Rev.Phil.Psych. (2011) 2:303–333
DOI 10.1007/s13164-011-0060-5



# Shared Representations as Coordination Tools for Interaction

**Giovanni Pezzulo**

Published online: 29 May 2011


**Abstract** Why is interaction so simple? This article presents a theory of interaction based on the use of shared representations as "coordination tools" (e.g., roundabouts that facilitate coordination of drivers). By aligning their representations (intentionally or unintentionally), interacting agents help one another to solve interaction problems in that they remain predictable, and offer cues for action selection and goal monitoring. We illustrate how this strategy works in a joint task (building together a tower of bricks) and discuss its requirements from a computational viewpoint.

## 1 Introduction

Consider two agents that play a "tower game", consisting of building towers with "bricks" of different colors. How do they coordinate their actions without previous agreements or conventions? How do they achieve their goals, such as building a tower of red or blue bricks? They can adopt different strategies: individualistic, social-aware, or interactive.

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreements 231453 (HUMANOBS) and 270108 (Goal-Leaders).

G. Pezzulo
Istituto di Linguistica Computazionale "Antonio Zampolli", CNR, Via Giuseppe Moruzzi, 1, 56124 Pisa, Italy

G. Pezzulo (✉)
Istituto di Scienze e Tecnologie della Cognizione, CNR, Via S. Martino della Battaglia, 44, 00185 Roma, Italy
e-mail: giovanni.pezzulo@istc.cnr.it

### 1.1 Individualistic Strategy

First, each agent can simply perform the tower building task individually, irrespective of the actions of the other agent. A drawback of this strategy is that the actions of the two agents will often interfere with one another. For instance, if the two agents try to put a brick at approximately the same time, or to pick up the same brick, they will hinder one another, causing the interaction to fail. This situation is illustrated schematically in Fig. 1.

Looking at real-word interactions more closely, however, reveals that their situated nature offers some advantages. In interactive settings, actions performed by one agent produce cues that other agents can use, even when this is not intentional; for instance, an agent can refrain from picking up a brick if it sees the other agent's hand in its target position. In this case, the perceptual cue (i.e., the observed hand) enhances cooperation even if the performer agent did not intend to signal anything, and even if the observer agent does not recognize the agentive nature of the hand movements. In this sense, adapting to the actions of other individuals is similar to adapting to the external environment and its dynamics. A second example of how joint action automatically produces cues for the others is moving together a table. As studied experimentally by van der Wel et al. (2010), when an agent is physically coupled to another, such as for instance when they are moving together a table, the movements done by one agent are also implicit "signals" in that they are (haptically) sensed by the other agent and vice versa; thus, both agents can use them as cues to adjust their individual actions. (Note that humans use many other cues, such as facial expressions and direction of gaze.)

It has been suggested that this implicit form of cueing could be sufficient to ensure successful coordination, confining mental state inference to the cases when interaction fails and a corrective action has to be planned (for this argument in linguistic domains, see Horton and Keysar 1996). However, this proposal disregards the fact that humans can adopt proactive strategies to ensure that interaction *will be* successful, rather than only adjusting postdictively their actions when interaction fails, and that they can use many strategies for coordination rather than only the cues offered by another's actions. Nevertheless, this proposal has the merit to emphasize cueing and signaling phenomena,

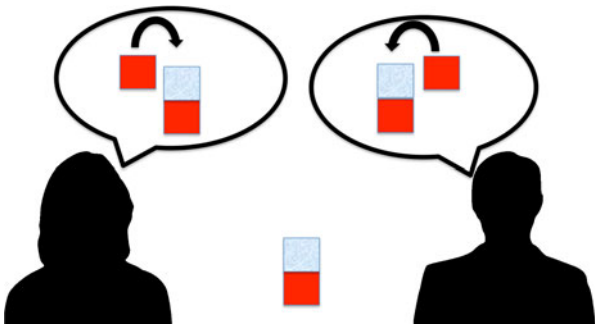

**Fig. 1** Individualistic strategy. See main text for explanation

which (we will argue) despite being insufficient as complete accounts of joint action, play nevertheless an important role in it.

### 1.2 Social-aware Strategy

An alternative, social-aware strategy consists in taking into consideration the mental states and actions of the other agent into one's own plan. Mental state inference can be done at different levels: at the level of current actions and their immediate goals (e.g., grasping a brick), at the level of distal intentions (e.g., building a red tower), and at the level of the underlying beliefs (e.g., if I build a red tower I will get an ice cream; to build a red block it is necessary to pile red bricks). As an example of the first kind, if an agent predicts that the other agent will try to put a brick at a given moment, it can wait until its action is completed, and then put its own brick. This strategy represents a first form of coordination, and can give rise to emergent phenomena such as turn-taking. In joint action scenarios that require multiple steps to be completed (e.g., the tower game), it can be advantageous to infer another's distal goals and subgoals, too. By doing so, an agent can help (or at least not hinder) another agent's goals, and ultimately maximize the success of the joint task. Finally, inferring beliefs can be advantageous when it becomes relevant to know the presuppositions for action of the other agent.

The social-aware strategy is potentially more efficacious than the individualistic one, in that it takes into consideration the effects of one's own and another's actions, and then it can select efficacious long-term strategies rather than just adapting to the environmental state, as in the case of the individualistic strategy. At the same time, the social-aware strategy requires predicting the behavior and inferring the mental states of the other agent. What follows is a brief description of how these processes might be implemented in the human brain.

Recent research in computational motor control has elucidated the requirements of successful action prediction, emphasizing the role of internal forward models. A popular view is that, in order to execute non-routine, goal-directed actions, it is necessary to derive and use an internal (predictive or forward) model of the observed dynamics rather than simply learn to react to external cues, since prediction entails better adaptivity than reaction (Wolpert et al. 1995). Similarly, internal models that encode the dynamics of actions performed by others are required to predict them so as to achieve good coordination.

However, it is worth noting that the actions of other individuals are far less predictable than environmental dynamics on the basis of purely perceptual inputs. The study of the large portion of the human and primate brain specialized for recognizing and predicting social actions (as distinct from other physical phenomena) indicates that actions are not only recognized at the level of their kinematic and dynamic regularities and trajectories, but also at the *agentive* level of action goals. A specialized neural machinery exists in humans and other primates for achieving 'parity' of performer and observer at the level

of their motor representations, so as that the observed actions are directly mapped into the observer's motor actions and their associated goals (Rizzolatti and Craighero 2004). Theoretical and empirical considerations suggest that this could be achieved by reusing the same internal models that they typically use during action execution so as to *emulate* the observed actions (Grush 2004). In other words, by reenacting their motor programs and associated internal models, humans can map the observed actions into their own motor repertoire and representations. During interaction, this gives advantages in terms of action and movement prediction (Wilson and Knoblich 2005), and at the same time this facilitates the recognition (and imitation) of the action goal (Cuijpers et al. 2006; Wolpert et al. 2003). In turn, recognition of the intended action goals, both short-term (e.g., grasping a brick) and long-term (e.g., grasping the brick as part of a stacking action), provides an additional advantage in terms of prediction efficacy and, as a consequence, it facilitates coordination.

In cases where recognition of the actions of another agent is not sufficient for predicting its behavior (and coordinating with it), it is possible to perform a deeper inferential process, and to estimate it's hidden (i.e., not directly observable) cognitive states: beliefs and intentions. Estimation of another's cognitive states, as distinct from the recognition of a perceived action and its goal, is often referred to as *mindreading*. The deeper mindreading is, the more it facilitates decoding and predicting actions performed by others, including their future actions and actions that cannot be (fully) observed. It is currently disputed if mindreading is implemented on top of the same mechanisms of internal modeling for action recognition as described before, or if complementary mechanisms of *rational inference* (Frith and Frith 2006; Gergely and Csibra 2003) are in play.

The social-aware strategy does not only use mindreading for action prediction and intention recognition. Indeed, with its actions, each agent also influences the (present and future) behavior of the other agent, being it willing or not.[1] Since its own actions influence the continuation of the interaction, a rational cognitive agent should always take into consideration their effects on the future actions and cognitive states of the other agent; this process is called *recipient design*. Recipient design permits to act strategically: to perform actions that are intentionally aimed at *influencing* the other agent, its behavior and cognitive states (i.e., its beliefs and intentions), such as for instance helping or hindering it, convincing it, or informing it. At the same time, since the effects of performed actions (e.g., requesting something, placing a brick in a certain position) on another agent are context-dependent, recipient design requires sophisticated mental state inference abilities.

To summarize, contrary to the individualistic strategy which emphasizes reaction to perceptual cues, the socially-aware strategy is essentially predictive,

[1] Influences can be direct, (e.g., moving the arm of the other agent, putting a block in a position that prevents other blocks to be added on), or indirect, so as to change another's cognitive representations (e.g., asking one to stop, looking repeatedly to its actions to express disapproval or to require help, signaling which brick to take if one is dubious).

and this entails numerous advantages. At the same time, it requires that people maintain separate representations of their own and their co-actor's actions, and use specialized predictive mechanisms (forward models) to predict both, recursively. To understand how this form of *recursive mindreading* works, consider again the tower building task. Before deciding which brick to take, an agent has to mind-read the co-actor so as to recognize its intentions (including communicative intentions) for the sake of discovering possible opportunities or conflicts with its own plans. In addition, because the agent's actions influence the future intentions and actions of the co-actors, and vice versa, the mutual dependencies have to be predicted recursively. As schematically illustrated in Fig. 2, an actor has to infer how its action will influence the co-actor's choice of the next action, then how the co-actor's next action will influence its own choice at the successive step, and so on indefinitely. As illustrated in Fig. 2, this requires using separate models of self and the co-actor.

This kind of recursive mental state inference is often assumed in game-theoretic accounts of reasoning in the social domain (e.g., in the iterated prisoner dilemma). Furthermore, a cognitive account of social reasoning has been proposed that is based on (bounded) recursive mindreading (Yoshida et al. 2008). However, despite monkeys and humans are equipped with sophisticated action prediction and mindreading abilities, which make them able to perform mental state inference during natural interactions (see e.g., Brown-Schmidt et al. 2008), the kind of recursive mindreading required by the socially-aware strategy is extremely demanding from a computational viewpoint, so it remains to be demonstrated that it is applicable to real-world situations.

A possible alternative (or add-on) to recursive mindreading, at least in cooperative scenarios, is making heavy use of explicit communication, so as to communicate one's own actions and intentions rather than letting the other

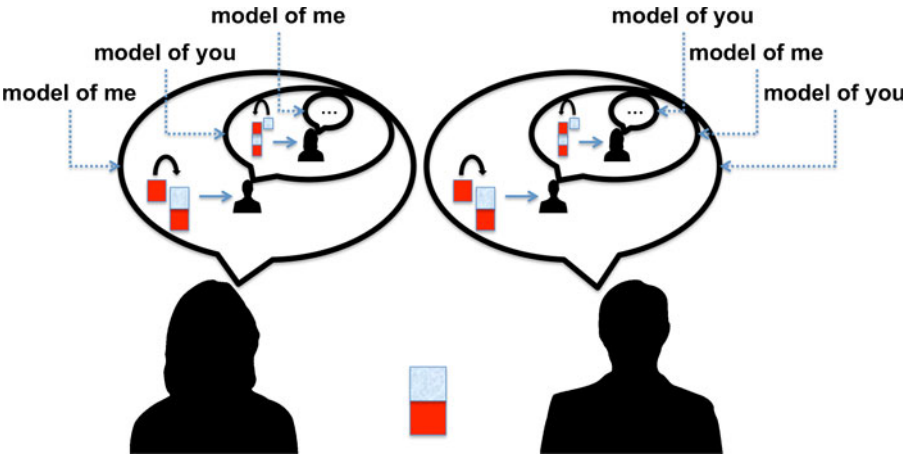


**Fig. 2** Social-aware strategy. See main text for explanation

agent infer it. For instance, research in the coordination of AI systems has reused constructs from discourse theory to model communicative dynamics during interaction (Grosz and Sidner 1990). Although communication is certainly an important part of interaction, viewing communication from the viewpoint of the social-aware strategy has two main problems: first, in the general case a common ontology is necessary from the beginning of the interaction; second, this method leads to an overestimation of how much should be really communicated and shared. The cumbersome method of performing a series of requests and responses until a full consensus is reached is the one actually followed by most artificial systems, such as for instance operating systems and web applications, which achieve coordination by exchanging a series of messages (with a predefined ontology), as well as by some dialogue systems, usually with silly effects.

Given these premises, one could ask why human (and animal) interaction is apparently so simple, and at the same time whether or not humans really adopt the social-aware strategy. One possibility is that do not even use internal modeling or mindreading, but simply learn rewarding routines, similar to model-free methods in computational reinforcement learning. Although potentially efficacious in highly standardized situations, this explanation falls short explaining the versatility of human behavior during interaction. A second explanation is the use of social conventions, such as traffic rules, or social "scripts", which constitute a mutually acknowledged set of rules whose compliance helps coordination. Again, although these methods play a role in many human interactions, they are less likely to be used in the unconstrained and novel interactive situations that people face every day.

In the rest of the article, we will argue in favor of a different strategy, the *interactive strategy*, which is social in its nature but does not use recursive mindreading. Rather than using priorly established conventions, this strategy forms a common ground of shared representations interactively, and uses it for facilitating coordination.

### 1.3 Interactive Strategy

A limitation of the individualistic strategy is that it disregards the actions and mental states of other agents, and only considers their (observable) movements as cues. In turn, the socially-aware strategy puts too much effort into inferring the mental processes of other agents, disregarding considerations of interactivity. Here we propose a third, *interactive strategy*, which considers joint action as a cooperative activity, in which cognitive agents help one another to solve interaction problems, rather than just considering one's own contribution to the joint action in isolation.

The interactive strategy simplifies the demands of interaction by implementing the maxim that agents should *remain predictable* by others. As discussed above, a key requirement for successful interaction is being able to predict accurately the actions of the other agent. In turn this could require (recursive) mindreading because prediction is too difficult in social contexts. The inter-

active strategy holds the promise that, if all agents maintain their behavior predictable, it makes mindreading less necessary.

A first illustration of this strategy is offered by the fact that, humans engaged in joint actions adapt their actions so that they are more easily discriminated, understood and predicted by others. In this vein, Vesper et al. (2010) discuss how people perform signaling actions and modify their behavior, such as for instance their action kinematics (e.g., slowing down action performance), for the sake of facilitating another's perceptual processing. Similar to what happens in the individualistic strategy, observer agents use cues produced by performer agents for facilitating prediction, except that in this case cues can be intentionally produced for this sake.

Signaling is not the only way people remain predictable. Indeed, within the interactive strategy, signaling is part of a broader strategy aimed at the formation and use of *shared representations (SR)* (Sebanz et al. 2006). As we will discuss, shared representations facilitate interactions by providing a firm ground to select what action to take (i.e., a predictable action), and as a source of evidence for predicting actions performed by others for the remaining of the interaction (especially if it is multi-step, such as for instance when two agents build together a "tower").

Shared representations formed during interaction can be used as *coordination tools* which lower cognitive load. To understand how this is possible, consider the case of a real-world coordination tool, a roundabout. Drivers arriving at a roundabout with a given goal (e.g., taking first or second exit) can coordinate without considering the mental states of the other drivers. What a driver needs to do is to select a course of actions that achieves its goals and does not conflict with the actions of the other driver, and to do so (in most cases), it is sufficient to predict the behavior of the neighbor drivers, and only for a limited period of time. All this works well because roundabouts facilitate action selection and prediction. First, a driver can use the perceptual characteristics of the roundabout itself (e.g., turning in circle) and the movements of the other cars as cues for its action selection.[2] Second, the mere presence of a roundabout makes actions of other drivers easier to predict and understand, because the courses of actions of the other drivers are subject to the same constraints as one's own, and then prediction is facilitated by the same perceptual cues regulating action selection. Should mental state inference become necessary, the number and placement of exits and cars greatly constrains the number of goals that can be attributed to another driver (which in turn constrains the predictive processes implied in action observation). Third, the same constraints that regulate action planning and prediction can be used for error monitoring, as surprising events often signal that one of the drivers is not doing the right thing. This latter process is important for taking corrective actions and avoiding collisions.

[2]Here we are disregarding conventional aspects of this activity, such as traffic rules (e.g., turning clockwise or counterclockwise, giving priority to the right of left).

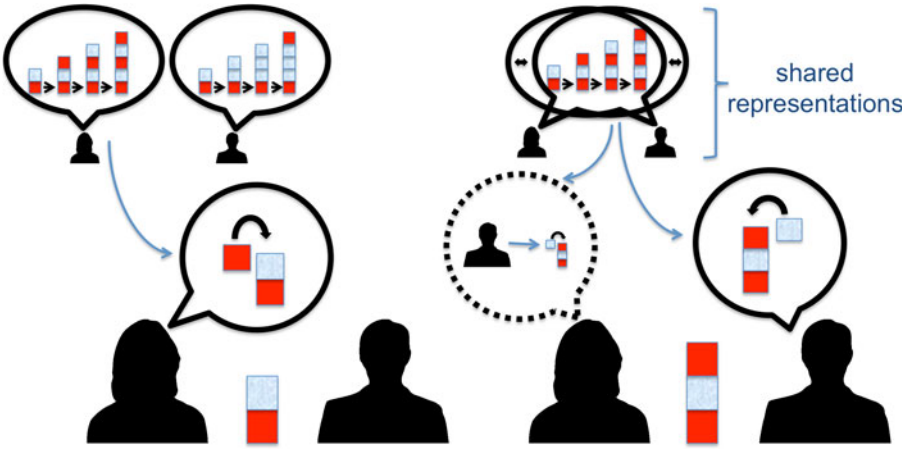


**Fig. 3** Two steps of the interactive strategy. *Left* agents have different task representations. *Right* during the course of the interaction, representations align (as shown by the *horizontal double-edges*), entailing coordinated patterns of action prediction (agent on the *left*, *dotted circle*) and selection (agent on the *right*). See main text for explanation

In the interactive strategy, shared representations play a similar role as coordination tools such as roundabouts, as they entail coordination without the need to represent the mental states of the other, interacting agents, as it is necessary in the socially-aware strategy. A sample illustration of how this can be done is shown in Fig. 3. When representations are not aligned (left), individual task representations guide action selection. Successively (right), if representations become aligned during the course of interaction, individual processes of action selection and prediction become coordinated, too, as they can be derived from the same (shared) task representations. In turn, this entails better coordination, and an easier monitoring of the (joint) task achievement.

Note that this strategy requires that representations become aligned during the course of the interaction, not that agents represent the shared part as being shared, or maintain a model of the others. In other words, contrary to the socially-aware strategy, in which people maintain separate representations of their own and their co-actor's actions, in the interactive strategy people have only representations of their own actions, which come to be aligned during interaction with the co-actor's representations of her or his actions. We use the term *shared representations* in this minimalistic sense; the terms *mutual representations* or *common ground* are more frequently used in relation to the second, meta-representational view. We will discuss this point more extensively in Section 5.2.

### 1.4 Aim and Structure of the Article

Our sketch of the interaction strategy lets three main questions unanswered. First, what is the content of shared representations, or what is shared during

a joint task such as the tower game? Performing a truly collaborative task could require sharing more information than necessary when two drivers cross a roundabout at the same time, such as for instance knowledge of common goals. Second, how do representations align during the course of the joint task? Indeed, differently from roundabouts and other physical tools, shared representations are not available perceptually, so it remains to be explained what mechanisms are responsible for their alignment, and what is the nature of such mechanisms (automatic or deliberate). Third, what are shared representation good for? Beyond what is intuitively suggested in Fig. 3, it remains to be explained if (and how) shared representations entail better planning and prediction abilities, and when is mental state inference necessary. The main aim of this article is answering these questions from conceptual and computational viewpoints.

The rest of the article is structured as follows. In Section 2 we sketch a model of action organization in which we define what representations are (can be) shared during joint action. In Section 3 we discuss how shared representation are formed. In Section 4 we discuss how, once formed, they are used as a coordination tool for implementing the *interactive strategy*, and how this activity constrains signaling and communication during interaction. Then, we draw our conclusions in Section 5. In Appendix A we offer a sketch of the interactive strategy from a computational viewpoint.

## 2 Action Organization, and What can be Shared

This theory starts from the premise that what is shared during an interaction is (part of) the individual representations for action that the two agents use for planning, execution and prediction. Although this action-based stance could seem too restrictive, research in cognitive psychology and neuroscience has shown that representations for action are surprisingly richer than believed before, as the action system is organized in such a way that low-level control of action and high-level representations, beliefs and intentions, form a continuum rather than being disconnected systems.

Current neuroscientific models describe action as hierarchically organized (Hamilton and Grafton 2007; Pacherie 2008; Wolpert et al. 2003). In keeping with this view, our computational (Bayesian) model distinguishes the levels of intentions ($I$), action representations ($A$) and motor primitives ($MP$) (see Fig. 4). The higher level represents the agent intentions and the outcome of actions. The intermediate level includes the actions necessary to achieve the intention. Note that actions are not simply movements, but are associated to a given goal (e.g. the grasping of an object rather than a hand movement per se), which they can satisfy flexibly; for instance, actions can be multimodal, and include speech, gesture, facial expressions, direction of gaze, etc. At the same time, actions are not completely specified in terms of what effector they use, how fast they are executed, what is the trajectory of the associated movement, etc. This additional level of details pertains to motor primitives and

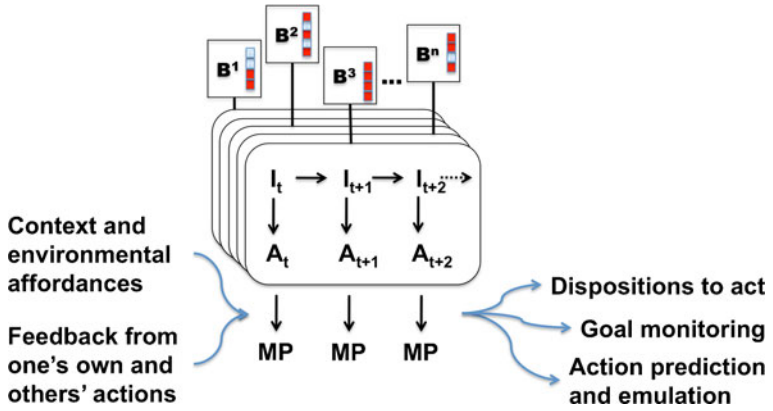


**Fig. 4** Conceptual model of action organization. Each "plate" is associated to a belief ($B$) and corresponds to the choice of a sequence of intentions ($I$) and actions ($A$). Beliefs, intentions and action representations are the "cognitive variables" that can be shared during a joint action. See main text for explanation

their parametrization. Motor primitives here correspond to the (hierarchically organized) "vocabulary" of recombinable motor elements, (e.g., for executing basic motor actions such as grasp, reach, and lift) found in premotor cortex of monkeys and humans (Rizzolatti et al. 1988).

In the model, all the variables $B$, $I$, $A$, and $MP$ can be considered as not directly observable (hidden) by an external agent; indeed, what is observable is only the overt movements, (including e.g., facial expressions, gesture and speech) that result from the choice of a given action and motor primitive. As explained in more detail in Appendix A, action understanding and mindreading refer to the estimation of hidden variables based on what is observed, or in other words the inference of what hidden variables could have generated the observed behavior of an agent. Edges encode (probabilistic) relations between the variables $I$, $A$, and $MP$; intuitively, this means that certain intentions make the choice of certain actions and motor primitives more likely.[3] Using this architecture, a given goal (e.g., building a red tower) is realized by generating a sequence of intentions $I_t$, $I_{t+1}$, $I_{t+n}$, which map into realization of (sub)goal states (e.g., having a red block over a blue block). In turn, this high-level structure leads to selection of a sequence of actions $A_t$, $A_{t+1}$, $A_{t+n}$ (e.g., grasping and lifting a specific red block) and activation of motor primitives $MP$ (e.g., power grasp with right hand). Note that only three time steps are shown, but the model can have indefinite length.

On top of the action hierarchy, beliefs ($B$) play a supporting role for selection of action and practical reasoning (for related models, see Bratman 1987;

[3]See Appendix A and Pezzulo and Dindo (2011) for a more complete description using the formalism of Dynamic Bayesian Networks.

Pacherie 2008; Wolpert et al. 2003). In this example, beliefs $B^1, B^2, \ldots, B^n$ are associated to task representations, and map intuitively into "what tower am I building". A more complete picture should include additional beliefs and contextual information, which govern the selection of intention and action sequences, such as for instance knowledge of means-ends relations (e.g., bricks can be piled, one can only pile bricks that he can reach), knowledge of different plans that achieve the same goal (e.g., two ways of piling red blocks), or even conventional information (e.g., red bricks always follow blue bricks). Here we are assuming that all this knowledge is implicit in the action structure, the $I \rightarrow A$ and $A \rightarrow MP$ relations. Our model represents the selection role played by beliefs explicitly, as a specific sequence of intentions and actions (i.e., a "plate") is associated to each belief $B^1, B^2, \ldots, B^n$. In Fig. 4, beliefs are represented at different "heights"; for instance, $B^2$ is highest. This intuitively means that $B^2$ is at the moment the most preferred (note that probabilistic models can also represent the entire probability distribution and (un)certainty of the hypotheses).

In our model, sharing representations means that part of the cognitive representations in the "plate" (beliefs, intentions, and action representations) are the same in two (or more) agents. In Section 3 we will discuss the processes through which representations become shared during the course of an interaction; ideally, in the tower game this process could lead the two agents to share an entire "plate", (or in other words the same $B$). Note that in this example this implies converging to the same goal (i.e., building the same tower), and the same choice of intentions and actions to achieve it. However, the same model can be used to represent other situations as well, such as for instance the case of two agents that already share a common goal; in this case, the "alignment" process could involve choosing the same plan among these available to achieve it (different from our tower game example, in which we assumed only one "plate" for each goal). Furthermore, not always sharing should include goals, intentions and action representations simultaneously. Consider the case of two agents moving together a table towards a goal location. In this case, alignment at the level of beliefs, goals and intentions is advantageous, but task achievement can have contrasting requirements for the two agents in terms of action (e.g., one pushes, and one pulls), and so what is required is that the patters of actions of the two agents share only certain characteristics (e.g., their temporal profile and amplitude).

To understand how this model supports joint action execution and alignment of representations, three characteristics are especially relevant. First, the same action structure is used for action planning and execution, for monitoring goal achievement, and for action prediction and emulation. This is compatible with the view that the same action representation underlies goal-directed action execution, planning, simulation, observation, and imitation (Jeannerod 2001, 2006). A computational description of this process has been proposed in relation to the MOSAIC architecture for action execution and understanding (Wolpert et al. 2003), which is akin to the one described in Fig. 4. In this architecture, the same internal (inverse and forward) models implied in action

performance are also reused for prediction of observed actions, simulation, etc. (see also Demiris and Khadhouri 2005; Dindo et al. 2011; Grush 2004; Pezzulo 2011). Second, activation of motor primitives is influenced by factors that are external to the usual belief- and intention-based mechanisms, and in particular emulation of observed actions and environmental affordances. This is compatible with the idea of an automatic perception-action linkage (Prinz 1990), by which motor primitives are activated following the observation of goal-directed actions and even affordances, and more in general with the role of automatic, "bottom-up" processes in social cognition (Frith and Frith 2008). Third, although edges in the model are top-down oriented, a characteristic of generative Bayesian models is that computations can proceed in either way: as one or more "values" are assigned to the variables, the value of all the other variables change accordingly. Normally, during action execution it is the choice of beliefs and intentions that determines the choice of action and motor primitives; however, "fixing" the value of motor primitives does affect backward the value of cognitive variables (beliefs, intentions and action representations) as well (see Kilner et al. 2007 for a related model that uses predictive coding for explaining the mirror-neuron system).

As we will see, these three elements, the reuse of the action execution system for emulation, the automatic perception-action linkage, and the possibility of motor primitives to affect backward the choice of cognitive variables, are responsible for automatic alignments of behavior and of cognitive representations during interactions.

## 3 Formation of Shared Representations

Shared representations can be formed automatically or deliberatively. What follows is a description of these two modalities.

### 3.1 From Automatic Entrainment of Behavior to Shared Representations

Research in social and cognitive science has revealed numerous examples of *entrainment* of behavior during interaction. Examples of automatic entrainment in language use are reported in Pickering and Garrod (2004), showing that people engaged in a dialogue align at several levels and tend to use the same syntactic forms; this evidence has lead to the 'Interactive Alignment Model'. Social psychologists have studied automatic entrainment of behavior for decades; one popular example is the chameleon effect (Chartrand and Bargh 1999), or the evidence that people tend to assume the same pose. Entrainment and synchronic behavior are ubiquitous phenomena when two agents are interacting, and affect their turn taking, walking speed, eye movement patterns, etc.

It has been proposed that entrainment of behavior can be explained by (automatic) priming or mutual emulation (Garrod and Pickering 2009), which are manifestations of an automatic perception-action linkage at the neural level

(Prinz 1990, 1997). In terms of our model of action organization, this produces a form of "motor contagion" in which activation of the motor primitives of one agent makes equivalent motor primitives of the other agent active as well. It has been suggested that alignment and synchronization of motor processes, kinematic characteristics, and speed profile of executed actions could enhance cooperation, in that it makes interactions more predictable (see e.g., Knoblich and Sebanz 2008; Pickering and Garrod 2004). At the same time, studies that use a paradigm of interference between executed and observed actions reveal that motor contagion and the bias to imitate another's actions can be detrimental for individual performance when the demands of the observed and simultaneously executed actions are different (Kilner et al. 2003). Indeed, emulation and motor contagion seem to be at odds with joint action, in which actions executed by two agents are typically *not* the same.

However, a case can be made that automatic processes of alignment of behavior could, as a by-product, produce alignment of cognitive variables (beliefs, intentions and action representations) and of dispositions to act. Besides those artificial scenarios that are studied experimentally, this could be favorable for joint action. In other words, if this is the case, an observer agent can be first behaviorally entrained, and then it would tend to assume the same cognitive variables as the performer agent, and as a consequence the same dispositions to act. For instance, an agent could first assume the same facial expression as observed in another agent, and then assume the same mood and mindset (from which the facial expression derives). By reviewing numerous empirical studies, Ferguson and Bargh (2004) propose that automatic alignment mechanisms produce a strong bias towards assuming another's intentions, not only its behavior. In a similar vein, Aarts et al. (2004) have proposed that perceiver that infers the goal of a performer can automatically adopt it, leading to "goal contagion".

In terms of our architecture, emulation of another's actions and the reuse of motor primitives equivalent to those used by the observed agent could produce alignment of cognitive representations (see Fig. 5). A first proposal of how this could happen, formulated by Blakemore and Frith (2005) in relation to the MOSAIC architecture (which has the same generative Bayesian structure as described before), is that motor contagion could serve to set the prior probability (priors) of observed actions and to predict them more easily. Applying the same logic to the model we have proposed, which includes sequences of task-related representations (in the "plates") rather than single actions, motor contagion could increase the priors of compatible sequences of intentions and action representations (and of associated beliefs). In this vein, by influencing the activation of motor primitives, alignment of behavior can indirectly entail alignment of disposition to act in such a way that is functional to task achievement. Another way to look at the same phenomenon is that, actions performed by others are informative of what task they are executing, and as they are interpreted by emulating them, they become mirrored in one's own representations for action, and produce a cascade process that results in an automatic alignment of task representations.

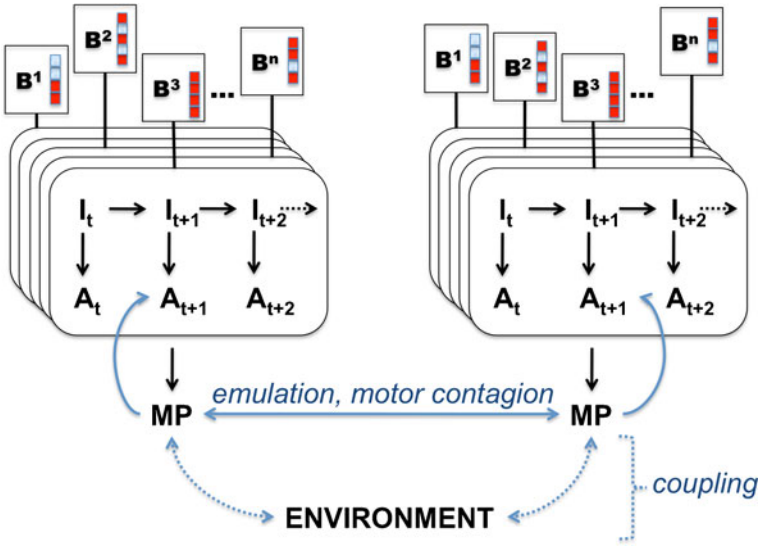


**Fig. 5** From automatic alignment and entrainment of behavior to shared representations. Emulation and motor contagion (*horizontal edge*) and environmental coupling (*dotted edges*) bias the choice of motor primitives, and this in turn influences backward the choice of cognitive variables (*bottom-up edges*). See main text for explanation

A second way cognitive representations can align automatically is through physical coupling between agents (see Fig. 5), and the resulting entrainment of their behavior so as to form coherent patterns of action. Figure 4 shows that the choice of motor primitives is influenced by feedback from another's actions. Consider again the previously described example of two agents moving together a table: actions of one agent influence the choice of actions of the other, and vice versa. The mutual adaptation to another's actions produces an entrainment of behavior of the two agents (Kelso 1995), and biases the choice of compatible motor primitives (e.g., for pushing and pulling) and of actions that have compatible characteristics (e.g., temporal profile and amplitude). As we discuss below, goal monitoring processes creates a feedback loop that, in turn, makes entrainment even stronger. Once two patterns of actions are entrained, this can influence backward the choice of other cognitive variables (e.g., "plates" with the same sequence of intentions), in the same way as described before for the case of emulation and motor contagion.

### 3.2 Deliberate Formation of Shared Representations

In addition to automatic mechanisms, interacting agents can implement deliberate strategies with the goals of aligning their representations, monitoring and fixing what is shared during the course of the interaction. This phenomenon

has been studied mainly in the context of dialogue[4] in relation to "common ground" (Clark 1996; Stalnaker 2002). However, it is not an exclusively linguistic phenomenon, as humans and other primates use multiple mechanisms for this sake, such as those for orienting social attention and gesture (Tomasello et al. 2005), as well as the aforementioned signaling strategies that consist in modifying action kinematics (Pezzulo and Dindo 2011; Vesper et al. 2010). At the same time, it does not necessarily imply common knowledge in the sense it is typically assumed in this literature (i.e., a proposition is common knowledge if every agent knows that it is true, that the others know it to be true, and so on). Rather, in our model it is sufficient to share representations for action, not meta-representations (see Section 5.2).

Besides relying on automatic mechanisms, during an interaction agents can perform signaling and communicative actions with the aim to modify the cognitive variables of the observer agent, and change what is shared. To understand this process, it is first useful to distinguish among actions that achieve communicative effects as part of their normal execution, and actions that are constitutively communicative.

In our tower game example, almost every pragmatic action, such as for instance placing a red brick, achieves both pragmatic goals (i.e., contributing to building the tower) and communicative goals (i.e., informing the other agent of what one is doing, and whether or not it is compatible with it's current hypotheses on what should be done). This is true even if the performer agent does not intend to achieve any communicative goal, since it capitalizes on the tendency of observers to interpret bodily movements and facial expressions as signs of the performer's actions and intentions.

However, agents can also *strategically* use their actions so as to convey communicative intentions. For instance, by looking at an object, pointing at it, or referring to it linguistically an agent can attract another's attention and show that it is the focus of discourse; by approaching it slowly with a hand, it can manifest its intention to lift it, or even require help, etc. These actions can be considered as communicative in that, would the performer agent not had the communicative intention, they would not have been executed, or at least they would have been executed in different ways, with a different speed profile, etc (as we will discuss below, these actions have a *cost* in terms of task performance).

In joint action contexts, the main role of these actions is modifying what is shared. The same way acting in such a way that is compliant (and predictable) given the currently shared representations implicitly confirms that they are

[4]Dialogue is itself a joint (communicative) action, and has similar characteristics as the tower game. Clark (1996) discusses extensively how utterances convey communicative intentions, but also serve to negotiate a common ground, resolve conflicts and miscommunication, keep communication in track, and ultimately to achieve (joint) communicative intentions, in addition to providing (common) ground to act jointly in the external world.

appropriate, violating the predictions that would be more natural given the shared representations implicitly signals that they have to be revised. For instance, by piling a brick in an odd way, one can signal disapproval, impatience or playfulness, whereas piling a brick in a standard way would typically not be interpreted as a prima facie communicative action (although it also implies a corroboration of the ongoing shared representations).

Coming back to the tower game example, one can signal that what is currently shared about the task goal is not correct (or that it wants to change task) by doing something that is unpredictable given it. For instance, by piling two or three red blocks an agent can communicate that it want the tower to be red, or that it is in charge of piling the red blocks (and the other agent the blue ones). Recognition a communicative intention depends on prior content of the shared representations; for instance, if the goal (color of tower) is uncertain, the former is the more plausible message; on the contrary, if both know that the tower must be red and blue, the latter message is more plausible. In turn, selection of a communicative intention has to be informative of how the observer agent has to change its representations; or, in terms of our model, what "plate" (and associated $B$) the observer agent should assume to explain (i.e., make predictable) the observed action. To resume, the deliberate sharing of representations consists in performing communicative actions inform the other agent on how it should change it's own representations to generate good predictions for the rest of the interaction.

It is worth noting that, according to our analysis, successful interaction and communication during interaction have contrasting requirements. Indeed, for the interaction to proceed smoothly, agents should aim to be predictable; in order to communicate, they should be unpredictable. This means that communicative actions have a cost for the interaction, since they interfere with the standard pragmatic goals. One implication of this view is that what is communicated, and even what is shared during the interaction, is only what is really *relevant for the ongoing interaction*, since communicating and sharing irrelevant things would hinder the interaction or at least make it more costly. This analysis suggests that parsimony and the "maxim of relevance" (Grice 1975) could be side-effects of informational dynamics and the costs of communicating during the interaction. Conversely, communicative intentions should be (and are) by default interpreted as being relevant for the ongoing interactions (in terms of signaling what should be shared), rather than, say, pertaining to different circumstances. In other words, there is a strong bias for observer agents to interpret communicative acts as being intended to change the shared representations.[5]

[5]Observers can, at the same time, infer information to revise their model of the performer agent (e.g., how skilled it is).

## 4 How Shared Representations are Used During Interactions

We have argued that the interaction strategy relies more on (automatic and deliberate) processes of sharing and alignment of representations than mental state inference. Indeed, once formed, shared representations can be used as coordination tools to facilitate action planning, prediction, and goal monitoring. Put in simple terms, an advantage of shared representations is that two (or more) agents can use the same representations for action (as encoded for instance in a "plate" in Fig. 4) to select their own actions, to predict others' actions, and for monitoring goal achievement. In this sense, shared representations plays the role of coordination tools and ensure coordination at the level of action execution and prediction of many agents.

### 4.1 Using Shared Representations for Action Planning and Execution

When two agents have shared (task) representations, the performer agent can simply use its representations for action rather than first performing recipient design. In fact, because the observer agent uses the same, shared representations, it will predict them successfully. For instance, when two agents are building a red-and-blue tower, as shown in Fig. 4, the performing agent has just to stick to the most obvious (and shared) course of action, that is grasping a red brick and placing it on top of the tower. As we have discussed, in this case good predictions of the observer agent implicitly confirm that the shared representations are reliable.

At the same time, if the representations of two agents are not aligned, the performer agent's actions will be surprising (*un*predictable) for the observer. Imagine for instance that the observer agent expects the performer agent to pile a blue brick, and it piles a blue brick instead. The same, unpredictable action counts both as a pragmatic action, which contributes to fulfilling the pragmatic goal of the performer (e.g., continue the tower with all blue bricks), and as a communicative action, indicating that the two agents are acting on different grounds, and what they share is not appropriate to continue the interaction well (of course, the way representations will be revised depends on the powers of the agents; see later).

In some cases, the interactive strategy can be more proactive, by incorporating elements of mindreading. For instance, should an agent that is monitoring goal achievement (see below) discover that the interaction is proceeding in a wrong direction (say, the other agent is piling a blue rather than red brick), it can (try to) remedy and "fix" what is shared by taking a communicative action. Examples or communicative actions are stopping the action, signaling the error linguistically, or even replacing the blue brick with a red brick, while at the same time emphasizing the communicative role of this action (e.g., doing it slower, orienting attention of the other agent towards it). The choice of communicative actions should be executed is done on top of considerations

of informativeness, in that from that action the observer agent should infer how its representations should be revised (see below).

### 4.2 Using Shared Representations for Action Observation

In most cases, observer agents can successfully predict actions of others by using their (shared) representations, and do not need to use mindreading. In other words, sharing representations permits to use one's own representations for action as a model of another's actions during perceptual processing, which is computationally similar to theories of emulation (Grush 2004; Wilson and Knoblich 2005), except that it implies higher-level (task) representations rather than only action representations.

Should the observed actions become unpredictable, then mindreading becomes necessary to extract the communicative intention and, in case, understand what part of the shared representations should be changed (but note that now mindreading is *postdictive*). Indeed, as we have discussed, being unpredictable is the main way performer agents can convey their communicative intentions.

To decide how to change its representations (e.g., passing from $B^1$ to $B^2$, $B^3$ or $B^n$ in the example of Fig. 4), an observer agent can infer what $B$ could have generated the choice of the observed action, or in other words under which conditions the observer action would have been less surprising. Because actions are informative of the hidden variables that could have generated them, this process is computationally similar to what described before in relation to action recognition and the MOSAIC architecture, except that the inference is done at the level of task representations rather than action goals. This inference could be realized using mechanisms that "simulate" what would be implied by the different $B$s (Demiris and Khadhouri 2005; Pezzulo and Dindo 2011; Wolpert et al. 2003), or via "inverse inference" (Baker et al. 2009), by considering how what is observed diverges from what would be rational.

### 4.3 Using Shared Representations for Goal Monitoring

Monitoring goal achievement, or assessing if the task is proceeding towards the goal, is another fundamental process in joint action, as it is in individual action performance (Botvinick et al. 2001). In individual set-ups, monitoring can be performed by using predictive mechanisms and a *comparator* between expected results of actions and desired goal; the use of predictive mechanisms ensures that actions expected to be detrimental can be stopped on time (Pezzulo and Castelfranchi 2009). A similar mechanism can be adopted in joint tasks, because the same shared representations and the same predictive mechanisms can be reused for monitoring both individual and shared goals. For instance, when two agents lift together a table, they can monitor the (joint) goal of "taking the table horizontal" using shared representations of (sequence of) intentions and actions, $I_t$, $I_{t+1}$, $I_{t+2}$, independent of who is realizing the actions that satisfy these intentions at time $t$, $t+1$ or $t+2$. This makes it

possible to coordinate the patterns of actions of two agents having different skills, or doing different parts of a task (e.g., in the table moving example, one pushes and another pulls). In this case, it is not necessary that agents represent their own and the other's actions, at least as long as they share intentions and monitor them with success.

An important side-effect of monitoring joint goals is that each agent can take corrective actions, such as for instance raising or lowering the table to keep it horizontal, and as these actions are sensed by the other agent, they can determine a feedback loop which stabilizes joint performance and keeps the table horizontal. To understand how this is possible, we discuss below how coordination of behavior depends on both shared representations and considerations of situatedness.

### 4.4 Coordination of Behavior: Shared Representations Plus Situatedness

Sharing representations could not be sufficient per se to determine success of real-world interactions. Most joint tasks, such as moving together a table, should incorporate elements of real-time coordination, which in most cases are not (or cannot be) explicitly represented or shared. For this reason, in the interactive strategy shared representations work in combination with elements of situatedness, and in particular the constraints introduced by context and environmental affordances on the one hand, and actions of the other agents on the other hand.

Environmental affordances can provide a first mean of coordination, in that both agents will take them into consideration into their choice of actions and intentions, raising the probability that (part of) their representations are indeed aligned. For instance, if two agents that transport a table are passing through a door, physical constraints such as amplitude of the door will influence the choice of intentions and actions of both at the same time (in most cases, this does not need to be explicitly represented). Not only the environment, but also the fit between the characteristics of the co-actors (e.g., one of the two can be stronger or taller) imposes constraints on the joint task, influencing how it unfolds in time (Isenhower et al. 2010). Furthermore, actions performed by one agent induce constraints for the other agent as well, as they change the context in which interaction takes place, create novel affordances, and offer cues for coordination. For instance, a turn-based structure can emerge if an agent refrains from picking a block after seeing that the hand of another agent is already picking it up; if this is done repeatedly, a pacing of actions can emerge without being explicitly represented (for related arguments in dynamical systems theory, see Kelso 1995).

These considerations of situatedness are further magnified when two agents do not only operate in the same environment, but use shared representations, as in this case not only their actions but also their processes of planning, prediction and goal monitoring become entrained. Consider the case of two agents sharing (even partially) the same "plate". When the first agent sees $A_t$, it can use the "plate" to foresee achievement of $I_t$ (independent of who

caused it), and then start planning $A_{t+1} \rightarrow I_{t+1}$, so as to execute it when the environmental conditions become appropriate. Similar to our previous considerations on turn-based structures, this example suggests that "division of labor" can emerge without representing who does what, and involve agents using different skills and action representations, which coordinate at the level of intentions.

### 4.5 Possible Advantages of the Interactive Strategy

We have argued that the interactive strategy could make interaction simpler and potentially more successfully than the other two strategies, individual and social-aware. Although a direct comparison of these strategies is beyond the scope of this article, it should be noted that the formation of shared representations, despite being costly in the short run, could provide advantages in terms of task achievement and cognitive load.[6] First, with this strategy it is sufficient to maintain only individual representations for action (that in the tower game example are essentially a model of the ongoing task) rather than two (or multiple) models (i.e., a model of oneself and one for each other agent participating in the interaction), because the same structure can be used for action planning and execution, prediction, and goal monitoring. Second, coordinating based on what is shared permits, in most cases, to produce appropriate behavior and good predictions without mindreading others. Third, shared representations are a good ground for conveying (and recognizing) communicative intentions, as unpredictable actions are informative that what is currently shared is not appropriate, and offer cues on what should be shared that could make the observable actions predictable.

## 5 Conclusions

So far, most models in economy, game theory and AI have assumed, implicitly or explicitly, that coordination and joint action make use of the social-aware strategy, or its variants. In other words, when two agents (performer and observer) interact, they need to maintain separate models of their own and their co-actor's actions, and mind-read one another recursively, for each turn of the interaction. The performer agent uses mindreading for *recipient design* so as to derive the effects of its actions on the observer agent, and the observer agent uses mindreading for *intention recognition* so as to infer and predict the observed actions. However, this method could have heavy demands in terms of continuous, recursive mindreading or, alternatively, put heavy burden on communication, in terms of a common ontology and the necessity of exchanging many messages in order to reach a full consensus.

[6] Forming shared representations could be useful in competitive scenarios, too. Indeed, channelizing predictions of the adversary facilitates tricking and feinting it when necessary.

We have described an alternative, *interactive strategy*, which explains why interaction is so easy, and how coordination and task allocation can be realized without too much mental state inference, communication and conventions. For most interactions and joint actions to proceed smoothly, it is sufficient that all agents select their actions using their representations for action implied in action execution and observation, because when they are shared, they can be considered a (joint) model of the ongoing interaction. As cognitive representations of two agents align, they begin generating coherent patterns of action execution and prediction, which facilitate successful task completion and monitoring. In other words, by sharing (part of) their representations for actions, agents can help one another to solve interaction problems, and facilitate coordination of actions and intentions; see Vesper et al. (2010) for the related idea of "coordination smoothers".

This proposal is therefore different from the idea that interacting agents should share "scripts" or conventional knowledge. However, we do not consider these alternatives to be mutually exclusive, as conventional knowledge is plausibly used in many real-world scenarios, such as for instance interior designers making plans for furnishing a house, or a team of athletes participating in a competition.[7] One advantage of this proposal is that, it spans from abstract formulation of joint action plans, to low-level coordination of actions and movements.

We have discussed how sharing of representations results from both automatic mechanisms of mutual priming, emulation, and physical coupling, and deliberate strategies of signaling and communication. At the same time, our analysis indicates that often communication has contrasting requirements than pragmatic action, and then it has to be used with parsimony in order not to hinder the interaction. In turn, this implies that only the minimum amount of representations for action that affords task achievement should be shared.

Another consequence of our analysis is that, during interaction, mental state inference could be less necessary than commonly believed. However, we are not suggesting that it is not used at all. We believe that the interactive and social-aware strategies are not mutually exclusive but may act in concert. It turns out from our analysis that mindreading is necessary in at least four cases. First, for conveying communicative actions, so as to derive what action is more informative of what should change in the shared part. Second, for interpreting ambiguous and unpredictable situations.[8] Third, for performing *helping or hindering actions*; indeed, in order to help or hinder other agents during the interaction it is first necessary to infer their goals (unless they are already part of the common ground). Fourth, when there is a failure in the interaction,

[7]Note also that not all representations for action need to be "mental" objects; action plans can include external representations (Kirsh 2010) as well, such as maps, diagrams to which both agents can refer (e.g., linguistically or by pointing at them), of objects that are under the attention of both agents (such as the two-bricks towers in Figs. 1, 2, and 3).

[8]More precisely, when the source of errors in prediction cannot be attributed to random events, but is more plausibly interpreted as depending on the other agent's beliefs and intentions.

and it is necessary to derive the specific role of each agent, or *who should have done what* (e.g., for complaining). In this case, mindreading can be used postdictively.

### 5.1 What is Shared During Interaction

According to our analysis, the term *shared representations* does not indicate that the mental representations of the two agent are always the same. It is worth noting that representations are only partially *shared*, and in case agents have beliefs on what is shared, these beliefs can be different. This is one of the main sources of failure of interaction and requires continuous adjustments and 'alignments' of the shared part. In turn, a positive aspect is that the alignment process can be more rapid as the interactive strategy requires that only a part of representations is aligned. Behind automatic effects, aligning representations has a cost, and so this is done only as long as it produces advantages in terms of facilitating prediction and ultimately achieving coordination. In addition to that, it is often the case that two agents do not want to share certain goals or beliefs, especially (but not only) in the case of competitive scenarios. Therefore, although interacting agents have significantly similar background knowledge even before any interaction (especially if they belong to the same culture), it is plausible to assume that during interaction only a limited amount of task-related representations are really shared.

Our model is minimalistic in that it only includes essential ingredients to perform joint actions in a shared, situated environment, such as for instance building a tower collaboratively. However, there are many kinds of joint actions, which could require different parts of the intention-to-action hierarchy to be shared. For instance, in the tower game, in which the two actors have the same set of available actions, it is plausible that they align their action representations. In other circumstances, such as for instance when they must perform different (e.g., complementary) actions, or combined actions that neither can execute by itself, alignment and monitoring could be done at the level of intentions, not of actions.

Furthermore, joint actions that do not require real-time coordination (e.g., two architects designing a new house), or which extend for long periods of time (e.g., planning a football game), could require additional constraints, such as for instance explicit agreements on who does what, who should move first, and increasingly use linguistic communication for sharing beliefs. These considerations suggest that formation of shared representations (like the formation of common ground) is highly context- and task-dependent. In addition, it depends on powers and role of the actors, as in certain circumstances it makes sense that one agent "aligns" its representations to those of another, but in other circumstances the two agents can find a compromise solution. An interesting line of research that addresses the search for such compromise solution is the study of how "artificial" semiotics arise in experimental situations in which participants have to create ex novo a linguistic code for communicating (Galantucci 2009).

### 5.2 Implications of our View for Philosophy and Cognitive Science

The issues we have addresses touch lively debated issues in philosophy and cognitive science. The first issue is whether or not having shared representations implies representing oneself as sharing representations (i.e., having a meta-representation on what is shared), or knowing that we know that we share the representation. Our model does not use meta-representations of what is shared (also called *mutual representations*), but only individual representations that happen to be shared and aligned—being the agents aware or not. In other words, the interaction strategy relies on the fact that representations are shared, not that the shared part is explicitly represented; similarly, in a roundabout we don't need to explicitly represent the fact that others know that we are all in a roundabout (although this could be the case), and in Figs. 1, 2, and 3 an agent does not need to represent the fact that the other agent sees the two piled blocks in front of them. However, the current scope of our theory is restricted to the kind of joint actions that, similar to the "tower game", are situated, are executed within a short time span, and do not require (too much) conventional knowledge. Meta-beliefs could play a role in more complex cases of joint action that are not captured by our model, such as for instance designing together a building or planning together a holiday.

A second issue is whether or not interacting people have *we-intentions* (Searle 1995) as distinct from individualistic intentions. Also related to this, one can ask whether or not performing joint actions implies team reasoning (Sudgen 2003) or seeing ourselves as part of a team (Bacharach 2006). Our theory shares with team reasoning the idea that agents tend to take into consideration another's payoffs rather than just their own, and play a joint strategy. At the same time, in our model success of interaction depends on the fact that *individual* representations for action are aligned, not on collective representations or meta-knowledge that is disconnected from action execution. Furthermore, our theory analyzes how representations become aligned, whereas theories of team reasoning tend to emphasize conventions.

Finally, our theory is related to Bratman's (1987) idea that *intentions* help both inter-temporal coordination within the same agent, and coordination among multiple agents (note that in this theory an intention maps roughly to the content of one of the "plates" in our model, not to a cognitive variable as we defined it). We have proposed a mechanistic explanation of how this could be possible.

### 5.3 Implications of our View for the Study of Social Cognition

Our proposal of an interactive strategy has implications for research in cognitive (and social) psychology neuroscience. In the analysis of the results of most experiments, especially those involving observation of actions performed by others, it is often tacitly assumed that the two agents are implementing a social-aware strategy, or one of its variants. This leads to the idea that the "social brain" is continuously engaged in mindreading others, mapping its

actions into one's own action repertoire, or inferring what action is it doing according to rationality principles. The perspective that we have proposed is slightly different in that it suggests that, in most cases, these operations can be replaced by simpler predictions afforded by (shared) representations for action.

The idea that the social brain adopts predictive mechanisms is certainly not novel (Frith and Frith 2006; Gallese 2009; Pezzulo and Castelfranchi 2007, 2009; Umiltà et al. 2001; Wolpert et al. 2003). However, our perspective suggests that the (main) use of such predictions is as part of action planning for coordination and joint action rather than, say, as a part of a mindreading process or of imitation, which could be more typical in "passive" action observation setups that are commonly used in the aforementioned experiments. In other words, in the interactive strategy the burden of neural computation during social cognition and interaction lies in *deciding what should I do to coordinate with you (and when)* rather than in *inferring what you are doing*, and prediction could be more functional to the former than the latter (Pezzulo 2008). We believe that this theoretical perspective could guide ongoing research on social cognition and especially joint action. In a series of recent studies, where the focus is on interaction rather than passive action observation, evidence begins to accumulate that the brain uses resources to encode another's actions in terms of complementary actions—an encoding that functional to the continuation of ongoing interactions–rather than only in terms of observed actions (Newman-Norlund et al. 2007, 2008). Indeed, an "interactive" encoding of actions and tasks could be extremely useful from computational and evolutionary perspectives. Moreover, because certain actions cannot be executed by one single agent but require collaboration, it would be useful to represent (parts of) the joint task in terms of actions to be performed jointly rather than individually (see Tsai et al. 2011 for a relevant study). Overall, this body of evidence suggests that interaction rather than passive observation is the most typical scenario to which our social brain could be adapted. "Resonance" mechanisms of the so-called social brain could have evolved to facilitate alignment of representations, given the adaptive advantages of collaboration. At the same time, they makes us humans potentially more prone to deception, but also to other forms of misbehavior, which for instance manifest as (undesired) "team behavior" when two or more individuals are simply aware of one another during task performance (Sebanz et al. 2005), in which collaboration is not advantageous. In these cases, it is possible that, because agents tend to share their representations even when this is not necessary or even useful, they could align to a common goal that is indeed detrimental for their real task.

Finally, we have proposed (together with many others) that an important activity during interaction and joint action consists in the formation and maintenance of shared representations. According to our analysis, not only do they align automatically via mutual emulation and priming, but they are also deliberately changed as part of the interactive strategy. We have suggested that communicative actions (verbal, gestural, etc.) are used to trigger changes

in the shared representations. Studying how communicative actions are used so as to modify shared representations during the course of interaction is an important direction of research for social cognition, we believe, and could gather confirming or disconfirming evidence for the interactive model we have proposed.

**Acknowledgements** Research funded by the EU's FP7 under grant agreement n. FP7-231453 (HUMANOBS). I want to thank Haris Dindo and David Kirsh for fruitful discussions, and two anonymous reviewers for helpful comments.

## Appendix A: The Interactive Strategy from a Computational Viewpoint

From a computational perspective, we can model each agent as a generative (Bayesian) system in which hidden (i.e., not visible) cognitive variables, beliefs ($B$) and intentions ($I$), determine the selection of actions ($A$) and then motor primitives ($MP$); in turn, they determine the agent's overt behavior, which becomes part of the world state ($S$). As the world state is only partially observable, we distinguish observables ($O$) from the full, unobservable state of the world $S$.[9] Note that, different from what shown in Fig. 4, we assume that beliefs are not only relative to the task, but are more generally contextual knowledge that influences intention and action selection.

### A.1 Action Planning and Execution

For the sake of simplicity, we can assume that each action $A$ executed at time $t$ achieves a certain action goal at time $t+n$, or in other terms it determines a future goal state $S_{t+n}$ (in doing this, we skip the level of motor primitives $MP$ that are involved in action performance). If we also assume that there is a way to map an agent's intentions ($I$) into goal states $S_{t+n}$ (e.g., the intention of realizing a tower of red bricks can mapped in a state of the world in which there are 5 stacked red bricks), then planning consists in the choice of an action, or a sequence of actions, conditioned to the agent's belief and intentions, which (is expected to) realize the future goal state $S_{t+n}$ (and typically, as a consequence, produce some reward, or even maximize reward), which can be done for instance with probabilistic planning methods (Bishop 2006).

In the passage from plans to action execution, however, the mapping from desired goal states to (sequences of) actions is typically ill-posed and difficult, and for this reason it has been proposed that the brain makes use of internal (inverse) models to solve it. In addition to that, during action execution internal (forward) models could be adopted as well to adapt the motor plan

[9]This formulation is typical of POMDP (Kaelbling et al. 1998). See Bishop (2006) for reference on Bayesian generative systems, and Pezzulo and Dindo (2011) for a more complete treatment of the interactive strategy from a computational viewpoint.

to the fast dynamics of the environment, in the cases in which feedback is too slow (Desmurget and Grafton 2000; Kawato 1999).

### A.2 Prediction and State Estimation

As we have discussed, forward models serve (sensory and state) prediction during action execution. Formally, forward models permit to map state and action information into a (sensory or state) prediction at the next time step, or more than one time steps in the future (i.e., $S, A \rightarrow S_{t+n}$).

However, in dynamic environments, the effects of actions ($S, A \rightarrow S_{t+n}$) are not easy to determine, for many reasons: first, in general agents can only access the observable part of the environment ($O$) and not its "true" state ($S$); second, the environment changes over time (i.e., $S_t$ is different from $S_{t+1}$). Since actions can have different effects when executed in different contexts, and this can hinder the achievement of goals, a solution of this problem consists in *estimating* the true state of the environment ($S$) rather than acting based on $O$, and at the same time learning to predict the dynamics of the environment ($S_t \rightarrow S_{t+1}$). In generative Bayesian systems, the probabilistic inference $P(S|O)$ can be done, for instance, via iterative methods such as Kalman filtering (Kalman 1960) (other, more sophisticated methods are necessary for non-linear cases).

### A.3 Action Prediction, Understanding and Mindreading

In interactive scenarios, the situation is even more complex, since the behavior of other agents is an extra source of dynamics. Again, forward models can be used to predict the interactive dynamics and to adapt to them. However, remind that agents have only access to the observable part of the environment ($O$), and this makes prediction (and hence action planning and coordination) difficult. In analogy with state estimation, prediction accuracy can be ameliorated by estimating the "true" state ($S_t$) behind its observable part ($O$). Now, the "true" state ($S_t$) is determined by both environmental dynamics ($S_{t-1} \rightarrow S_t$) and actions ($A$) of performed by the other agent(s). Therefore, in addition to modeling the environmental transitions ($S_{t-1} \rightarrow S_t$), it is also advantageous to model and predict the actions performed by the other agent(s). (Note that agents have at their disposal also information of what parts of the current state are produced by *their own* past actions, therefore can "cancel out" the self-produced part from the stimuli and the explicandum Blakemore et al. 1998; Frith et al. 2000.)

More in general, by noting (intentionally or unintentionally) that actions ($A$) executed by other agents are selected on the basis of its hidden cognitive variables, its beliefs ($B$) and intentions ($I$), an even more complete solution consists in inferring these cognitive variables, too. From a computational viewpoint, then, mindreading can be conceptualized as the estimation of the hidden (cognitive) variables of the other agent, rather than simply the observation and prediction of its overt behavior.

Computational methods for mindreading under this formulation have formal similarities with state estimation, but are more complex, because the generative architecture that generates the observables is richer. What makes mindreading easier is the assumption that the observed agent has a similar generative architecture as one's own; indeed, constraining the structure of a generative model makes it easier to infer the value of its variables.

Under this formulation, mental state inference can be performed at different levels (e.g., estimation of actions, intentions or beliefs), can use whatever *prior* information available, such as for instance knowledge of the preference of an agent for certain actions and not others (i.e., $P(A)$), and whatever source of information available. Current implementations of mindreading range from *inverse planning*, which compares an agent's actions against a normative, rational principle (Baker et al. 2009), to *motor simulation* (Demiris and Khadhouri 2005; Wolpert et al. 2003), which compares an agent's actions with the putative effects of one's own (derived by re-enacting one's own motor system 'in simulation'); see also Cuijpers et al. (2006).

Note that, in this process, motor contagion or the alignment of motor primitives in performer and observer agents is highly advantageous, in that another's $MP$ can be considered as "observations" and used to facilitate inference of the underlying beliefs, intentions and action goals that could have generated them. A second element that facilitates estimation is a consideration of what environmental affordances are available. Indeed, availability of affordances and other contextual information (e.g. where the interaction takes place) can be considered as factors that raise the prior probability $P(A)$ of actions that are associated to given affordances and contexts. If this information becomes available to an observer, it can help in the inferential process.

### A.4 Advantages of Using Shared Representations and the Interactive Strategy

Within our model, shared representations ($SR$) can be considered as the aligned subset of beliefs, intentions and actions of two (or more) agents, or in other terms the equivalence of a subset $I^{\text{agent1}}$ with a subset of $I^{\text{agent2}}$, a subset of $B^{\text{agent1}}$ with a subset of $B^{\text{agent2}}$, and a subset of $A^{\text{agent1}}$ with a subset of $A^{\text{agent2}}$. It is worth reminding that it is not necessary (and not even true) that *all* the cognitive variables are shared. This also entails that, if agents have meta-beliefs on what is shared, they can be different. First, only cognitive variables pertaining to task achievement are generally shared. Second, as shown in Fig. 3, representations can become shared during the interaction; an indeed from a computational viewpoint the guarantee of successful interactions is that (only) representations that afford coordinated patterns of action execution and prediction become aligned.

At the beginning of the interaction only some beliefs and intentions are shared, mainly because of past experience, reliance on a common situated context, and the recognition of social situations and agreements. During the course of interaction, both agents can perform actions that have as their goal

changing $SR$ rather than achieving goals defined in terms of external states of affairs. These are *communicative* actions that typically change another's beliefs and intentions.

One advantage of using shared representations during action planning and observation (including planning and observation of communicative actions) is that they are a novel source of information, and one that is *available* to both agents, and *reliable*, since both agents put their effort in maintaining it and signaling when it has to be updated. By using shared representations, an agent can more easily predict and understand the actions of the other agents even without estimating the "true" state of the world or its "true" cognitive variables (i.e., perform $P(A|SR)$ rather than $P(A|O)$ or $P(A|I, O)$). In turn, this can be done either by appealing to a principle of rationality (i.e., what would be the best action given $SR$), similar to the method in Baker et al. (2009) (with the difference that not the inference is much more constrained than asking what is the best action in general), or by using forward search and motor simulation, similar to Demiris and Khadhouri (2005), Wolpert et al. (2003), and also by using $SR$ as *priors* that bias the search.

Note also that, as suggested in Section 4.5, in order to implement the aforementioned form of planning it is not necessary to maintain a model of the other agent, or a model for each of the other agents, but only individual representations for action tied to task achievement, which can be used for both action planning and prediction of actions of the other agents.[10]

Shared representations give advantages also in the planning of communicative actions. As we have highlighted, the standard way of expressing communicative intentions under this formulation consists in violating what would be predictable given $SR$. Here, again, the inference of what is surprising given $SR$ is easy and does not require any complex inferential mechanism, since the necessary source of information (the $SR$) is already available to the planner agent. In turn, choosing *when* and *what* to communicate (or how much to share) is more complex; indeed, most systems inspired to the social-aware strategy implicitly assume that it is necessary to share a lot of information, such as for instance my own and your part of a common plan, my short-term and long-term intentions, etc.

However, as suggested in Section 4.5, communicative actions that change $SR$ can be interpreted as a form of "teaching" for the other agents how to generate good predictions, or in other words communicating it what is the value of $SR$ that is more accurate in predicting $P(A|SR)$. This gives rise to a mutual form of supervised learning. For this form of learning to be efficacious, the "learning episode" (i.e., the message) should not be casual, but selected on purpose by the other agent to be *on time* and *maximally informative*.

In this context, being "on time" means that I should communicate when I know that your future inference $P(A|SR)$ would be wrong, or in other terms I

[10]This does not exclude that task representations can be richer; this is the case, for instance, when only one of the participating agents is able to perform a certain action.

know that the common ground is not sufficient for you to do good predictions. This criterion is used to decide when it is necessary that I perform a communicative action, so as that my actions and intentions are not misunderstood. (In addition to that, occasionally I change the $SR$ as part of my pragmatic actions, for instance adding a blue brick rather than a red one. In this case, however, the communicative implicature is automatic and no further planning is needed.)

To select "what" to communicate, instead, the criterion is informativeness, and consists in lowering uncertainty and entropy of the $SR$ (which, in our example of Fig. 4, is equivalent to lowering uncertainty on $B$) and raising the probability that the next predictions of the other agent will be more accurate. This entails that not necessarily I have to share with you all that I believe or intend to do, but only create a good ground for you to predict my actions well, to detect violations of the $SR$, and ultimately to conduct a successful interaction. Any additional information should be better not shared, since sharing it has a cost in terms of the success of the interaction.

## References


Aarts, H., Gollwitzer, P., and Hassin, R. 2004. Goal contagion: Perceiving is for pursuing. *Journal of Personality and Social Psychology* 87:23–37.

Bacharach, M. 2006. In *Beyond individual choice*, eds. N. Gold, and R. Sugden, Princeton, NJ: Princeton Univ. Press. http://press.princeton.edu/titles/8174.html.

Baker, C.L., Saxe, R., and Tenenbaum, J.B. 2009. Action understanding as inverse planning. *Cognition* 113(3):329–349.

Bishop, C.M. 2006. *Pattern recognition and machine learning*. Springer.

Blakemore, S.-J., and Frith, C. 2005. The role of motor contagion in the prediction of action. *Neuropsychologia* 43(2):260–267.

Blakemore, S.-J., Wolpert, D.M., and Frith, C.D. 1998. Central cancellation of self-produced tickle sensation. *Nature Neuroscience* 1(7):635–640.

Botvinick, M.M., Braver, T.S., Barch, D.M., Carter, C.S., and Cohen, J.D. 2001. Conflict monitoring and cognitive control. *Psychological Review* 108(3):624–652.

Bratman, M. 1987. *Intentions, plans, and practical reason*. Harvard University Press.

Brown-Schmidt, S., Gunlogson, C., and Tanenhaus, M.K. 2008. Addressees distinguish shared from private information when interpreting questions during interactive conversation. *Cognition* 107(3):1122–1134.

Chartrand, T.L., and Bargh, J.A. 1999. The chameleon effect: the perception-behavior link and social interaction. *Journal of Personality and Social Psychology* 76(6):893–910.

Clark, H.H. 1996. *Using language*. Cambridge University Press.

Cuijpers, R.H., van Schie, H.T., Koppen, M., Erlhagen, W., and Bekkering, H. 2006. Goals and means in action observation: A computational approach. *Neural Networks* 19(3):311–322.

Demiris, Y., and Khadhouri, B. 2005. Hierarchical attentive multiple models for execution and recognition (hammer). *Robotics and Autonomous Systems Journal* 54:361–369.

Desmurget, M., and Grafton, S. 2000. Forward modeling allows feedback control for fast reaching movements. *Trends in Cognitive Sciences* 4:423–431.

Dindo, H., Zambuto, D., and Pezzulo, G. 2011. Motor simulation via coupled internal models using sequential monte carlo. In *Proceedings of IJCAI 2011*.

Ferguson, M.J., and Bargh, J.A. 2004. How social perception can automatically influence behavior. *Trends in Cognitive Sciences*, 8(1):33–39.

Frith, C.D., Blakemore, S.J., and Wolpert, D.M. 2000. Abnormalities in the awareness and control of action. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 355(1404):1771–1788.

Frith, C.D., and Frith, U. 2006. How we predict what other people are going to do. *Brain Research* 1079(1):36–46.
Frith, C.D., and Frith, U. 2008. Implicit and explicit processes in social cognition. *Neuron* 60(3):503–510.
Galantucci, B. 2009. Experimental semiotics: A new approach for studying communication as a form of joint action. *Topics in Cognitive Science* 1(2):393–410.
Gallese, V. 2009. Motor abstraction: A neuroscientific account of how action goals and intentions are mapped and understood. *Psychological Researchs* 73(4):486–498.
Garrod, S., and Pickering, M.J. 2009. Joint action, interactive alignment, and dialog. *Topics in Cognitive Science* 1(2):292–304.
Gergely, G., and Csibra, G. 2003. Teleological reasoning in infancy: the naive theory of rational action. *Trends in Cognitive Sciences* 7:287–292.
Grice, H.P. 1975. Logic and conversation. In *Syntax and semantics*, eds. P. Cole, and J.L. Morgan, vol. 3. New York: Academic Press.
Grosz, B.J. and Sidner, C. 1990. Plans for discourse. In *Intentions in communication*, eds. P.R. Cohen, J. Morgan, and M.E. Pollack, MIT Press.
Grush, R. 2004. The emulation theory of representation: Motor control, imagery, and perception. *Behavioral and Brain Sciences* 27(3):377–96.
Hamilton, A.F.d.C., and Grafton, S.T. 2007. The motor hierarchy: From kinematics to goals and intentions. In *Sensorimotor foundations of higher cognition*, eds. P. Haggard, Y. Rossetti, and M. Kawato, Oxford University Press.
Horton, W.S., and Keysar, B. 1996. *Cognition* 59(1):91–117.
Isenhower, R.W., Richardson, M.J., Carello, C., Baron, R.M., and Marsh, K.L. 2010. Affording cooperation: Embodied constraints, dynamics, and action-scaled invariance in joint lifting. *Psychonomic Bulletin & Review* 17(3):342–347.
Jeannerod, M. 2001. Neural simulation of action: A unifying mechanism for motor cognition. *NeuroImage* 14:S103–S109.
Jeannerod, M. 2006. *Motor cognition*. Oxford University Press.
Kaelbling, L.P., Littman, M., and Cassandra, A.R. 1998. Planning and acting in partially observable stochastic domains. *Artificial Intelligence* 101:99–134.
Kalman, R.E. 1960. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering* 82(1):35–45.
Kawato, M. 1999. Internal models for motor control and trajectory planning. *Current Opinion in Neurobiology* 9:718–27.
Kelso, J.A.S. 1995. *Dynamic patterns: The self-organization of brain and behavior*. MIT Press, Cambridge, Mass.
Kilner, J., Paulignan, Y., and Blakemore, S. 2003. An interference effect of observed biological movement on action. *Current Biology* 13:522–525.
Kilner, J.M., Friston, K.J., and Frith, C.D. 2007. Predictive coding: An account of the mirror neuron system. *Cognitive Processing* 8(3):159–166.
Kirsh, D. 2010. Thinking with external representations. *AI & Society* 25(4):441–454.
Knoblich, G., and Sebanz, N. 2008. Evolving intentions for social interaction: From entrainment to joint action. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 363(1499):2021–2031.
Newman-Norlund, R.D., Bosga, J., Meulenbroek, R.G.J., and Bekkering, H. 2008. Anatomical substrates of cooperative joint-action in a continuous motor task: Virtual lifting and balancing. *Neuroimage* 41(1):169–177.
Newman-Norlund, R.D., van Schie, H.T., van Zuijlen, A.M.J., and Bekkering, H. 2007. The mirror neuron system is more active during complementary compared with imitative action. *Nature Neuroscience* 10(7):817–818.
Pacherie, E. 2008. The phenomenology of action: A conceptual framework. *Cognition* 107:179–217.
Pezzulo, G. 2008. Coordinating with the future: The anticipatory nature of representation. *Minds and Machines* 18(2):179–225.
Pezzulo, G. 2011. Grounding procedural and declarative knowledge in sensorimotor anticipation. *Mind and Language* 26(1):78–114.

Pezzulo, G., and Castelfranchi, C. 2007. The symbol detachment problem. *Cognitive Processing* 8(2):115–131.
Pezzulo, G., and Castelfranchi, C. 2009. Thinking as the control of imagination: A conceptual framework for goal-directed systems. *Psychological Research* 73(4):559–577.
Pezzulo, G., and Dindo, H. 2011. What should i do next? using shared representations to solve interaction problems. *Experimental Brain Research*. http://www.springerlink.com/content/v1626220237466x2/.
Pickering, M.J., and Garrod, S. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences* 27(2):169–90; discussion 190–226.
Prinz, W. 1990. A common coding approach to perception and action. In *Relationships between perception and action*, eds. O. Neumann, and W. Prinz, 167–201. Berlin: Springer Verlag.
Prinz, W. 1997. Perception and action planning. *European Journal of Cognitive Psychology* 9:129–154.
Rizzolatti, G., Camarda, R., Fogassi, L., Gentilucci, M., Luppino, G., and Matelli, M. 1988. Functional organization of inferior area 6 in the macaque monkey. ii. area f5 and the control of distal movements. *Experimental Brain Research* 71(3):491–507.
Rizzolatti, G., and Craighero, L. 2004. The mirror-neuron system. *Annual Review of Neuroscience* 27:169–192.
Searle, J. 1995. *The Construction of social reality*. New York: The Free Press.
Sebanz, N., Bekkering, H., and Knoblich, G. 2006. Joint action: Bodies and minds moving together. *Trends in Cognitive Science* 10(2):70–76.
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(6):1234–1246.
Stalnaker, R. 2002. Common ground. *Linguistics and Philosophy* 25:701–721.
Sudgen, R. 2003. The logic of team reasoning. *Philosophical Explorations* 16(3):165–181.
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(5):675–91; discussion 691–735.
Tsai, J.C.-C., Sebanz, N., and Knoblich, G. 2011. The groop effect: groups mimic group actions. *Cognition* 118(1):135–140.
Umiltà, M.A., Kohler, E., Gallese, V., Fogassi, L., Fadiga, L., Keyers, C., and Rizzolatti, G. 2001. I know what you are doing. A neurophysiological study. *Neuron* 31(1):155–65.
van der Wel, R., Knoblich, G., and Sebanz, N. 2010. Let the force be with us: Dyads exploit haptic coupling for coordination. *Journal of Experimental Psychology: Human Perception and Performance*. http://www.ncbi.nlm.nih.gov/pubmed/21417545.
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* 131:460–473.
Wolpert, D.M., Doya, K., and Kawato, M. 2003. A unifying computational framework for motor control and social interaction. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 358(1431):593–602.
Wolpert, D.M., Gharamani, Z., and Jordan, M. 1995. An internal model for sensorimotor integration. *Science* 269:1179–1182.
Yoshida, W., Dolan, R.J., and Friston, K.J. 2008. Game theory of mind. *PLoS Computational Biology* 4(12):e1000254+.