Phenom Cogn Sci
DOI 10.1007/s11097-013-9318-y

Making our ends meet: shared intention, goal adoption and the third-person perspective

Luca Tummolini

Abstract Mind reading (i.e. the ability to infer the mental state of another agent) is taken to be the main cognitive ability required to share an intention and to collaborate. In this paper, I argue that another cognitive ability is also necessary to collaborate: representing others' and ones' own goals from a third-person perspective (other-centred or allocentric representation of goals). I argue that allocentric mind reading enables the cognitive ability of goal adoption, i.e. having the goal that another agent's achieve *p* because and as long as another agent has that goal that *p*. Having clarified the relevance of mutual goal adoption for acting jointly, I argue that when an intention is shared between several agents, each individual has an intention in favour of the joint action and one in favour of a joint mode of reasoning. This mode of reasoning is allocentric reasoning. Finally, I elaborate on the consequences of this view for the scientific study of human collaboration.

Keywords Shared intention · Goal adoption · Third-person perspective · Collaboration · Allocentric representations · Social cognition

Introduction

Being able to collaborate is one of the most important behavioural traits displayed by our species (Tomasello et al. 2005). For a long time, this aspect of human sociality has received only scant attention probably because the overall alignment of incentives that is typical of joint actions seemed to suggest that collaboration is unproblematic, at least from an evolutionary perspective. This traditional neglect is nowadays under revision. Contemporary theories of the evolution of human intelligence emphasize the significance of being able to coordinate in joint action contexts (Sterelny 2007; Moll and Tomasello 2007) and hypothesize that the human brain is especially adapted for making decisions in a social world (Frith and Frith 2010).

It is thus understandable that lately in philosophy (Searle 1990; Bratman 1993; Tuomela 1995; Sugden and Gold 2007) and in psychology (Knoblich et al. 2011),

L. Tummolini (✉)
Istituto di Scienze e Tecnologie della Cognizione del CNR, Via S. Martino della Battaglia, 44, 00185 Rome, Italy
e-mail: luca.tummolini@istc.cnr.it

Published online: 23 May 2013

there has been much interest in developing new conceptual frameworks to explore the cognitive pre-requisites for joint action and collaboration. Given that the peculiar cognitive abilities of collaboration have been only limitedly explored in the past, the interplay between philosophy and psychology on this topic is particularly relevant.

Though different theories vary a lot in their details, there is a growing consensus that in order to further our understanding of how we are able to collaborate, the ability to *share* mental states (e.g. beliefs, desires and intentions) is of paramount importance.[1] In this perspective, two agents are able to collaborate if they are able to share an intention to act together. Thus, understanding what it means to share an intention and how we can do it would explain collaboration.

In this paper, my aim is to offer a new analysis of the mental attitudes that constitute a shared intention: what the nature of a shared intention is. To do so, I first briefly summarize the relevant concept of intention, the relation between shared intention and joint action and the two constitutive roles of a shared intention. Then I identify a cognitive ability whose role in supporting joint action is often overlooked: In order to engage in joint intentional action, two agents should be able to act in order to promote a goal of another agent *because and as long as* such goal is another agent's goal.[2] This is the cognitive ability of *goal adoption* (Castelfranchi 2003). While the role of mind reading is well recognized (see Apperly 2012 for a comprehensive overview), I argue that goal adoption is enabled by a combination of mind reading and the ability to represent somebody else from a third-person perspective. Having cleared the stage, I then argue that two individuals can rationally form an intention that *they* carry out a joint action if they engage in mutual goal adoption ("Intending that we reason about how to *J* from the third-person perspective" section). However, goal adoption is also crucial to understand how agents that share an intention are prepared to independently choose means that are compatible with each other. To show that it is so, I avail myself of the game theoretical analysis of coordination problems. Applying some basic tools of bargaining theory to coordination games makes the relevance of the third-person perspective explicit. On these grounds, I argue that goal adoption plays a crucial role in enabling agents to share an intention: It is required in order to form both the intention in favour of the coordinated action and that in favour of a joint mode of reasoning. In order to assess this proposal, I then compare it with two other prominent ones (Bratman's analysis and team-reasoning approaches. I conclude by briefly elaborating on the consequences of this analysis for the scientific study of human collaboration.

[1] Exactly what we share and how we can share it is still, however, matter of debate. See for instance the various contributions in Butterfill and Sebanz (2011).

[2] For the aim of this paper, the term 'desire' and 'goal' are used interchangeably to refer to the same kind of mental state with propositional content: a desire that *p* is synonymous with a goal that *p*. To have a 'desire' or 'goal' means to be in a state with which the world must fit (see Smith 1987 for the notion of 'direction of fit' and its application to desires). While the term 'desire' is conventionally adopted in philosophy of mind and action, the term 'goal' is mostly used in psychology of motivation and neuroscience of action. For an extended discussion of the concept of goal as the prototypical conative mental state that considers desire as a sub-case, see Castelfranchi (2012).

The relation between joint action and shared intention

Not every kind of social interactions between two or more persons is usefully understood as a case of joint action. Contrast, for instance, two friends taking a walk together with two strangers who happen to walk alongside without bumping into each other. Behaviourally, these two cases might look similar. But, at a closer inspection, one can notice that only in the former situation their *reciprocal coordination is intended* by the participants, and this intentional fact makes their coordinated movements a case of *joint action*. Alternatively one can say that when two friends take a walk together, they *share* an intention to do it, while two pedestrians walking alongside to each other do not seem to share a similar attitude towards their coordinated action. Still, an appeal to a shared intention to explain joint action is vacuous, if an analysis of what it takes to share an intention is not offered.

There are indeed many competing theories of shared or collective intentionality (see for instance Gilbert 1990 and Bratman 2009a for two alternative proposals that start from the previous paradigmatic example). However, in this paper, I will only focus on the approach developed by Michael Bratman (1993, 2007, 2009a) for three main reasons.

The first reason is pragmatic: Besides its philosophical merits, Bratman's theory has been also useful to organize empirical data in developmental and comparative psychology (Tomasello et al. 2005; Warneken et al. 2006) and to formulate new comprehensive theories amenable to be empirically tested (Pacherie 2008, 2012), thus fostering the interdisciplinary dialogue between philosophy and psychology. The second is that, besides action, Bratman's framework takes into account aspects of reasoning and deliberation that are often neglected by those who are interested in the cognitive foundations of joint action but, as I will argue, are crucial to properly analyze collaboration.[3] Finally, Bratman's characterization of shared intention is grounded on his planning theory of intention (Bratman 1987), an approach that will be endorsed as a background in this contribution.

The planning theory of intention

To clarify the conceptual tools that frame the discussion that follows, it is useful to provide a short summary of the main tenets of the planning theory of intention (Bratman 1987). Bratman's theory of intention is a variety of functionalism (Lewis 1972): An intention is identified with a kind of mental state that plays a particular set of *roles* in a cognitive system. Specifically, in this theory, an intention is crucially linked to our *ability to pursue goals that*, *in order to be achieved*, *require the temporal coordination of several actions* (Bratman 1987). This temporal dimension is crucial to understand intentions,

[3] The importance of modes of reasoning to understand shared or collective intentions is also characteristic of team-reasoning approaches to joint action; see for instance Sugden and Gold 2007 and "Beyond Bratman's semantic strategy" section below.

which are viewed as embedded in larger future-directed plans with specific *guiding*, *coordinating* and *organizing* roles in our practical reasoning and action.[4]

According to the planning theory, associated with these roles, there are characteristic norms of instrumental rationality that are implicitly accepted by an agent. Typical norms are those of consistency, agglomeration, means-end coherence and stability. For instance, intentions are supposed to be internally consistent and consistent with one's beliefs. An agent should be able to agglomerate her different intentions into a larger one that is consistent in these ways. Moreover, having an intention demands that the agent settle on relevant means when appropriate. Finally, an intention should be stable over time and should involve resistance to reconsideration and change (see Bratman 1987, 2009b for a thorough defence of these norms).

The appeal to these roles and norms allows discriminating intentions from other mental states like beliefs and desires or goals. Ordinary goals are not subject, for instance, to norms of consistency and means-end coherence: One can have several goals that are inconsistent with each other in the sense that they cannot be realized at the same time. Moreover, simply having a goal is not by itself enough for settling on a means to achieve it. Moreover, the roles and norms characteristic of intentions fit with each other in the sense that a general conformity to the norms of consistency, agglomeration, means-end coherence and stability contributes to how the guiding, coordinating and organizing roles of intentions are fulfilled.

Grounding shared intentions on personal intentions

Such an approach to the nature of intention has been employed to also understand how an intention can be *shared* between two or more agents, e.g. from Alice's *personal* intention to take a walk to Alice's and Bob's *shared* intention to take a walk together. Like in the individual case, an intention that is *shared* between (at least) two agents is supposed to be a state that plays the same *coordinating*, *guiding* and *structuring* roles in their *joint* action and reasoning. In particular, according to Bratman, the characteristic roles of a shared intention to do a joint action *J* include the *interpersonal coordination of action and planning in the pursuit of J* and the *structuring of related bargaining and shared deliberation concerning how to J* (see Bratman 1993).

Again, as in the case of personal intentions, these (social) roles are associated with such (social) norms as social agglomeration and consistency, social coherence and social stability (Bratman 2009a). That is, agents sharing more than one intention should be able to agglomerate them into a larger social plan that is consistent, that specifies relevant means in a timely way and that is stable over time in a way that resists to constant reconsideration. Agents who comply with these norms will share

[4] Thus, the only of kind of intention that will be relevant for the aim of this paper is the one that Pacherie (2008) has called *distal intention* (pp. 182–184). Pacherie usefully distinguishes between distal, proximate and motor intentions. Distal intentions are relevant for the rational guidance and control of action, proximate intention for situational guidance and control and motor intention for the motor ones (p. 188). I assume that a similar distinction is appropriate for shared intention as well. Hence, the present discussion will only concern shared *distal* intention.

an intention that plays the distinctive roles of coordinating, guiding and structuring their reciprocal planning and acting.

Bratman's peculiar strategy, at this point, is to understand the nature of shared intention by relating the roles and norms at the level of personal intentions with the *social* roles and *social* norms at the level of the agents' shared agency. In this perspective, a shared intention is reduced to *a web of interdependent and interlocked personal intentions of the individual participants*. In particular, Bratman argues that such interdependence and interlocking is achieved through the *contents* of the participant's personal intentions. In other words, each participant's personal intention refers to the joint action, to the way the joint action is supposed to unfold and to the way that sub-plans of each of the participants should combine (see "Beyond Bratman's semantic strategy" section below for a critical appraisal of this strategy). When this structure of personal intentions in a context of common knowledge functions properly (i.e. as specified by the planning theory of intention), it realizes the characteristic social roles of an intention that is shared between the agents. Thus, a shared intention is identified with whatever plays these social roles, i.e. *with the set of personal intentions of the individual participants*, *the relation between these intentions and the relevant beliefs* (see for an extended argument in favour of this approach Bratman 1993, 2007, 2009a, c).

Since the planning theory of personal and shared intentions exemplifies a functionalist approach, it is left open whether the peculiar structure of personal intentions that, according to Bratman, constitutes a shared intention is the only possible realization of a shared intention or whether it can be realized in a different way (see Bratman 1999, p. 144). Indeed, an alternative (and more transparent) realization of a shared intention is precisely what this paper aims to offer.

The enabling role of goal adoption as a cognitive ability

Given the role that knowledge of others' beliefs and intentions plays in Bratman's account, his theory clearly presupposes that agents sharing an intention are equipped with a sophisticated level of mind reading ability. Bracketing for the moment possible critiques of requiring such cognitive sophistication to explain joint action (see for instance Tollefsen 2005, Pacherie 2011 and "Beyond Bratman's semantic strategy" section below), here I want to suggest that, besides standard mind reading, two agents sharing an intention to do a joint action J (e.g. paint a house together) should display the more general ability of *goal adoption*.

By goal adoption, I mean the ability to form the goal that another agent's goal is achieved *because and as long as* it is the goal of that other agent (Conte and Castelfranchi 1995, pp. 44–56; Castelfranchi 2003). More precisely, goal adoption is a *mental act*: It is the formation of a conditional goal, that is, the goal that another agent achieve p is *conditional* on the belief that the other agent has the goal that p.[5] Being a mental act, goal

[5] I will use the expression 'adopted goal' to refer to the goals that an agent form when engaging in goal adoption.

adoption is a mental event analogous to judgment or decision; if a judgment is the mental act of forming a belief and a decision is the mental act of forming an intention, goal adoption is the mental act of forming a goal that another agent's goal is achieved.

Goal adoption is the terminal stage of a practical reasoning process in which, for instance, if an agent has the goal that *q* and believes that he will not achieve it unless he adopts another agent's goal that *p*, then he will form the goal that the other agent achieves *p*. This particular case of practical reasoning will be named 'adoptive reasoning'.

Importantly, by adopting another agent's goal, the adopting agent forms a goal with a characteristic *social content* (i.e. that another agent achieves his own goal). Moreover, goal adoption is also a basic process to come to have goals with a *social origin* (i.e. another agent). Defined in this way, goal adoption is clearly different from imitation in which one inherits a new goal from another agent: In imitation, for instance, one is not *motivated* by the goal that another agent achieve something.

Moreover, and more importantly, even if adoptive reasoning is a kind of practical reasoning that is crucially socially oriented, it does not necessarily imply any benevolence or altruism with respect to the other agent. For instance, a typical non-altruistic goal adoption underlies the exchange of goods in the marketplace. A simple act of trade between two agents exemplifies reciprocal goal adoption: For Alice to obtain something she wants from Bob the butcher, she has to adopt Bob's goal to get a proportionate amount of money and act in order that Bob achieve it, for instance, by paying, similarly for Bob who has to adopt Alice's goal to get the meat and has to act in view of its fulfilment, i.e. by giving the meat to her (Conte and Castelfranchi 1995, p. 51; Sugden 2011).

In general, one can distinguish two different reasons for adopting somebody else's goal. In the case of market exchange, goal adoption might be simply *instrumental*: If Alice's goal is to get the meat and she believes that unless she adopts Bob's goal to have some money in return, she will not get it, Alice has an instrumental reason to adopt Bob's goal.[6] Goal adoption, however, can also be based on altruistic motivations. Suppose Alice is benevolent with regard to her friend Bob, meaning that she is motivated by a general attitude towards Bob: She wants Bob to obtain what he wants in a terminal way (i.e. this goal of Alice's is not instrumental to other goals of hers). If she decides to help him with some of his projects, she will engage in *benevolent* goal adoption, i.e. goal adoption for purely altruistic reasons.

In both cases, the kind of social orientation that is implicit in successful goal adoption is a form of de-centred or other-centred cognition. Thus, the relevant opposition is not between self-*regarding* (egoistic) and other-

[6] Even if exchanging goods is not motivated by altruism, this does not mean that the underlying motivation should be necessarily selfish, in that it can be driven by a motivation of mutual advantage. On the fact the market exchange might also be construed as a form of collaboration, see Sugden (2009). For an early analysis of exchange and goal adoption, see Castelfranchi and Parisi (1984).

regarding (altruistic) motivations in which one cares about another agent's welfare,[7] but between self-*centred* (egocentric) and other-*centred* (allocentric) ones in which one comes to have a goal *from the third-person perspective of somebody else* be that either for instrumental or for benevolent reasons.[8]

Even if exchanging goods does not require altruistic motivations, it does require goal adoption: the ability to see a choice from a perspective that is different from one's own, to understand what option someone else values and to instrumentally act in view of such knowledge in order to fulfil someone else's goals. *The adopted goal is an other-centred goal that is a goal that one entertains from an allocentric perspective.*

In order to see the relevance of this point more clearly, consider the difference between a situation in which Alice has the key to a door, wants to go out and thus forms the goal to open the door and a situation in which Alice still has the key and Bob is in front of the door. In this latter scenario, Alice might instrumentally adopt Bob's goal to exit and open the door in order to let him exit just because she cannot stand his presence in the house. Though the content of the goal is the same in the two scenarios (i.e. that the door is opened), in the former case Alice's has the goal from her own egocentric perspective, while in the latter Alice's goal is held from the allocentric perspective of Bob's and she acts in order that Bob's achieve it. Thus, even such a trivial case of instrumental goal adoption demands that Alice smoothly *switches* between an egocentric and allocentric perspective in order to get what she wants.

The importance of goal adoption and the cognitive sophistication it requires in terms of other-centred representations, is, unfortunately, widely underestimated in current debates on cooperation and collaboration. Indeed, it is striking that, humans aside, the ability to intentionally promote a goal of another animal, even when doing so is instrumental to create the conditions to promote one's own goals, is a behaviour that seems not to be available to other animals; chimpanzees, for instance, despite their quite sophisticated mind reading abilities, are remarkably incapable of instrumental helping (see for instance Warneken and Tomasello 2006 and "Conclusion" section below).

Mutual goal adoption enables the intention that *we* do the joint action

Consider now a complex goal like that of having a whole house painted, an outcome for which two agents, say Alice and Bob, knowingly depend on each other. As we have seen, the constitutive role of a shared intention is to enable

[7] The key aspect of an other-regarding preference is that 'one's evaluation of a state depends on how it is experienced by others' (Bowles 2004, p. 109). This aspect makes social or other-regarding preferences the standard model of altruistic motivations.

[8] An allocentric representation in social cognition is the adoption of a third-person perspective when representing somebody else instead of a first-person egocentric and self-related one (Frith and de Vignemont 2005; Frischen et al. 2009). The egocentric representation of another's goal is the representation of such goal and of the means to achieve it *in a way that is relevant for oneself* (i.e. that satisfies one's own goals). The allocentric representation of another's goal is the representation of such goal and of the means to achieve it *in a way that is independent from oneself*.

the agents to achieve this kind of goal since having a whole house painted requires a temporally extended joint action. In order to successfully perform a joint action, the interpersonal coordination of action and planning of several agents and the structuring of possible bargaining and deliberation concerning how to do it should be ensured.

Suppose that, given their mutual dependence with respect to the common goal that the house is painted, both Alice and Bob independently formulate the goal *to paint the house together*. Having the goal to do a joint action *J* is not, however, the same as having decided to pursue it (i.e. forming the intention) because one cannot intend something that is beyond one's personal control, and an intention that *we* paint the house seems to be such a case (see on this point Velleman 1997).

Consider, however, that Alice has the goal that she and Bob paint the house together and believes that unless Bob have that same goal, the house will not be painted (i.e. they are mutually dependent). Alice will form the goal that Bob too has the goal that they paint the house together. If Bob is aware of this goal, he might decide to engage in adoptive reasoning towards Alice. This means that Bob can independently realize that unless he adopts Alice's goal that they paint their house together, he will not have the whole house painted. If Bob engages in instrumental goal adoption, he will form the goal that Alice's goal is achieved *because and as long as* it is her goal. Thus, assuming that the same is true of both, they might also *adopt* the goal that we do the joint action from each other for purely instrumental reasons, e.g. Bob adopts Alice's goal that they paint the house together only because doing so is instrumental to his own goal to have the house painted and vice versa for Alice with respect to Bob. Thus, even if each of them might originally have independent reasons to have the goal of doing the joint action, reciprocal goal adoption would offer additional support to such personal goals: that the other's goal to do the joint action is fulfilled. Reciprocally adopting their goals to do the joint action creates a new structure of *interdependence* between such goals, which are now entertained also *because and as long as* the other one has it: Their *adopted* goals to do the joint action are conditional on their reciprocal beliefs that each of them has that goal.[9]

As a consequence, once the participants commonly know that each has the goal to do the joint action also because and as long as the other has it (i.e. their reciprocal goal adoption is commonly known), each participant can decide to pursue it and thus rationally form *a personal intention in favour of the joint activity*. This is now rationally possible because each individual intention to do

[9] Interlocking intentions or interdependence between the intentions that we *J* is also a core feature of Bratman's analysis (Bratman 2009a, pp. 159, 161). In my strategy, intentions are interlocking in virtue of the interdependence between the *adopted* goals that we *J*. Since each agent adopts the other's goal, each goal is conditional on the belief that the other has that goal. If one were to revise this belief, one would not have the adopted goal. As a consequence, one could not rationally intend that we *J*. Bratman, on the other hand, postulates that each agent has an additional intention that *the joint action go in part by way of the relevant intention of each of the participants*. This semantic strategy is critically assessed below.

the joint action leads to *their* jointly acting in part also by *supporting the intention of the other one*.[10]

A personal *intention* that we paint a house together is, hence, rationally possible thanks to the adoption of each other's *goal* that we paint the house. Once such an intention is formed, it will play the distinctive guiding, coordinating and organizing roles specified by the planning theory of intentions. If each participant has an *individual* intention-like commitment towards the *joint* activity, the joint action will be an element in the participants' *individual* plans. The norm of consistency requires that the personal intentions of the agents be internally consistent and consistent with their beliefs. The norm of means-end coherence mandates that, individually, each agent should settle on relevant means for the joint action (e.g. aim at the coordinated use of some paint in order to paint the house together), and such personal plans should be consistent with the joint action. Each participant's personal intention in favour of the joint action should resist reconsideration. That is, each individual agent will be *responsive* to the *joint* action and not simply to their *individual* shares in it.

How do we make the right choice in a coordination problem?

Such individual responsiveness to the joint action is not, however, enough to fully support the interpersonal coordination of action and planning in the pursuit of *J* and the relevant bargaining and deliberation regarding the most appropriate means. In particular, the intentions in favour of the joint activity do not ensure *social agglomeration* since the sub-plans that each may intend to select in order to *J* may be incompatible with each other.

To see exactly why it is so, suppose, for instance, that both Alice and Bob have already formed an intention to paint a house together (i.e. an intention that we *J*). At some point, Alice and Bob should settle on relevant *means* to paint it, e.g. choose a specific colour with which the house should be painted.

The interactive decision problem that Alice and Bob would face can be described as follows. To make it simple, assume that they already are in a situation of common knowledge of each other's preference ranking. They both already know (and know that the other knows and so on) that each of them likes blue houses. The only other available possibility is to use a yellow colour. But they both know (and know that they both know and so on) that they both prefer living in a blue house than living in a

[10] My reply to Velleman's objection is close in spirit to the one offered by Bratman (1999) himself (p. 154). Highlighting the role of instrumental goal adoption, however, has several advantages. In order to rebut Velleman's objection, Bratman assumes a 'kind soul' condition. According to Bratman, you can individually form an intention to do a joint action, on the assumption that the other is 'kind soul'. That is, when the other fellow recognizes your intention that we *J*, he will come to have a corresponding intention that we *J*. This implies that knowledge of your intention that we *J* is a reason for the other to form a similar intention when he has altruistic motives (the 'kindness' of the soul). This assumption, however, is both not necessary and not sufficient. It is not necessary because it is evident that we can share intentions also for instrumental reasons (the joint option is in the self-interest of both and we know it). It is not sufficient because it does not discriminate cases in which an agent is completely self-centred and ignores the other in pursuing the joint action, provided that the other is altruistic. Egocentricity in the pursuit of a joint action is not acceptable when one is collaborating. See below section for a defence of this claim.

		Alice	
		Paint Blue	Paint Yellow
Bob	Paint Blue	2,2	0,0
	Paint Yellow	0,0	1,1

Fig. 1 The choose-the-colour game

yellow house. Finally, it is also common knowledge between them that the worst possible situation is to have half of the house painted in blue and the other half painted in yellow.

The matrix in Fig. 1 shows Alice's and Bob's reciprocal ranking of possible outcomes, i.e. how the world would be after their choices. The best outcome for both is the one in which they both use the blue paint. An acceptable, but less preferred one, is when both use the yellow paint. What would be really annoying is mixing the colours. This is the 'choose-the-colour' game.

Selecting an equilibrium in the choose-the-colour game

The choose-the-colour game is an instance of a class of strategic interactions that are known as *common-interest games*. These are situations in which the agents (or players) have a common interest in acting in a coordinated way. Common interest games represent the prototypical case of interactive decisions in a joint action context.

As it is exemplified in the choose-the-colour game, the agents have a symmetrical (or at least compatible) preference ranking such that the strategy that is favoured by one of them (what is best to do, given what the other would do) is the same strategy preferred by the other. This means that the strategies favoured by both—their strategy profile—are also a Nash equilibrium.[11]

We can safely assume that in common interest games, each player has an intention that we coordinate; however, coordination is at risk because there are *multiple equilibria*. Since coordination in such situations is not guaranteed, players need to solve a 'coordination problem' (Schelling 1960).

Consider for example a coordination game like in Fig. 2. Players 1 and 2 are playing the game of matching their sides of pennies. If they both show the head side of their coin at the same time, they win. The same is true, if they both turn the tail side of their coin. However, each of them loses if they fail to match their pennies, or, in other words, if they fail to coordinate. So if the agents intend to coordinate with each other, what should they do in this situation?

If player 1 thinks that player 2 will choose head, then the best reply is to play head too. At the same time, if player 2 thinks that player 1 will choose head, then, again, the best reply is to play head. So both expectations are correct and are confirmed by best-replying agents: Head–Head is a Nash equilibrium. However, the same is true of the Tail–Tail profile, which is for the same reasons an alternative Nash equilibrium of

[11] A strategy profile is a Nash equilibrium if given what the other can do, the best response of one agent is the same response the other would choose adopting identical reasoning towards him or her. For the limited aims of this paper, it is enough to consider only Nash equilibria in pure strategies. For the distinction between Nash equilibrium in pure and mixed strategies, see Osborne and Rubinstein (1994).

		Player 1	
		Head	Tail
Player 2	Head	1,1	0,0
	Tail	0,0	1,1

Fig. 2 The game of matching tail and head

this game. Thus, the logic of Nash equilibrium—beliefs and instrumental rationality are conducive of facts that match the beliefs—is insufficient to suggest a definite course of action in these situations.

The choose-the-colour game that has been introduced above is actually a variant of this game and, in its more abstract version, is known as the *Hi-Lo* game.[12] The only difference with a coordination game in which there is complete indifference between alternative Nash equilibria is that in the Hi-Lo game, one equilibrium is Pareto superior to the others.[13] That is, in the choose-the-colour game, both Alice and Bob are better off when they both choose the blue paint than when they both choose the yellow one.

However, if being in a coordination problem can in fact threaten coordination when there is indifference between equilibria, it seems strongly intuitive that—even when the participants cannot communicate with each other—the only 'right' solution of this game is to choose the outcome that is most preferred by both (e.g. the outcome in which both use the blue paint). However, the logic of Nash equilibrium considers the profile of strategies in which both paint a yellow house just as good as that in which both paint a blue house. Since both are Nash equilibria, the Nash equilibrium solution concept does not predict what the agents in this game are going to do. By the same logic, even in a Hi-Lo situation, the fact that both Alice and Bob might have the intention that they coordinate over some relevant means (i.e. the intention that we *J*) does not by itself suggest any specific course of action.

The behavioural and judgmental facts of the Hi-Lo game

While the choose-the-colour game seems problematic when viewed through the lens of game theory, this kind of social interaction is not puzzling at all in our everyday life. Discussing the Hi-Lo game in general, Bacharach (2006, p. 42) has pointed out two generally accepted facts about this game: the *behavioural* and the *judgmental* facts.

The behavioural fact is the idea that real people, when facing a social interaction of the Hi-Lo kind, almost always have no difficulties in choosing the outcome that is best for all (i.e. they easily coordinate on the Hi-Hi equilibrium and systematically ignore the Low-Low one).[14] The judgmental fact, on the other hand, refers to the widespread intuition that the Hi-Hi profile of strategies is the obvious 'right' choice for agents in this situation even when the interaction does not allow for repetition.

[12] See Hodgson (1967), Sugden (1993) and Bacharach (2006).

[13] The fact that one equilibrium is Pareto superior means that at least one agent is better off in that outcome, while the other agent scores at least as good as in the other equilibrium.

[14] This intuitive result has also been experimentally verified; see Bardsley et al. (2010).

Thus, the Hi-Lo game as such has been considered mainly as a problem for the descriptive and normative adequacy of the game-theoretical framework itself, which should be designed in a way that delivered the choice of the Hi-Hi outcome both as an empirical prediction and as a normative judgement. In the next sections, my aim is to offer an explanation of both facts in order to make the cognitive ability required to solve this problem explicit.

Beyond personal preferences: empathetic preferences

The standard treatment of preferences in game theory is limited to the *personal* preferences of the individual decision makers. Adopting the revealed preference approach, an agent's personal preferences are discovered by observing the choices the agent makes when solving one-person decision problems. Moreover, by observing the choices the agent makes between *risky* prospects, it is possible to represent such preferences in terms of numerical utilities that also measure *how much* an outcome is preferred over another one (for the standard procedure, see Von Neumann and Morgenstern 1944).[15]

An agent, however, may have preferences besides her personal ones. Extending the standard view, Harsanyi (1977) and (Binmore 1994, 2005) have introduced the idea that, in some contexts, one can also infer *empathetic preferences* from the agent's choices and that these preferences should be kept apart from an agent's personal ones.[16]

In the normative project of Harsanyi, the peculiar contexts that may require an appeal to empathetic preferences are those in which an agent wishes to make a moral judgement (i.e. to answer the question 'what is the right thing to do?'). Making this judgement, it is suggested, requires one to take an impartial viewpoint.

Following this line of thought, but turning the project into a descriptive one, Binmore has contended that these special contexts are those in which *we have to solve coordination problems* but we are unable to discuss face to face on what to do—thus cases analogous to the choose-the-colour game discussed above. In these situations, Binmore suggests, we may act according to a 'do-as-you-would-be-done' principle (Binmore 2005, pp. 17, 130). The way this principle works in practice presupposes that agents facing a coordination problem may *independently* select an equilibrium on which to converge by *empathizing* with their fellows, that is, by putting themselves in the position of another agent to see a choice from this different point of view. In this view, acting according to the do-as-you-would-be-done principle would reveal empathetic preferences.

While standard personal preferences are defined over outcomes, empathetic preferences differ because they are defined over pairs (i, x) where i is a person and x is a social situation (Sugden 2001, p. F227). The idea is that a given individual k in a social situation x can have a preference between 'being the individual i in social situation y' and 'being the individual j in social situation z'. Since one cannot choose

[15] In the revealed preference interpretation, preferences and utilities are considered as descriptive concepts of what an agent would do when facing certain decision problems. In this view, these constructs are not used to explain a choice or to refer to what causes an agent to act in a certain way. For a different interpretation of preferences and utility, see the discussion in Sugden (1991).

[16] The expression 'empathetic preferences' is Binmore's (1994: 56–61). Harsanyi (1977) names the same construct 'extended preferences' (p. 53).

to be another agent, these alternatives are to be considered as *imaginary* ones and manifest the ability of an agent to conceive in his or her imagination what it would be to see a choice from another agent's point of view and assessing an outcome in light of another agent's preferences.

Harsanyi (1977, pp. 54–55) has shown that, assuming that such preferences are consistent and that the agents are very good at empathizing, thereby knowing each other's real preferences, it is possible to represent such preferences in a single *empathetic utility function*. The only additional assumption that is needed to obtain this result is that, in his or her imagination, an agent considers that it is equiprobable that, after a choice is made, one turns out to be either oneself or the other person in the chosen state. Though this is again only an imagined possibility, it is taken to express the *impartiality* of the point of view.

The empathetic utility function describes how an agent *compares* the utility that she and the other one gets from different outcomes and combines them in an overall assessment of outcomes that is simply the average of the empathetic utilities of all. Thus, an empathetic utility function implicitly determines a standard for making interpersonal comparison of utility: It determines a rate at which the utility units of one are traded against those of the other. In order to compare the utilities of different agents, each agent must be associated with a *weight* that specifies the worth of an outcome for each one (Binmore 2005: 28–29).[17] The weight of the agents may vary in different contexts and could take into account for example the effort, talent and social status of each agent. However, in the simplest cases of sociality that are relevant for this paper, the agents engage in small-scale egalitarian joint actions in the absence of asymmetric authority relations (see Bratman 2009a; Pacherie 2012). Moreover, we can assume that they bring the same talent and effort to the joint project. This egalitarian aspect can be expressed by the fact that they have equal weight (e.g. 1:1). Given this simplification, when making a judgement according to the do-as-you-would-be-done principle and thus acting according to their empathetic utility functions in the context of a joint action, the agents are independently seeking to maximize *the sum of their reciprocal weighted personal utilities*.

Allocentric utility function and the impartial viewpoint

As we have seen, both Harsanyi and Binmore attach great explanatory importance to the ability of empathy. Here I want to suggest empathy suggests a role of emotions that is not necessary while that the ability that is required is a more general form of mind reading: *allocentric mind reading*.

According to common usage in psychology (Hoffman 1978; Meltzoff 2002), empathy is primarily an emotional response that is caused by someone else's experiencing a similar emotion. Empathy is different from the phenomenon of emotional contagion in that an agent who empathizes with someone else's emotional state is ascribing the emotion one is feeling *to* someone else. For instance, suppose that Alice sees that Bob is in pain and as a consequence she starts experiencing some pain herself. If Alice is empathizing with Bob's pain, she feels an emotional response that is similar to the one Bob is feeling, and she experiences it because of Bob.

[17] Binmore names this weight the 'social index' of the agent.

Moreover, even though she is feeling something similar to what Bob is feeling, she does not mistake this emotion as a feeling of hers but she ascribes it to Bob: Alice knows that she is feeling pain because of Bob but that pain is Bob's pain and not hers.[18] If one were not to acknowledge the special role of emotion in empathy, one could not distinguish empathy from the more general ability of mind reading in which an agent comes to have a mental state (e.g. a belief) because another agent has a similar mental state while ascribing that mental state to the other.

The important point of empathy, though, is that it is an *other-centred* or *allocentric* process (Hoffman 1978; Meltzoff 2002): When Alice empathizes with Bob's pain, she feels an emotion from the third-person perspective of Bob: She knows that *this* is how Bob is feeling. But since the emotional component of empathy does not play any role in Harsanyi's and Binmore's theories, we can dispense with empathy altogether. What is really implied by an agent acting according to the do-as-you-would-be-done principle is the ability to engage in *mind reading from an other-centred or allocentric stance* and not from the standard self-centred or egocentric perspective. What we have to assume that the agents are very good at is switching to *allocentric mind reading* in the appropriate contexts. Assuming that an agent is a perfect allocentric mind reader means that she is able to understand the other's goals without error and represent them from a third-person or allocentric perspective.

Moreover, an allocentric mind reader can decide to *adopt* such goals and thus form the further goal that the other achieve them. If we also assume that an agent can switch to an allocentric stance also towards her own personal goals,[19] when deciding what to do, such an agent would trade her own personal allocentric goals against those adopted from the other one. The *allocentric utility function* and the corresponding *allocentric preferences* are a way to formally represent this possibility.[20]

In sum, an agent who acts according to her allocentric preferences is making an *impartial choice*, which is nothing but a choice taken from a third-person perspective in which one balance one's own and the other's goals from an impartial viewpoint. The impartial viewpoint is just the third-person or allocentric one.

Bargaining from an impartial viewpoint in simulation

In order to appreciate the role of allocentric mind reading and allocentric preferences in finding a solution in coordination problems, however, it is useful to understand what would happen if the agents were able to bargain face to face on what is the outcome to pursue. If we restrict our attention to the simplest possible case of the choose-the-colour game, the analysis is straightforward. In this trivial game, the set of

[18] For extended discussion of empathy, see Vignemont and Singer (2006) and Vignemont and Jacob (2012).

[19] Frith and de Vignemont (2005) indeed distinguish between two attitudes towards the self: egocentric representations of the self that derive from direct knowledge attached to the self in the first-person perspective and allocentric representations of the self that derive from detached knowledge of the person one happens to be as if one was looking at oneself from a third-person perspective (p. 725). Here I am considering only the special case of personal goals that can be either egocentrically or allocentrically represented.

[20] Thus, instead of the expression 'empathetic' preferences and 'empathetic' utility function from now on I will use the expression 'allocentric' preferences and 'allocentric' utility function. Despite the terminological difference, I am referring to the same phenomenon discussed in the previous section.

possible agreements is represented by the outcome in which Alice and Bob have a blue house (with payoffs of 2:2) and that in which they have a yellow one (with payoffs of 1:1). All the outcomes in which they do not use the same colour can be considered as their disagreement point. The game theoretical analysis of bargaining predicts that, when two agents negotiate face to face, having common knowledge of their personal characteristics and of the nature of the bargaining problem, they will agree on the *Nash bargaining solution*. The Nash bargaining solution is the outcome in which the product of the players' gains over their disagreement payoffs is largest.[21] In the choose-the-colour game, this is the outcome in which both Alice and Bob paint a blue house.

Granted this, however, the choose-the-colour game has been introduced to model the social situation in which Alice and Bob have to decide what to do *without* having discussed this matter beforehand. In this situation, they may ask themselves the question 'what is the right thing to do?' separately. Thus, if both are good at allocentric mind reading, both are able to see their choice also from the perspective of the other one, and so they will exploit their allocentric preferences to solve this problem. By acting according to the do-as-would-be-done principle, each will individually compare their allocentric utilities and make a decision that maximizes the sum of the weighted personal utilities of both.[22] This means that both will choose to use the blue paint, which corresponds, in the trivial case of the choose-the-colour game, like in all the garden variety of Hi-Lo games, to the Nash bargaining solution of the corresponding bargaining game. That is, applying their ability to switch to allocentric mind reading and their allocentric preferences in this context entails that they coordinate on the deal that they *would* agree if they were to bargain in ignorance of their reciprocal identities (Binmore 2005).

Explaining the behavioural and judgmental facts in the Hi-Lo game

As discussed above, even if standard game theory is unable to predict and prescribe the choice of Hi in one-shot Hi-Lo problems, people almost always choose Hi in real decisions (the behavioural fact) and are ready to acknowledge that Hi is the 'right' choice in this context (the judgmental fact). The account that I have sketched suggests that choosing Hi in the Hi-Lo game is explained as the output of a *simulated* process of bargaining. The Pareto superior outcome is the obvious deal that the agents would choose if they were to discuss the matter face to face. Understanding this, however, requires that one is able to switch to an allocentric perspective in order to see a choice from the third-person viewpoint and to take into account the goals of another one by adopting them. The limiting case of the Hi-Lo game makes this choice trivial due to the perfect alignment between the agents' interests.[23] The pretended ignorance of each other identities when computing the possible deal is a way

[21] See Chapter 7 of Osborne and Rubinstein (1994) for an introduction to bargaining games.

[22] This process works only if we assume that the agents use the *same* standard for interpersonal comparison of utility. In the context of an egalitarian joint action and assuming that both agents bring the same talent and effort to the joint project, I have suggested before that this implies that both agents weight their reciprocal utilities equally (the weight is 1:1) in their allocentric utility functions.

[23] However, in more complex situations in which a level of conflict between the personal preferences is introduced, the same mechanism can ease coordination on some form of compromise; see Binmore (2005) for an extended discussion of these more interesting situations.

to model, in these situations, the allocentric perspective. Thus, the appeal to allocentric mind reading and goal adoption, and to non-standard preferences like allocentric preferences, is offered as a descriptive account of the behavioural fact: This cognitive ability is involved in the actual choices of Hi in real contexts.

The same approach, moreover, offers an explanation of the judgmental fact. Actually, the framework of allocentric preferences has been developed by Harsanyi (1977) precisely to model the normative judgment of agents in similar situations. The emphasis on morality both in Harsanyi (1977) and Binmore (2005) is due to the fact that similar problems also arise when coordinating on multiple equilibria in which the interests of the agents are not perfectly aligned. If making a judgment according to the do-as-you-would-be-done principle in a Hi-Lo context does not require any personal sacrifice, in other contexts it may correspond to an appeal to a *fairness* criterion. In the Hi-Lo case, choosing Hi is the *right* choice because it the choice that conforms to this principle and, at the same time, is the one that, when viewed from a third-person perspective, is the best choice for everyone.

Intending that we reason about how to *J* from the third-person perspective

Let us take stock. I have argued that, besides generic mind reading that is clearly needed to infer *what* goals and beliefs each agent has, two agents sharing an intention to do a joint action *J* (e.g. paint a house together) need to switch to an allocentric representation of each other. Moreover, the agents facing a coordination problem should properly exploit their *allocentric utility functions*, that is, they should choose by also taking into account how each option is valued by their co-actors *from a third-person perspective*. When seen in this way, coordination on an outcome is a form of simulated bargaining which parallels the way in which the agents would bargain face to face if they had the opportunity. Thus, the interpersonal coordination and the structuring of bargaining and shared deliberation that are constitutive roles of shared intention are strictly linked to the allocentric abilities and preferences of the agents. Finally, since preferences are used here to *describe* the observable choices of the agents (i.e. preferences are revealed by choices), the allocentric utility function that trades one's own allocentric utilities against those of another agent presupposes the ability of *goal adoption*.

Thus, the two constitutive roles of a shared intention imply that the agents display mutual goal adoption also regarding the choice of the *means* that are relevant for the joint action. In other words, besides the intention in favour of the joint action (i.e. the intention that we *J*), each agent should have an intention in favour of *mutual goal adoption* in the pursuit of the joint action, i.e. the intention that *we adopt each other's goals when deciding how to J*.

An intention in favour of joint goal adoption is simply a commitment to *reason* in a strategic decision-making context by appealing to the agents' allocentric preferences and not to their personal preferences. This is a mode of reasoning in which one reasons adoptively towards the other and represents one's own goals from a third-person perspective too. To stress this aspect, I will name it *allocentric reasoning*. For

reasons already discussed, such mode of reasoning promotes interpersonal coordination and structure bargaining and shared deliberation, and so it is needed to fulfil the roles that a shared intention is supposed to play.

Of course, an intention in favour of a joint form of *reasoning* is not different from one in favour of a joint *action*. As for the latter intention, an individual cannot hold the former one alone: the way in which two agents reason together is not under the control of any of them separately. Hence, in order to intend this joint mode of reasoning, such intention must be formed on the grounds that the agents are engaging in mutual goal adoption. Though they may have independent reasons as well, each of them comes to have the goal to engage in joint allocentric reasoning on how to do the joint action *because and as long as* the other one has such a goal, and each is thus disposed to shift to a more egocentric mode if and when the other drops out.

I think that the adoptive origins of the intention that we engage in joint allocentric reasoning when choosing the appropriate means for a joint project nicely correspond to our experience of collaborating with others. It is common experience that when we are part of a collaborative enterprise we are disposed to take into account how our partners value certain options, provided that—and also because of—their similar reasoning towards us. Whenever this does not happen or ceases to happen, collaboration typically goes awry and one is left wondering whether the two of us were 'really' collaborating in the beginning.

Shared intention: the adoptive approach

We now have all the conceptual resources to offer a new analysis of shared intention. When two or more agents share an intention to act together, each has (1) an intention in favour of a joint action—*an intention that we J*—and (2) an intention in favour of joint allocentric reasoning in the pursuit of the joint action—*an intention that we reason about how to J from a third-person perspective*. Even if these intentions are mere personal intentions of the participants, (3) the goals on which they are based are *interdependent* because both (3a) the goal to do the joint action and (3b) the goal to engage in joint allocentric reasoning over the means for the joint action are, in part, (4) *the output of a mutual goal adoption process*. When (5) the structure of interdependency between these goals and the relevant intentions are common knowledge, a shared intention between two or more agents is realized.

This shared intention has the constitutive roles of supporting interpersonal coordination and of structuring relevant bargaining and shared deliberation. The social norms of social agglomeration and consistency, social coherence and social stability emerge from an appropriate functioning of the personal intentions of the participants in a context of common knowledge. A shared intention with specific social properties is, thus, realized only thanks to individual mental attitudes and processes.

Beyond Bratman's semantic strategy

The adoptive approach to shared intention is here proposed as an alternative realization of a shared intention as first analyzed by Michael Bratman (1993). Since it

presupposes the planning theory of intention (Bratman 1987) and the two distinctive roles that identify a shared intention, it is not offered as a challenge to Bratman's own analysis (see Bratman 1993, 1999, 2009a, c, 2007). Notwithstanding so, the adoptive approach presents some advantages.

Bratman's strategy to ground a shared intention in the personal intentions of the agents mainly exploits the *semantic interconnection* between the intentions of the agents (see for instance Bratman 2009a, p. 157). Taking this strategy means that, in order to specify a structure of interrelated personal intentions able to play the roles of interpersonal coordination and structuring of bargaining and shared deliberation, Bratman has to postulate, in addition to the intention in favour of the joint activity, two other personal intentions that explicitly refer (1) to the role of each other intentions in the joint action (e.g. each of the participants should have an intention that they act jointly by way of each other's intention to act jointly) and (2) to the fact that possible sub-plans are mutually compatible (e.g. each of the participants should have an intention that they act jointly by way of meshing sub-plans). These highly complex contents have limited both the transparency of Bratman's proposal and the domain of applicability of the theory. Tollefsen (2005), for instance, has suggested that young children lack the sophisticated mind-reading ability that is required to infer mental states with such a complex content (see also Knoblich and Sebanz 2008). Since, however, 2-year-old children already manifest the behavioural traits characteristic of collaboration (Warneken et al. 2006), there is a need for an analysis of shared intention that is compatible with this evidence.

The adoptive approach to shared intention is thus a valuable alternative since is possibly less demanding as far as understanding of each other's intentions is concerned. It requires that the participating agents have the ability to form goals with social content, but this content is just that the other one achieves what he wants. It implies that agents are able to form allocentric representations of others' goals and are able to switch between egocentric and allocentric perspectives. It does not need interconnectedness in the content of intentions because the interdependence is due to the fact both participants' relevant goals (i.e. that we act jointly and that we choose the means from the third-person perspective) are adopted from each other. Though we do not know enough about allocentric mind reading and its development (Frith and de Vignemont 2005), in the context of spatial cognition, the ability to form an allocentric representation of object position is already displayed by 2-year-old children (Ribordy et al. 2013). Interestingly, Warneken and Tomasello (2006) have shown that children of this age are already quite proficient 'goal adopters'.

Finally, Bratman's analysis looks like a still image of the participants' reciprocal mental attitudes without any understanding of the *process* that underlies the formation of such complex intentions. Castelfranchi and Paglieri (2007) have argued that a model of the process of intention generation helps to clarify the nature of intentions as distinctive mental states. The present focus on goal adoption and on the relation between reciprocal goal adoption and the intentions in favour of a joint action and a joint mode of reasoning is in the same spirit.

Before team-oriented approaches

The adoptive approach to shared intention bears some similarity also with team-oriented approaches to collective intention (see for instance Gold and Sugden 2007 and Sugden 2011). Rejecting any attempt to identify a collective intention with a list of constitutive mental states, Gold and Sugden (2007) suggest that an intention has the property of being *individual* or *collective* only relative to the mode of practical reasoning that the agents endorse to infer it. They argue that the distinctive mode of reasoning that generates a collective intention is *team reasoning* in which agency is attributed to groups and not to individuals (see also Sugden 2000, 2011 and Bacharach 2006).

When two agents face a situation of mutual dependence and have a common interest, each player might identify with a group. According to Bacharach (2006), for instance, the interdependency of the players and the perfect alignment of interests in the Hi-Lo game facilitate that a player thinks about herself and her co-player as 'us'. From the perspective of the group, one is thereby identified with, each player can conclude that we should choose Hi-Hi (because that is best for us) and that, if I am part of us, I should choose Hi.

When viewed from the perspective advance here, Bacharach's proposal seems to suggest that the agents are disposed to switch from their personal egocentric perspectives to a group-level ego-centric one that can be called *we-centric*. Appealing, however, to such we-centric viewpoint seems to be particularly relevant when one has to adopt goals that are *not* of any participant but of the group itself. In such cases, an agent has to act *as* a group member, and this shift of identity can also have dramatic consequences. A paradigmatic example is offered by a team leader who has to select people under budget constraints: Being forced to a choice, an agent thinking of herself *as* a chief has to make a choice in the group's interests without taking into account the third-person perspective of those he has to fire.

A theory that is able, in fact, to account for the possible discrepancy between ego-centric and we-centric preferences is Sugden's (1993, 2000) theory of team preferences.[24] Sugden's theory allows the agents, who take themselves to be members of the team, to care for their team in a way that is not reduced to the maximization of the utilities of the individual members. The core of the proposal is that preferences should be attributed to a unit of agency, *which is not necessarily an individual*. Accordingly, if the unit of agency is the team, one can identify a consistent set of preferences over possible outcomes that the team has the power to achieve. Hence, when one acts as a team member, one is acting *for* the team with the aim of maximizing a team utility function. Such team preferences are held by individuals that identify with their team. This process yields independent preferences across interacting participants, and by endorsing team reasoning, the team preferences

[24] In a recent contribution on these issues, Sugden has sketched a different account of team reasoning that appeals to intuitions coming from cooperative game theory. The agents that engage in team reasoning are taken to choose the profile of actions that correspond to the one they would agree on if their agreement were enforceable. This new approach is very similar in spirit to the one defended here; see Sugden (2011).

of the agents allow them to choose actions that may even overtly contrast with what they would choose as individuals.

When compared to my combination of allocentric preferences, allocentric utility functions and joint allocentric reasoning, Sugden's proposal might point to a different mechanism to solve coordination and cooperation problems. The adoptive approach to shared intention suggests that, in the kind of small-scale sociality that characterizes simpler cases of joint actions like that of painting a house together, an appeal to a different unit of agency is not actually needed for an appropriate explanation. Even if in small-scale projects without structured groups, the unit of agency is still the individual agent with his or her own personal egocentric preferences, this individualistic aspect does not prevent one from switching to allocentric mind reading and allocentric preferences, to take into account both one's own and another agent's perspective in an impartial way and so to reason impartially from both perspectives at once.

Moreover, the appeal to team preferences dramatizes the interaction between an individual and the group one might belong to. In contrast, allocentric preferences focus on a similar interaction *across* individuals.

Despite their differences, the adoptive and the team-oriented approach might be, in fact, complementary and focus on different processes. A more complete theory of joint action should explain how these two mechanisms could be combined.

What is special of human collaborative ability?

Before concluding, I want to briefly elaborate on the possible consequences of the adoptive approach to shared intention for the scientific study of human collaboration. I have argued that two crucial interconnected cognitive abilities are essential to explain human collaboration: allocentric mind-reading and (on this basis) goal adoption. Hence, this approach supports a general prediction: Possible differences between typical and atypical humans and between human and non-human primates with respect to their collaborative abilities might be in part due to different levels in (1) the ability to represent the interacting partner from an allocentric perspective and (2) to switch between the egocentric perspective and the other-centred, allocentric one. Though a full development of this egocentric–allocentric–switch hypothesis is beyond the scope of this contribution, consider these two kinds of evidence.

Frith and de Vignemont (2005) have recently suggested that people with Asperger syndrome (usually considered as a mild form of autistic disorder combined with high verbal activity; Frith 2004) might be characterized precisely by an improper functioning of the switch between the egocentric and allocentric kinds of social cognition. People with Asperger syndrome have a high level social competence, but they fail precisely in tasks that require full blown collaborative attitudes: Children with this syndrome are not interested in cooperative play and adults have difficulties in being selected as team members and to smoothly participate in group action. Somewhat similarly, a vast amount of evidence collected by Michael Tomasello and his collaborators (Tomasello et al. 2005; Tomasello and Herrmann 2010) suggests that an analogous pattern characterizes great apes' cooperation. In fact, while humans and great apes share many cognitive skills, the main difference lies in social cognitive ones. Even if great apes are quite proficient mind readers, this ability is triggered

especially in competitive (egocentric) contexts (Hare et al. 2006). Differently from humans, great apes do not seem to engage in mind reading in cooperative contexts and do not collaborate between themselves or with their human partners in the way humans do: They are not interested in cooperative play, they do not have human typical social responsiveness to the joint action and they do not help each other even when it might be instrumentally useful to reach a common goal (Warneken et al. 2006; Warneken and Tomasello 2006). To explain this difference, Tomasello has suggested that humans have evolved a specific set of 'skills and motivations for shared intentionality'. Since, however, such skills and motivations are not explained further, it is not clear what actual difference is at stake here. The adoptive approach to collaboration taken in this work suggests that either allocentric mind reading itself or the more sophisticated ability to switch between egocentric and allocentric perspectives when reasoning about others' goals might be the core ability.

As a final note, it is interesting to remind that Piaget (1962/1995) himself, who has notoriously emphasized the importance of decentred cognition for cognitive development, has suggested that Vygotsky's objections to Piaget's individualism have failed to appreciate 'egocentrism itself *qua* obstacle to the coordination of viewpoints and to cooperation'. The consequences of this idea for an appropriate explanation of human collaborative abilities are still to be fully untapped.

Conclusion

Understanding the cognitive underpinnings of joint action is an area that is, lately, attracting much interdisciplinary attention. Even if the evolutionary relevance of joint action and collaboration and the underlying cognitive complexity have been underestimated in the past, this situation is nowadays quickly changing.

The interplay between philosophy and other behavioural sciences, like psychology and economics, is thus particularly timely because there is a need for integrative conceptual frameworks that might support the formulation of new scientific theories and facilitate interdisciplinary discussion. Philosophy can contribute to this enterprise.

The adoptive approach to shared intention is offered as a new conceptual framework. It suggests that human collaboration is based on the ability to *share* mental states like beliefs, goals and intentions. Focusing on shared intention in particular, it contends that when two agents share an intention to act together, each individual has an intention in favour of the *joint action* and in favour of a *joint mode of reasoning*. Such mode of reasoning is identified with *allocentric reasoning*, that is, the ability to reason and act in view of one's own (allocentrically represented) goals and the goals adopted from somebody else. Since personal intentions with respect to joint activity (both action and reasoning) cannot be formed in isolation, it contends that these intentions are possible thanks to reciprocal goal adoption. Since adopted goals are conditional goals, reciprocal goal adoption creates a structure of *interdependence* between the personal goals of the participants. Thus, the ability of goal adoption emerges as a core cognitive ability whose role to understand human sociality should be adequately addressed. In this paper, I have explored how basic mind-reading abilities and allocentric representations might enable goal adoption and joint allocentric reasoning.

In order to argue for this view, I have borrowed tools from the planning theory of intention (Bratman 1986) and from the game-theoretical approach to bargaining problems. In particular, the extension of the standard framework as suggested by Harsanyi (1977) and Binmore (2005) has been useful to identify the role of allocentric mindreading and goal adoption in choosing the appropriate means for a joint action.

As such, the adoptive approach to shared intention contributes to the scientific study of joint action by identifying core *abilities* that the agents should display in order to collaborate in view of common goals. Even if the chosen level of explanation abstracts from underlying cognitive mechanisms, it is broadly compatible with contemporary approaches that are focused on the role of action in cognition and that emphasize the importance of off-line simulation to understand representing and reasoning abilities (Pacherie 2008, 2012; Pezzulo and Castelfranchi 2009). By pinpointing new precise abilities and their pre-requisite, the adoptive approach can identify constraints at the level of cognitive mechanisms and can directly inform empirical investigation in the domains of developmental and comparative psychology. Whether the adoptive approach can successfully systematize empirical data is a matter of future research.

Acknowledgments Earlier versions of this paper were presented at the International Conference on *Reciprocity*: *Theories and Facts* (February 2007) and at the Workshop on *Michael Bratman and the Structure of Agency* (University of Berne, September 2007). This paper has enormously benefited from numerous discussions in particular with: Giacomo Bonanno, Michael Bratman, Luigino Bruni, Cristiano Castelfranchi, Herbert Gintis, Reto Givel, Natalie Gold, Davide Grossi, Emiliano Lorini, Maria Miceli, Elisabeth Pacherie, Fabio Paglieri, Robert Sugden and two anonymous reviewers.

References

Apperly, I. A. (2012). What is "theory of mind"? Concepts, cognitive processes and individual differences. *The Quarterly Journal of Experimental Psychology, 65*(5), 825–839.

Bacharach, M. (2006). Beyond individual choice. In N. Gold & R. Sugden (Eds.), *Teams and frames in game theory.* Princeton: Princeton University Press.

Bardsley, N., Mehta, J., Starmer, C., & Sugden, R. (2010). Explaining focal points: cognitive hierarchy theory versus team reasoning. *The Economic Journal, 120*, 40–79.

Binmore, K. (1994). *Game theory and the social contract. Volume 1: playing fair*. Cambridge: MIT.

Binmore, K. (2005). *Natural justice*. Oxford: Oxford University Press.

Bowles, S. (2004). *Microeconomics. Behavior, institutions and evolution*. Princeton: Princeton University Press.

Bratman, M. E. (1987). *Intentions, plans and practical reason*. Cambridge: Harvard University Press.

Bratman, M. E. (1993). Shared intention. *Ethics, 104*(1), 97–113.

Bratman, M. E. (1999). *Faces of intention*. Cambridge: Cambridge University Press.

Bratman, M. E. (2007). *Structures of agency.* Oxford: Oxford University Press.

Bratman, M. E. (2009a). Modest sociality and the distinctiveness of intention. *Philosophical Studies, 144*, 149–165.

Bratman, M. E. (2009b). Intention rationality. *Philosophical Explorations: An International Journal for the Philosophy of Mind and Action, 12*(3), 227–241.

Bratman, M. E. (2009c). Shared agency. In C. Mantzavinos (Ed.), *Philosophy of the social sciences: philosophical theory and scientific practice* (pp. 41–59). Cambridge: Cambridge University Press.

Butterfill, S., & Sebanz, N. (2011) Joint action: what is shared? *Review of Philosophy and Psychology*, *2*(2), 137–373.

Castelfranchi C. (2003). Grounding we-intentions in individual social attitudes. In M. Sintonen, P. Ylikoski & K. Miller (Eds.), *Realism in action. Essays in the Philosophy of the Social Sciences, Synthese Library Vol. 321*. Dordrecht: Kluwer Academic.

Castelfranchi, C. (2012). Goals, the true center of cognition. In F. Paglieri, L. Tummolini, M. Miceli, & R. Falcone (Eds.), *The goals of cognition. Essays in Honor of Cristiano Castelfranchi* (pp. 825–870). London: College.

Castelfranchi, C., & Paglieri, F. (2007). The role of beliefs in goal dynamics: Prolegomena to a constructive theory of intentions. *Synthese, 155*(2), 237–263.

Castelfranchi, C., & Parisi, D. (1984). Mente e scambio sociale. *Rassegna Italiana di Sociologia, 1*, 45–72.

Conte, R., & Castelfranchi, C. (1995). *Cognitive and social action*. London: UCL.

de Vignemont, F., & Jacob, P. (2012). What is it like to feel another's pain? *Philosophy of Science, 79*(2), 295–316.

de Vignemont, F., & Singer, S. (2006). The empathic brain: how, when and why. *Trends in Cognitive Sciences, 10*(10), 435–41.

Frischen, A., Loach, D., & Tipper, S. P. (2009). Seeing the world through another person's eyes: simulating selective attention via action observation. *Cognition, 111*, 212–218.

Frith, U. (2004). Confusions and controversies about Asperger syndrome. *Journal of Child Psychology and Psychiatry, 45*, 672–686.

Frith, U., & de Vignemont, F. (2005). Egocentrism, allocentrism, and Asperger syndrome. *Consciousness and Cognition, 14*, 719–738.

Frith, U., & Frith, C. (2010). The social brain: allowing humans to boldly go where no other species has been. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 365*(1537), 165–76.

Gilbert, M. (1990). Walking together: a paradigmatic social phenomenon. *Midwest Studies, 15*, 1–14.

Hare, B., Call, J., & Tomasello, M. (2006). Chimpanzees deceive a human competitor by hiding. *Cognition, 101*, 495–514.

Harsanyi, J. (1977). *Rational behavior and bargaining equilibrium in games and social situations*. Cambridge: Cambridge University Press.

Hodgson, D. H. (1967). *Consequences of utilitarianism*. Oxford: Clarendon.

Hoffman, M. L. (1978). Empathy: its development and prosocial implications. In J. H. E. Howe & C. B. Keasey (Eds.), *Nebraska symposium on motivation: social cognitive development* (pp. 169–218). Lincoln: University of Nebraska Press.

Knoblich, G., & Sebanz, N. (2008). Evolving intentions for social interaction: from entrainment to joint action. *Philosophical Transactions of the Royal Society B: Biological Sciences, 363*(1499), 2021–2031.

Knoblich, G., Butterfill S. & Sebanz, N. (2011). Psychological research on joint action. In B. Ross (Ed.), *Psychology of learning and motivation* (Vol. 51, pp. 59–101). New York: Academic.

Lewis, D. K. (1972). Psychophysical and theoretical identifications. *Australasian Journal of Philosophy, 50*, 249–258.

Meltzoff, A. N. (2002). Imitation as a mechanism of social cognition: origins of empathy, theory of mind, and the representation of action. In U. Goswami (Ed.), *Handbook of childhood cognitive development* (pp. 6–25). Oxford: Blackwell.

Moll, H., & Tomasello, M. (2007). Co-operation and human cognition: the Vygotskian intelligence hypothesis. *Philosophical Transactions of the Royal Society—Series B: Biological Sciences, 362*(1480), 639–648.

Osborne, M. J., & Rubinstein, A. (1994). *A course in game theory*. Cambridge: MIT.

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

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

Pacherie, E. (2012). The phenomenology of joint action: self-agency vs. joint-agency. In A. Seemann (Ed.), *Joint attention: new developments* (pp. 343–389). Cambridge: MIT.

Pezzulo, G., & Castelfranchi, C. (2009). Thinking as the control of imagination: a conceptual framework for goal-directed systems. *Psychological Research, 73*, 559–577.

Piaget, J. (1962/1995). Commentary on Vygotsky's criticisms of language and thought of the child and judgment and reasoning in the child. *New Ideas in Psychology*, *13*, 325–340.

Ribordy, F., Jabès, A., Banta Lavenex, P., & Lavenex, P. (2013). Development of allocentric spatial memory abilities in children from 18 months to 5 years of age. *Cognitive Psychology, 66*(1), 1–29.

Schelling, T. (1960). *The strategy of conflict*. Cambridge: Harvard University Press.

Searle, J. (1990). Collective intentions and actions. In P. Cohen et al. (Eds.), *Intentions in communication* (pp. 401–415). Cambridge: MIT.

Smith, M. (1987). The Humean theory of motivation. *Mind, 96*, 36–61.
Sterelny, K. (2007). Social intelligence, human intelligence and niche construction. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 362*(1480), 719–730.
Sugden, R. (1991). Rational choice: A survey of contributions from Economics and Philosophy. *The Economic Journal, 101*(47), 751–785.
Sugden, R. (1993). Thinking as a team: toward an explanation of nonselfish behavior. *Social Philosophy and Policy, 10*, 69–89.
Sugden, R. (2000). Team preferences. *Economics and Philosophy, 16*, 175–204.
Sugden, R. (2001). Ken Binmore's evolutionary social theory. *The Economic Journal, 111*, F213–F243.
Sugden, R. (2009). Neither self-interest nor self-sacrifice: the fraternal morality of market relationships. In S. A. Levin (Ed.), *Games, groups, and the global good* (pp. 259–283). Dordrecht: Springer.
Sugden, R. (2011). Mutual advantage, conventions and team reasoning. *International Review of Economics, 58*, 9–20.
Sugden, R., & Gold, N. (2007). Collective intentions and team agency. *The Journal of Philosophy, 104*, 109–137.
Tollefsen, D. (2005). Let's pretend! Children and joint action. *Philosophy of the Social Sciences, 35*(1), 75–97.
Tomasello, M., & Herrmann, E. (2010). Ape and human cognition: what's the difference? *Current Directions in Psychological Research, 19*, 3–8.
Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: the origins of cultural cognition. *The Behavioral and Brain Sciences, 28*(5), 675–691.
Tuomela, R. (1995). *The importance of us: a philosophical study of basic social notions*. Stanford: Stanford University Press.
Velleman, J. D. (1997). How to share an intention. *Philosophy and Phenomenological Research, 57*, 29–50.
Von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton: Princeton University Press.
Warneken, F., & Tomasello, M. (2006). Altruistic helping in human infants and young chimpanzees. *Science, 31*, 1301–1303.
Warneken, F., Chen, F., & Tomasello, M. (2006). Cooperative activities in young children and chimpanzees. *Child Development, 77*(3), 640–663.