Manuscript Number: ROPP62

Title: Shared representations as a coordination tool for interaction

Article Type: Joint Action: What is Shared?

Keywords: internal model, prediction, interaction, joint action, mindreading, action simulation

Corresponding Author: Dr. Giovanni Pezzulo,

Corresponding Author's Institution: National Research Council

First Author: Giovanni Pezzulo

Order of Authors: Giovanni Pezzulo

Abstract: Why is interaction so simple? This article presents a theory of interaction based on the use of shared representations as a "coordination tool". By aligning their representations (unintentionally or intentionally), interacting agents remain predictable by one other. In turn, this facilitates interaction and lowers its cognitive load by making costly processes such as action prediction and mindreading less necessary, or at least much easier from a computational viewpoint.

Suggested Reviewers:

# Shared representations as a coordination tool for interaction

September 14, 2010

**Abstract**

Why is interaction so simple? This article presents a theory of interaction based on the use of shared representations as a "coordination tool". By aligning their representations (unintentionally or intentionally), interacting agents remain predictable by one other. In turn, this facilitates interaction and lowers its cognitive load by making costly processes such as action prediction and mindreading less necessary, or at least much easier from a computational viewpoint.

*Keywords:* internal model, prediction, interaction, joint action, mindreading, action simulation

# Shared representations as a coordination tool for interaction

September 14, 2010

### Abstract

Why is interaction so simple? This article presents a theory of interaction based on the use of shared representations as a "coordination tool". By aligning their representations (unintentionally or intentionally), interacting agents remain predictable by one other. In turn, this facilitates interaction and lowers its cognitive load by making costly processes such as action prediction and mindreading less necessary, or at least much easier from a computational viewpoint.

*Keywords:* internal model, prediction, interaction, joint action, mindreading, action simulation

# 1 Introduction

Consider two agents that play the game of building towers consisting of 'bricks' of different colors. How do they coordinate their actions without previous agreements or conventions? How do they achieve their goals, being them cooperative (e.g., both want to build a tower composed of red bricks) or competitive (e.g., building a red versus a blue tower)? They can adopt different strategies: individualistic, social-aware, or interactive.

## 1.1 Individualistic strategy

First, each agent can simply perform the tower building task individually, irrespective of the actions of the other agent. One 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. So, ultimately this strategy performs poorly.

## 1.2 Social-aware strategy

An alternative, social-aware solution consists in taking into consideration the actions of the other agent into it's own plan. For instance, if an agent sees (or 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 it's own brick. This strategy represents a first form of coordination, and can give rise to emergent phenomena such as turn-taking. In cooperative scenarios and joint action (e.g., with the common goal of building *together* a tower), a sophisticated extension of this strategy consists in considering another's *(sub-)goals* in one's own plans, too, so as to help (or at least not to hinder) them, and ultimately so as to maximize the success of the joint intention and not only of one's own goals.

The social-aware strategy is potentially more efficacious than the individualistic, but it requires more cognitive effort due to the necessity to understand and predict the behavior of the other agent. The easiest way to do so consists in using observations of another's behavior, facial expressions and gesture as cues, similar to the observation of any other environmental cue. In this sense, adapting to the actions of other individuals is similar to adapting to the external environment and its dynamics. Recent research in computational motor control indicates 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 individual 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 great advantages in terms of action and movement prediction (Wilson and Knoblich, 2005), and at the same time this facilitates the recognition (and also imitation) of the goal of the action (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, facilitates coordination.

An extension of this strategy, which could be supported either by emulation or by complementary mechanisms for *rational inference* (Frith and Frith, 2006; Gergely and Csibra, 2003), consists in assuming an *intentional stance* toward the other agent, and then predicting its behavior on the basis of (an estimate of) its cognitive states, beliefs and intentions. Estimation of another's cognitive states is often referred to as *mindreading* (which we use here as a suitcase word for the recognition of the immediate action, the long-term intentions, or the underlying beliefs of the other agent). 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.

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 it's 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 provides great advantages in the interaction since it 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 predictive and mindreading abilities, too.

To summarize, the social-aware strategy makes extensive use of mindreading abilities. However, despite monkeys and humans are equipped with sophisticated action prediction and mindreading abilities (possibly of different nature), beyond simple scenarios mindreading is often an ill-posed, costly and difficult task, and this makes the computational demands of the social-aware strategy huge. Consider again the tower building task. Before deciding which brick to take, an agent has to mind-read the other agent so as (1) to recognize its intentions (including communicative intentions) for the sake of discovering possible opportunities or conflicts with it's own plans, and (2) infer the effects of it's own actions on the observer agent and its future plans, and so on indefinitely. This method is extremely expensive in terms of cognitive operations required before taking any action, but it is nevertheless the standard way game-theoretic accounts conceptualize reasoning in the social domain (e.g., in the iterated prisoner dilemma). A possible alternative (or add-on), at least in cooperative scenarios, is making heavy use of explicit communication rather than mindreading, so as to communicate one's own actions and intentions rather than letting the other 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 they use much simpler methods that do not 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 conventions, agreements, coordination artifacts or facilities for suggesting a "framing" or "scripting" of situations, which provide a solid grounding for prediction and coordination, such as for instance semaphores in crossroads, but also social conventions and scripts. This is certainly the case in some situations, but again it lets unexplained the much

[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).

more 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, which we call an *interactive strategy*, which is far less demanding of the social-aware strategy but is still goal-directed and model-based, and does not necessarily make use of prior conventions or agreements, but creates temporary versions of these coordination tools interactively (and possibly erases them after the interaction).

### 1.3 Interactive strategy

The interactive strategy simplifies the demands of interaction by implementing a simple requirement: *remaining predictable* by the other agent(s). As discussed above, a key requirement for successful interaction is being able to predict accurately the actions of the other agent (in much the same way the success of individual goal-directed agents depends on prediction of external dynamics). Additional and deeper forms of mindreading are often functional to prediction (although they are also required to fine-tune interactions, such as for instance as a prerequisite for helping actions; see later). This means that, if both agents are able to maintain their behavior predictable by the other agent, performing complex inferences is less necessary.

How is the "remaining predictable" maxim realized? In a nutshell, the core of the *interactive strategy* is forming (intentionally or unintentionally) a common ground of *shared representations (SR)* (Sebanz et al., 2006), and then exploiting them as a basis for selecting what action to take (so as that it is easily predicted), and predicting actions performed by others.

Forming shared representations could be costly in the short run, but gives great advantages in terms of coordination, since (1) it creates a (shared) structure of expectations which facilitates predictions and in most cases makes intention recognition less necessary; (2) it provides a ground for planning that makes recipient design less necessary; (3) it simplifies the encoding and decoding of *communicative* intentions in addition to standard pragmatic actions. For all these reasons, we argue that shared representations can be used as a coordination tool, similar to a blackboard, to facilitate interaction and joint action.

**Aims and structure of the article** In the next sections we elucidate the *interactive strategy* from conceptual and computational viewpoints. In Section 2 we discuss how shared representation are formed (automatically and deliberatively). In Section 3 we discuss how, once formed, they are used as a coordination tool for implementing the *interactive strategy*, and how this activity constrains communication during interaction. Then, we draw our conclusions in Section 4, and in Appendix A we offer a sketch of the interactive strategy from the computational viewpoint of Bayesian generative systems.

## 2 Formation of shared representations

The first requirement for the interactive strategy is forming a common ground of *shared representations (SR)*[2]. Shared representations can be formed at least in two ways, automatically and deliberatively. The first method consists in an automatic alignment of behavior and even representations, which is achieved through a form of mutual entrainment and emulation of behavior. The second method consists in the intentional creation of shared representations during the interaction, which can be done either by explicit signaling and communication (e.g., attracting attention toward a certain object, requesting help) or by leveraging on the general disposition of agents to extract meaning from standard performative actions (e.g., putting a brick in a certain position so as to complete the tower and, at the same time, signal that it is now your turn).

[2] Other terms with partially related meaning have been used in the literature, including shared intentions, shared beliefs, shared attention, shared common ground, shared task representations. Although we are aware that these terms denote partially different processes, for the sake of simplicity here we refer collectively to shared representations (SR) or interchangeably as a common ground, although the latter term is more typically used in language and dialogue studies.

## 2.1 Automatic alignment of behavior

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, 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. From a mechanistic perspective, entrainment can be explained by (automatic) mutual emulation and mimicking (e.g., a weak form of imitation in which the goal of the actor is not explicitly copied, but only the form) (Garrod and Pickering, 2009). In turn, mutual emulation and alignment of behavior facilitates prediction by the other agent.

## 2.2 Automatic formation of shared representations

Not only the overt behavior, but cognitive representations (beliefs and goals) can align unintentionally during interaction. It has been reported that this process produces 'team behavior' when two or more individuals are aware of one another during task performance (Knoblich and Sebanz, 2008; Sebanz et al., 2005), even when this is not required or even useful[3]. A powerful drive for this effect is that most interactions take place in the same, situated environment, and situatedness poses heavy constraints on the actions and beliefs of agents, aligning them. In addition to that, the aforementioned "resonance" mechanisms of the so-called social brain could have evolved to facilitate alignment of representations, given the adaptive advantages of collaboration.

## 2.3 Deliberate formation of shared representations

In addition to automatic mechanisms, interacting agents can implement deliberate strategies with the goals of aligning their representations, forming a common ground, monitoring and fixing it during the course of the interaction.

Some actions perform both external goals (e.g., building the tower) and common ground maintenance (e.g., signaling that I have already put the brick and it is now your turn) at the same time, and then their goals cannot be readily disentangled. Other actions are purely communicative (rather than having pragmatic goals) and are specialized for the formation and maintenance of shared representations. For instance, I can look at repeatedly to a heavy red block, or point at it, in order to attract your attention, communicate you something about it, show you my intention to lift it, and eventually to ask you help in lifting it; or, I can communicate you explicitly by using language. The explicit formation of shared representations and common ground has been studied mainly in the context of dialogue[4] (Clark, 1996), but it is an ubiquitous phenomenon in interaction of humans and other primates, and it does not necessarily make use of language (Tomasello et al., 2005).

As we will discuss below in more detail, although actions that are functional to establish or maintain a common ground trade off short-term advantages (i.e., if i point at a brick, I cannot at the same time build the tower), they bring long-term advantages in terms of success of interaction.

Before continuing, two caveats are necessary. First, the term *shared representations* does not indicate that the mental representations of the two agent are the same. It is worth noting that representations are only partially *shared*, and often the two agents have different beliefs on what parts are really shared. This

[3] We have argued elsewhere that casting action planning and recognition within a Bayesian generative architecture explains how alignment of behavior can produce alignment of internal representations (Pezzulo, sub).

[4] Dialogue is itself a joint (communicative) action, and has the same characteristics as building together a tower of bricks. 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. Galantucci (2005) has studied the emergence of communication from non-conventional rules (i.e., without a predefined language).

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 of this fact is that no common ontology is necessary. Note that, 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 (we will elaborate on that theme in Section 3.3). 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.

Second, although the term *shared representations* suggests that they are all internal cognitive representations, there can be external representations (Kirsh, 2010) as well, such as maps, diagrams to which both agents can refer, of objects that are under the attention of both agents (indeed, mechanisms for orienting social attention are part of the common ground building mechanisms).

# 3 Use of shared representations as a coordination tool

Once formed, shared representations can be used as a coordination tool to facilitate prediction during action planning and action understanding, making mindreading less necessary.

## 3.1 Using shared representations during action planning

During action planning, the first and foremost rule of the interactive strategy consists in selecting actions and goals *only* from the common ground of shared representations. Since the same information is already available to the other agent(s), this rule makes it unnecessary to ensure that it will predict and interpret them accurately. For instance, if the common goal is building a red tower and the common ground is that it is now my turn, I should simply stick to the most obvious plan, grasp a red brick and place it over the tower. In addition to that, by acting according to the common ground, I am also at the same time communicating that it is reliable, and we should stick to it.

Should I need to do something that goes beyond what we have shared up to now (say, continue the tower with blue bricks), then I need to do two things: first, performing a pragmatic action (i.e., piling a blue brick), and second, communicating the change in the common ground. Importantly, these two goals are achieved by the same method: being *un*predictable given the common ground. In other words, by performing an action that is at odds with what should be assumed by default given the common ground, or a predictable action in an odd way (e.g., with some elements that are not functional to achievement of an expected pragmatic goal), not only I advance my pragmatic objectives, but I also communicate that what we have shared up to now has to be revised. Indeed, during interaction, a powerful way to convey communicative intentions is by violating the predictions given the common ground (more on this point below, in Section 3.3).

## 3.2 Using shared representations during action observation

During action observation, the observer has to simply *predict the performer's behavior based on the common ground*, or, in other words, calculate what would be plausible given the common ground (which is already known by definition), rather than mindreading the other agent (which would imply inferential processes). This strategy works well providing that the performer agent has sticked to the aforementioned rule.

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, mindreading is exactly what the performer agent wants from the observer agent: by being unpredictable, the performer agent conveys its communicative intentions and calls for intention recognition.

## 3.3 Communication during interaction: conveying communicative intentions by violating predictions

Communicative intentions are conveyed in subtle ways during interactions. First, during joint action, every pragmatic action achieve communicative goals (i.e., informing you about the common ground) in addition to its standard pragmatic goals (i.e., contributing to building the tower). 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 actions, bodily movements and facial expressions as meaningful.

However, what is more relevant is that agents can *strategically* use their pragmatic actions to convey communicative intentions; the main way to signal that the action has indeed a communicative intention is violating the predictions that would be more natural given the shared representations. For instance, by piling a brick in an odd way, I 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). By doing somewhat unrelated to the current task, I can signal that I want to change task (or change discourse during dialogue). By piling two or three red blocks I can tell you that I want the tower to be red, or that I am in charge to pile the red blocks, and you the blue ones. The encoding and decoding of the communicative intention, in turn, depends on prior content of the shared representations; for instance, if we did not share the color of the tower, the former is the more plausible message; on the contrary, if we share knowledge that the tower must be red and blue, the latter message is more plausible. The key insight for encoding or decoding communicative intentions in this context is considering what should be added or subtracted from the currently shared representations that could explain (i.e., make predictable) the observed (communicative or pragmatic) action–and then modifying shared representations accordingly.

Note that shared representations help encoding and decoding communicative intentions, too. First, the criterion of prediction success (or failure) helps distinguishing what actions are pragmatic, and what are communicative, which is not trivial in the general case. In addition to that, encoding and decoding communicative intentions does not (in the general case) require deep recipient design and mindreading, but its meaning is already salient given the ongoing interaction, and an agent can assume by default that the 'surprising' (i.e., not predictable and hence communicative) part of the action is a message about what should be added in the shared representation.

## 3.4 Cost of communicating, and why communication remains relevant for 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 are costly during 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, this also means that communicative intentions should be (and are) by default interpreted as being relevant for the ongoing interactions and to be put in the common ground, 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 common ground[5]. The situated context of the interaction is an important cue for intention recognition, and explains why interactions facilitate intention recognition.

[5] Observers can, at the same time, infer information that serves them to revise the model of the performer agent, and not the common ground. However, we will assume that by default intentionally delivered communicative actions change the common ground, whereas actions that have informative side-effects (not willed by the performer agent) can change the observer's model of the performer agent as well.

### 3.5 Advantages of the interactive strategy

The interactive strategy drastically lowers the demands and cognitive load of successful interaction, in both cooperative and competitive scenarios[6]. The first reason is that it replaces complex inferential processes and mindreading with simpler predictions (although mindreading remains necessary in some cases, see later).

The second reason is that with this strategy it is sufficient to maintain only a model of the ongoing interaction rather than two (or multiple) models (i.e., a model of oneself and one for each other agent participating in the interaction). As we have discussed, internal models are necessary for planning coordinated actions. However, in the interactive strategy planning requires only to predict if the interaction will proceed well, assuming the ongoing shared representations (and implicitly assuming that all the participating agents will use the interactive strategy, too), independent of who will do what. Should an agent predict a failure of the interaction, then the strategy prescribes to fix the shared representations (on which the future plans will be based) by performing communicative actions. Indeed, a good quality of shared representations is a prerequisite for the success of the interactive strategy, and so both agents need to put some effort in their maintenance during all the course of the interaction (in addition to executing the standard pragmatic actions aimed at building the tower).

Put in another perspective, maintenance of the shared representation is also a form of "teaching" for the other agent: by confirming or disconfirming your predictions, I am teaching you on what "ground" (the SR) you can generate good predictions about me and the rest of the interaction. Note that this is subtly but significantly different from telling you all the content of my mental representations. For instance, during joint action, it is not necessary that I tell you or that we share my part and your part of the plan; it is sufficient that I provide you enough ground to make you able to generate good predictions at the right time (although in some cases this implies that you infer, predictively or retrospectively, my part of the plan). We will elaborate more on this point in Appendix A.

## 4 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 mind-read one another, and to do it recursively (e.g., I infer what you believe, you infer what I believe you believe, etc.) and repeatedly, 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. Unfortunately, this method could have heavy demands in terms of continuous, mutual mindreading (or also a loopy mindreading, if during my mindreading I also consider you mindreading me) 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.

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 explicit reasoning, communication and agreements. This analysis is based on the idea that for most interactions and joint actions to proceed smoothly it is sufficient that all agents select their actions on the top of a model of the ongoing interaction that affords prediction of its (successful) continuation. Since actions of all the participating agents influence the ongoing interaction, in order to use the interactive strategy is necessary that they strive to *remain predictable*. We have proposed that shared representations are interactively built and exploited for this sake. An important part of the interactive strategy is communicating when the shared representations have to be revised, which is done by intentionally violating the observer's predictions and triggering his mindreading and inferential processes.

At the same time, our analysis indicates that often communication has contrasting requirements than

[6] Why should one bother forming a common ground in competitive scenarios? The reason is that, by doing so it becomes easier to achieve one's own individualistic goals; in addition to that, maintaining a common ground channelizes the predictions of the adversary and facilitates tricking it, such as for instance feinting in soccer. See (Pezzulo, sub) for a discussion.

pragmatic action, and then it has to be used with parsimony in order not to hinder the interaction. This implies that what is really shared during interaction could be significantly less than commonly believed, and ultimately two interacting agents could have different beliefs on what parts they really share–what is really essential is that they share the minimum amount of common ground that affords mutual prediction in the course of interaction, and ultimately its success. However, note that realistic scenarios of interaction are not as unconstrained as the "tower game" (which is itself quite constrained in natural scenarios, indeed). This is the case, for instance, of interactions that range from athletes guiding together a canoe to drivers coordinating in a crossroad. In these scenarios, and others, one can safely assume that social scripts and conventions are part of the shared representation and guide the interaction, at least unless one of the agents violates them (an occasion that leads the other agent(s) to revise their own model of the violator, and the common ground).

We have also suggested that, during interaction, mindreading and associated processes are less necessary than commonly believed. However, we are not suggesting that they are not used at all. We believe that the interactive and social-aware strategies are not mutually exclusive but may act in concert. In addition, it turns out from our analysis that mindreading is necessary in at least four cases. First, for encoding or decoding communicative actions (that are typically aimed at changing shared representations, and are typically interpreted in the same way). Second, for interpreting ambiguous situations, when prediction is not straightforward[7]. 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, as we have discussed the interactive strategy focuses on shared representations irregardless of the role of the other partners. However, in some cases, especially when the interaction fails, it is necessary to derive the specific role of each agent, and *who should have done what*; in this case, mindreading can be used postdictively.

## 4.1 Implications of our view for the study of social cognition

Before concluding the article, we want to highlight possible implications of our theory and the interactive strategy for research in cognitive (and social) psychology and 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 it is doing according to rationality principles. The perspective that we have proposed is slightly different in that it suggests that most of these operations could be safely skipped in favor of simpler predictions, which in turn are functional to achieving good coordination and the planning of one's own actions.

The idea that the social brain adopts predictive mechanisms is certainly not novel (Frith and Frith, 2006; Gallese, 2009; Pezzulo and Castelfranchi, 2007, 2009; Hamilton et al., 2007; 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 the mindreading mechanism, 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 the ongoing research on social cognition and especially joint action. In a series of recent these studies, where the focus in 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

[7] More precisely, not always mindreading is required when an action is unpredictable, but only when it is also *informative*, that is, when it cannot be explained by a random effect, so the most plausible explanation is that the observed surprising behavior has a rationale.

et al., 2007, 2008). Indeed, an "interactive" encoding could be extremely useful from computational and evolutionary perspectives.

In addition to that, 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 they align automatically via mutual imitation or similar unintentional dynamics, but they are also deliberately changed according to the demands 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.

# 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 ($A$), which in turn determine the agent's overt behavior, which becomes part of the observable state of the world (i.e., $O$ out of the full, unobservable state of the world $S$)[8].

**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 $S_{t+n}$. 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 to the fast dynamics of the environment, in the cases in which feedback is too slow (Desmurget and Grafton, 2000; Kawato, 1999).

**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).

**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 dinamicity. Again, forward models can be used to predict the interactive dynamics and to adapt to them. However, remind that agents have only

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

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 more complex. 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, mindreading 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. For instance, if one can only see the observable effects of an action, and wants to infer it, the inference has the form $P(A|O)$; if one knows also the underlying intention, the inference has the form $P(A|I, O)$. 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).

**Advantages of using shared representations and the interactive strategy** Within our model, shared representations ($SR$) can be considered the aligned subset of beliefs and intentions of two (or more) agents, or in other terms the equivalence of a subset $I^{agent1}$ with a subset of $I^{agent2}$, and/or a subset of $B^{agent1}$ with a subset of $B^{agent2}$. It is worth reminding that it is not necessary (and not even true) that *all* the intentions and beliefs are shared, or that agents agree on what representations are shared (in other terms, they can have partially disjoint beliefs on what is the content of the $SR$).

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 external states of affairs $S$ (although in some cases, the latter can entail the former). These are *communicative* actions that typically change another's beliefs and intentions—not beliefs and intention in general, however, but only those relative to the $SR$.

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 *observable* by 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 3.5, 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 a model of the ongoing interaction that includes knowledge of $SR$, which predicts what will happen next (and if it is good or bad) for the success of the interaction. The individualized models are not required since the aforementioned prediction can be done independent of knowledge of who does what, and when (although in specific cases this knowledge could be required, such as for instance when only one of the participating agents is able to perform a certain action).

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 observable and 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 3.5, 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 best value of $SR$ for performing $P(A|SR)$. 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 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 $SR$ 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

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

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 Netw.*, 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 Cogn. Sci.*, 4:423–431.

Frith, C. D., Blakemore, S. J., and Wolpert, D. M. (2000). Abnormalities in the awareness and control of action. *Philos Trans R Soc Lond B Biol Sci*, 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.

Galantucci, B. (2005). An experimental study of the emergence of human communication systems. *Cognitive Science*, 29:737–767.

Gallese, V. (2009). Motor abstraction: a neuroscientific account of how action goals and intentions are mapped and understood. *Psychol Res*, 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 Cole, P. and Morgan, J. L., editors, *Syntax and semantics*, volume 3. New York: Academic Press.

Grosz, B. J. and Sidner, C. (1990). Plans for discourse. In Cohen, P. R., Morgan, J., and Pollack, M. E., editors, *Intentions in Communication.* 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., Joyce, D. W., Flanagan, J. R., Frith, C. D., and Wolpert, D. M. (2007). Kinematic cues in perceptual weight judgement and their origins in box lifting. *Psychol Res*, 71(1):13–21.

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.

Kirsh, D. (2010). Thinking with external representations. *AI & Society.*

Knoblich, G. and Sebanz, N. (2008). Evolving intentions for social interaction: from entrainment to joint action. *Philos Trans R Soc Lond B Biol Sci*, 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. *Nat Neurosci*, 10(7):817–818.

Pezzulo, G. (2008). Coordinating with the future: the anticipatory nature of representation. *Minds and Machines*, 18(2):179–225.

Pezzulo, G. (sub). Grounding the pragmatics of language in the pragmatics of non-linguistic interaction. *submitted.*

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.

Pickering, M. J. and Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behav Brain Sci*, 27(2):169–90; discussion 190–226.

Rizzolatti, G. and Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience*, 27:169–192.

Sebanz, N., Bekkering, H., and Knoblich, G. (2006). Joint action: bodies and minds moving together. *Trends Cogn Sci*, 10(2):70–76.

Sebanz, N., Knoblich, G., and Prinz, W. (2005). How two share a task: corepresenting stimulus-response mappings. *J Exp Psychol Hum Percept Perform*, 31(6):1234–1246.

Tomasello, M., Carpenter, M., Call, J., Behne, T., and Moll, H. (2005). Understanding and sharing intentions: the origins of cultural cognition. *Behav Brain Sci*, 28(5):675–91; discussion 691–735.

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.

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. *Philos Trans R Soc Lond B Biol Sci*, 358(1431):593–602.

Wolpert, D. M., Gharamani, Z., and Jordan, M. (1995). An internal model for sensorimotor integration. *Science*, 269:1179–1182.