, Call, Josep, Behne, Tanja, and Henrike Moll. 2005. Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences,* 28: 675-735.

Tuomela, Raimo 2010. *The Philosophy of Sociality*. Oxford: University Press.

Vonk, Roos 1998. Effects of Cooperative and Competitive Outcome Dependency on Attention and Impression Preferences, *Journal of Experimental Social Psychology* 34: 265–288.

Wilby, Michael 2010. The simplicity of mutual knowledge. *Philosophical Explorations,* 13(2): 83-100.

Wooldridge, Michael, and Nicholas R. Jennings. 1995. Intelligent Agents: Theory and Practice, *Knowledge Engineering Review* 10(2): 115-152.