Editorial Manager(tm) for The Review of Philosophy and Psychology
Manuscript Draft

Manuscript Number: ROPP63

Title: SHARED INTENTION REVISITED: ON COLLABORATIVE REASONING IN JOINT ACTION CONTEXTS

Article Type: Joint Action: What is Shared?

Keywords:

Corresponding Author: Dr. Luca Tummolini, Ph.D

Corresponding Author's Institution: Consiglio Nazionale delle Ricerche

First Author: Luca Tummolini

Order of Authors: Luca Tummolini

Abstract: Two agents sharing the intention to do something are interlocked in a web of individual intentions with special social contents. This individualistic approach to shared intentional activity has been originally proposed by Michael Bratman. However, Bratman's 'shared intention' is of little help when interdependency of agents' reasons can threaten coordination as in common-interest games with multiple equilibria. In particular, in Hi-Lo games (coordination games with at least two equilibria one of which is Pareto superior), it has been argued that agents fail to coordinate when they reason from an individualistic perspective. In this article, an explanation of individualistic choice in Hi-Lo situations is proposed. The propensity of agents to choose Hi is interpreted as the output of both a preference and a reasoning effect. Choosing Hi presupposes both social preferences and a framing of the situation as a collaborative choice requiring usual individual practical reasoning (although social in its scope). Agents that can be motivated by an other-regarding cognitive process of goal adoption have rational grounds both to infer a shared intention in the first place and then to specify it by generating new shared intentions as time goes by.

Suggested Reviewers:

SHARED INTENTION REVISITED: ON COLLABORATIVE REASONING IN JOINT ACTION CONTEXTS

Luca Tummolini
Istituto di Scienze e Tecnologie della Cognizione (CNR)
Via San Martino della Battaglia 44
00185 - Roma - Italy
luca.tummolini@istc.cnr.it

SHARED INTENTION REVISITED: ON COLLABORATIVE REASONING IN JOINT ACTION CONTEXTS

Abstract

Two agents sharing the intention to do something are interlocked in a web of individual intentions with special social contents. This individualistic approach to shared intentional activity has been originally proposed by Michael Bratman. However, Bratman's 'shared intention' is of little help when interdependency of agents' reasons can threaten coordination as in common-interest games with multiple equilibria. In particular, in Hi-Lo games (coordination games with at least two equilibria one of which is Pareto superior), it has been argued that agents fail to coordinate when they reason from an individualistic perspective. In this article, an explanation of individualistic choice in Hi-Lo situations is proposed. The propensity of agents to choose Hi is interpreted as the output of both a preference and a reasoning effect. Choosing Hi presupposes both social preferences and a framing of the situation as a collaborative choice requiring usual individual practical reasoning (although social in its scope). Agents that can be motivated by an other-regarding cognitive process of goal adoption have rational grounds both to infer a shared intention in the first place and then to specify it by generating new shared intentions as time goes by.

1. Introduction

Consider the case of us (you and me) *sharing* the intention to paint a house together. According to Michael Bratman (1993, 2007, 2009), the best analysis of this situation is as follows:

We (you and me) intend to paint the house if[1]:

(1)	I intend that WE PAINT THE HOUSE	You intend that WE PAINT THE HOUSE
(2)	I intend that WE PAINT THE HOUSE IN PART BY WAY OF YOUR INTENTION THAT WE PAINT THE HOUSE	You intend that WE PAINT THE HOUSE IN PART BY WAY OF MY INTENTION THAT WE PAINT THE HOUSE
(3)	I intend that WE PAINT THE HOUSE IN PART BY WAY OF SUB-PLANS THAT MESH WITH YOUR SUB-PLANS	You intend that WE PAINT THE HOUSE IN PART BY WAY OF SUB-PLANS THAT MESH WITH MY SUB-PLANS
(4)	I intend to HELP YOU PLAYING YOUR ROLE IN OUR PAINTING THE HOUSE	You intend to HELP ME PLAYING MY ROLE IN OUR PAINTING THE HOUSE
(5)	I believe that IF WE CONTINUE TO INTEND THAT WE PAINT THE HOUSE AS A RESULT WE WILL PAINT THE HOUSE	You believe that IF WE CONTINUE TO INTEND THAT WE PAINT THE HOUSE AS A RESULT WE WILL PAINT THE HOUSE
(6)	I believe that I INTEND THAT WE PAINT THE HOUSE IF AND ONLY IF YOU INTEND THAT WE PAINT THE HOUSE	You believe that YOU INTEND THAT WE PAINT THE HOUSE IF AND ONLY IF I INTEND THAT WE PAINT THE HOUSE
(7)	All the conditions above (including this one) are common knowledge between me and you	

Adopting Bratman's approach, sharing an intention to do something together is nothing mysterious. It is sufficient that the *contents* of the usual individual intentions of each participant mention, in adequate ways, the intentions of the others. The appropriate "interlocking" of these individual intentions, as Bratman says, must be so that the individual intentions are able to realize

[1] Bratman's more recent contributions (1999 and 2007) defend the sufficiency and not the necessity of these conditions. For the aim of this paper I adopt the same view. Moreover, I use, whenever it is relevant, the capital letter to distinguish the *mental attitude* (i.e. the belief, the goal, the intention) from the *mental content* of this attitude.

the following roles: *they coordinate the intentional actions and the planning of the involved participants*, and *they structure relevant bargaining between them*[2]. A shared intention is realized by the individual attitudes of the participants or, to put it differently, emerges from a proper interrelation of such individual attitudes.

In this paper I will offer an argument to show that Bratman's conditions (within standard assumptions about preferences) are not able to account for one of the roles that a shared intention is supposed to play. In an important situation that occurs frequently when we engage in joint actions, sharing an intention in Bratman's sense is not enough to coordinate our actions. Given Bratman's commitment to functionalism[3] this objection is threatening for the overall soundness of the analysis. If it can be shown that the interlocking web of individual intentions cannot have the coordinating *role* that is supposed to have, a shared intention will not be realized (or emerge). This objection raises an additional worry. If sharing an intention cannot warrant the derivation of a new more specified web of intentions as time goes by, one can even question the possibility to share an intention in the first place. I will claim that if the former problem is not solved, the agents have no rational ground to infer a shared intention whatsoever. However, the solution I suggest to the problem of the rational derivation of new shared intentions is also intended to block the scepticism about shared intention formation in the first place. Finally, I will propose a revised analysis of the conditions that are sufficient for sharing an intention.

2. Deriving a new shared intention

Let's add to our opening simple example an artificial intelligence flavour.

Suppose a designer have constructed two robots: C3PO and R2D2. These robots have only bounded computational resources that can be allocated to deliberate on what to do, but the designer wants them to be able to achieve complex goals requiring actions that extend through time. This

[2] For a more extended explanation of what these conditions mean and imply see Section 2 below. For a complete clarification and defence see Bratman (2007).

[3] 'Functionalism' in the philosophy of mind is the doctrine according to which a mental state of a particular type is not identified by its internal constitution, but rather by the way it functions, or the role it plays, in the system of which it is a part; see for instance Lewis (1972).

ability entails that they must be able to coordinate their individual activities across time in a way that let them, normally, reach such goals. To meet this challenge, the designer has added to their computational minds a mental state devoted precisely to solve these problems. Given the specific roles such state plays in these computational systems (i.e. filtering new inconsistent options out of deliberation, being input for additional means-end reasoning and committing them to actions to be executed in the future in a way that resist constant revision of previously chosen plans of action), we can safely claim that the designer has succeeded in providing C3PO and R2D2 with genuine *intentions* in Bratman's future-directed sense (Bratman 1987)[4].

Granted this capacity of forming true individual intentions, let's suppose that both C3PO and R2D2 have the goal to paint a house together as a joint intentional action. How could they *share* something more committing to the joint endeavour action than a mere goal?

According to Bratman, they need to satisfy the seven conditions summarized above. So, because we assume that those conditions are satisfied, C3PO and R2D2 will *share* the intention to paint the house. Having a shared intention amounts to be in *an interlocking web of individual intentions* such that the robots are enabled (1) to coordinate their actions across time, (2) to further their planning in compatible ways, and, if needed, (3) to structure the relevant bargaining between them on which option to choose.

For example, each of them, individually, should be able to filter new individual or joint options that are inconsistent with their beliefs and intentions. To this end, condition (1) expresses *a commitment to the joint action such that new possible joint options that are inconsistent with it* (e.g. going for a walk together) *will not be considered in their deliberation*. Moreover as in every intention, knowledge of the other having such intention will provide reliable predictions supporting coordination between them. Condition (2) differently *expresses a commitment to their reciprocal interlocking*. And this means that, in case one of them ceased the intentional pursuit of the joint action, the other would immediately stop, and the shared intention would be dropped: they do not intend to coerce the other, and both are only interested in obtaining a painted house as a result of each other intending to paint it. Conditions (5) and (6) indicate that *the agents are mutually*

[4] See Bratman, Israel and Pollock (1988) for the original proposal for such a computational architecture.

responsive to each other intentions and action in a way that tracks the joint action they intend to do, and that *the persistence of their reciprocal commitment to the joint action is interdependent*. That is, according to former condition, the agents believe that their reciprocal intentions to paint the house will effectively conduce to the fact that the house will be painted. The latter one, on the other hand, ensures that the two robots persist in intending to paint the house together if and only if each of them continues to intend to do it. For instance, in case C3PO dropped the commitment to paint the house with its fellow, R2D2 would do the same.

Moreover, an intention as that of painting a house together is indeed a very general one. Just sharing this general intention does not amount to having a shared recipe on how to proceed in each time step. However having, both of them, an intention to proceed by way of the other intending the same (condition 2) and an intention to mesh one's own plan with the other (condition 3), they will proceed in further means-end reasoning in ways that maintain coherence with the already established commitment to the joint activity (conditions 1 and 2). Condition (3) is introduced precisely for ensuring that *the agents are committed to identifying sub-plans that "mesh"*, in the sense of being compatible and co-realizable.

Finally, if at some point one of them, say C3PO, faces some difficulty at playing his role in the joint action (e.g. C3PO risks to trip over the bucket while having his artificial hands busy with the brushes), Condition (4) ensures that *the other one is committed to help* if that is not too burdensome (e.g. by warning or by moving away the bucket).

At some point, for example, C3PO and R2D2 will need to choose a colour for the house. Supposing that the two robots have lived a long life together and have shared many adventures, they feel confident about each other's tastes when it comes to choosing a colour. To make it simple, we assume that they already are in a situation of common knowledge[5] 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 (say for some legal policy ruling house painting in the galaxy they ended up inhabiting) is to use a yellow colour. But they both know, that they both

[5] *Common knowledge* is traditionally understood as that collective epistemic state in which all agents believe that *p*, all agents believe that all agents believe that *p*, all agents believe that all agents believe that all agents believe that *p*, and so on *ad infinitum*. See the canonical, though different, analyses of David Lewis (1969) and Robert Aumann (1976).

know and so on, that they both prefer living in a blue house than living in a yellow house. Finally, what is also common knowledge between them is that the worst possible situation is to have half of the house painted in blue and the other half painted in yellow. Because the outcome they are pursuing depends on the choices of both, this is a strategic decision. If the agents choose what to do *by reasoning*, it seems natural to assume that they will engage in *strategic reasoning*, that is, reasoning about what is best for one to do, given that the other will do what is best for him to do.

Adopting the game theoretic notation for game in normal or strategic form, we could represent their situation with the following matrix:

		C3PO	
		Paint Blue	Paint Yellow
R2D2	Paint Blue	2,2	-5,-5
	Paint Yellow	-5,-5	1,1

Figure 1: The choose-the-colour game

This matrix shows the reciprocal ranking relative to possible outcomes of the two robots, i.e. how the world would be after their choices. The clear best outcome 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 to mix the colours and they clearly don't want to end in those outcomes (as it is emphasized for illustrative purposes by the negative utilities). Let's dub this situation the 'choose-the-colour' game.

Suppose now that because they have common knowledge of this situation (they already know to have the same preference ranking), they have considered useless to engage in an explicit deliberation on what colour paint to use for their joint endeavour. Without communicating then, from the shared intention to paint the house and from the common preference to paint it blue each of them derives (1) an intention that they paint a *blue* house, (2) an intention that they paint a *blue* house by way of the other intention that they paint a *blue* house, (3) an intention that they paint a *blue* house in part by way of sub-plans that mesh with the other, (4) an intention to help the other to play the of painting a *blue* house (5) a belief that if they will continue to intend that they paint a *blue* house, then as a result, they will paint a *blue* house, (6) a belief that they intend that they paint a *blue* house if and only if the other intends that they paint a *blue* house.

Everything being under common knowledge (even this very condition 7), then we (and they) can conclude that R2D2 and C3PO have derived *a new shared intention* to paint a blue house.

If this reasoning is sound, they can proceed in their planning and start painting. The question is of course whether it really is sound. In the next section I intend to show that there are no reasons for our robots to form a derived intention similar to the one expressed in condition (1) since they don't have a decisive reason to use the blue paint in the first place.

3. The problem

Unfortunately C3PO and R2D2 are caught in a problem that, although maybe ignored by them, is known to game-theorists as the *Hi-Lo Paradox* (Bacharach 2006)[6].

Let's provide some background. The choose-the-colour game is an instance of a class of strategic interactions that are known in the game theoretic literature as 'common-interest games'. These are situations in which the agents (or players*) have a common interest in acting in a coordinated way*; hence they are the prototypical situations in which a *joint action* is needed.

As it is exemplified in our choose-the-colour game, the agents have a symmetrical (or at least a compatible) preference ranking such that the favoured *strategy* for one of them (what is best to do, given what the other can do) is the same strategy preferred by the other. This means that the strategies favoured by both – the *strategy profile* – are also, adopting the game-theoretical framework, a *Nash equilibrium*. A strategy profile, in fact, 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.

In common interest games the players have the common goal to coordinate. However a threat to reciprocal coordination is posed by the fact that strategic situations are often characterized by *multiple* equilibria where each of the players is indifferent about which strategy to choose provided that the others choose the same. Because coordination in such situations is at risk and is not guaranteed, players need to solve a 'coordination problem' (Schelling 1960).

[6] The paradoxical nature of this game lies in the divergence between the 'intuitive' rationality of a choice and the inability of accepted theories of rationality to predict (or prescribe) that choice.

Consider for example a coordination game like the following:

		Player 1	
		Head	Tail
Player 2	Head	1,1	-1,-1
	Tail	-1,-1	1,1

Figure 2: The game of Matching Pennies

Players 1 and 2 are playing the game of matching their 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 what should they do in this situation?

Because the payoff matrix and the fact that they reason in a symmetric way[7] is common knowledge, they could try to predict what the other will do by engaging in *strategic reasoning*.

Consider for example Player 1's reasoning. Player 1 thinks that if Player 2 will play Head then her best reply is to play Head too. At the same time, if Player 2 will play Tail, the best response for Player 1 would be to play Tail. But when will Player 2 for example play Head? When he will be able to predict that Player 1 will play Head too. However, because Player 2's choice is symmetrically dependent on what Player 1 will do, both of them by pure reasoning are not able to arrive to a definite reason to turn one side or the other of the coin. In this situation there are two Nash equilibria (Head-Head and Tail-Tail) and none of the players has an *independent* reason to choose Head or Tail. What the payoff matrix makes explicit is that in such situations the players have *interdependent* reasons to choose a strategy. It is the interdependency of the reasons of the players in coordination problems that makes it impossible for them individually to choose (or select) an equilibrium in a definite way.

The problem that C3PO and R2D2 are facing is actually a variant of this game dubbed Hi-Lo[8]. The only difference with a 'pure' coordination game (one in which there is complete

[7] The fact that the agents know that they are 'symmetric reasoners' is a condition that is usually implicit in game theoretic models. On the importance of making this assumption explicit see Gintis (2009, Chapter 7).

[8] See also Hodgson (1967), Gauthier (1975), Sugden (1993, 2000) and Hollis (1998).

indifference between the two Nash equilibria) is that one of the two equilibria is also Pareto superior to other rather than equally good for both[9]. Our choose-the-color game is even stronger however because both C3PO and R2D2 are better off when they both choose the blue paint than when they both choose the yellow one.

So even if being in a coordination problem can in fact threaten coordination when there is indifference between equilibria (and in fact this does happen very often in our daily experiences[10]), it seems strongly intuitive that the evident rational solution for our robots is to choose the blue paint and stop worrying about it.

This is not so easy, however, because, as counter-intuitive as it can be, Hi-Lo situations have the same drawbacks of pure coordination games.

If it is difficult to step back from our intuition that it is obvious to 'directly' choose the most preferred outcome, we should remember that we are dealing with robots. The two robots have been constructed along the lines just described (so that what is intuitive for *us* is not intuitive for *them*), and so we have to ask if, from *their* point of view, they can settle to use (and hence to intend to use) the blue paint or not. Actually, notwithstanding that they both prefer the Paint_Blue-Paint_Blue equilibrium over the Paint_Yellow-Paint_Yellow one, in a one-shot situation without communication[11] (as it is the case in our thought experiment), the two robots are not able to choose rationally (i.e. instrumentally) the Pareto-superior outcome.

And this is so because the structure of their (strategic) reasoning is the same as in the case of the pure coordination game: even in the choose-the-paint case, the reason to choose the blue paint is an interdependent one. C3PO prefers to use the blue paint but is worried to end up in him using the

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

[10] The role of the *salience* of one equilibrium, to pick up an 'obvious' outcome in pure coordination games have been explored by many, see in particular Schelling (1960) and Sugden (1995). However a salient outcome is just one that is more probable that the other will notice. Salient outcomes then just turn a pure coordination game in a Hi-Lo game.

[11] There are arguments supporting the idea that if the interaction is iterated (Aumann and Sorin 1989) or if pre-play communication is allowed (Farrell 1988), the players can choose Hi in the Hi-Lo dilemma. But these solutions do not scale to the general case when iteration and communication are not allowed.

blue paint and R2D2 using the yellow one. C3PO is justified in its intention to achieve its most preferred outcome only if it can reliably predict R2D2's choice. Notwithstanding the fact that each of them prefers the blue house over of the yellow one, C3PO can intend to use the blue paint if and only if R2D2 is choosing the blue paint too. However the same is true for R2D2, and so they are both trapped in the coordination problem.

From the individual perspective, having a common preference for an outcome over another is not a sufficient reason to choose it and thereby forming the corresponding intention.

4. Do intentions in the background help?

One may suggest that having the interlocking web of intentions in the background typical of Bratman's shared intention can help the two robots in this kind of situations since intentions have a role in deliberation that is underestimated by game-theorists[12]. Intentions offer a background for deliberation over the means in the sense that they usually filter out new options that are inconsistent with already settled intentions (Bratman 1987).

In our example a relevant background intention is that specified in condition 3: i.e. the intention (that is common knowledge that each of them has) to paint the house by way of meshing sub-plans. If the structure of the payoffs is such that in the choose-the-colour game there is an implicit common interest in coordination, this intention can be seen as an explicit version of this common interest.

What such intention does is prompting each of the robots to search for sub-plans that are jointly co-realizable in order to avoid the outcome in which they don't coordinate (i.e. when the plans are not meshing with each other). This intention can do so because it is part of its *role* to filter out the inconsistent outcomes without explicit deliberation. However both the Paint_Blue-Paint_Blue equilibrium and the Paint_Yellow-Paint_Yellow one are meshing sub-plans. And so this background intention doesn't help in this situation because it fails to suggest a more specific course of action.

[12] For an attempt to embody (individual and not shared) 'intentions' in the game theoretic framework see Roy (2009).

If one assumes that in order to specify their intentions, the agents endorse in strategic reasoning, I think that this kind of situation offers a simple but clear-cut counter-example to the following statement: "our shared intention will coordinate our actions in part by ensuring that my planning about my role in the house painting is coordinated with your relevant planning, and vice versa" (Bratman 1999: 112).

At this point, one can argue that coordination can indeed fail: people simply don't always coordinate. Although this is true in more complex cases, it is strongly intuitive that two agents that share an intention to paint a house and have common knowledge of their common preference ranking as to which colour they both prefer should be capable of deriving a new shared intention such as: let's paint the house with our most preferred colour!

Is this difficulty a critical problem for Bratman's analysis of shared intention? I think it is for two reasons.

The first one originates from the commitment to functionalism. If having a shared intention in the background doesn't help us to coordinate our actions even in presence of a commonly known common preference-ranking, then the crucial coordination function that shared intentions are supposed to play is missing (at least in a relevant, trivial and recurrent situation). Without being able to explain how a shared intention is able to fulfil its role, one is forced to admit, from a functionalist viewpoint, that a shared intention is not realized after all.

The second reason springs from the explanatory role that a theory of shared intention should have. Because our intuition tells us that when we share an intention in such situations we would definitely choose the obvious most preferred outcome, a correct understanding of what sharing an intention is should explain what we do when we reason in such an evidently correct way.

Although I think these two reasons pose a serious problem, I suggest in the following sections a way to deal with them in a way that is coherent with the overall approach to shared agency.

5. From independent to dependent preferences

Before proceeding in the analysis, one must be sure that the notion of common preference ranking is at least consistent with Bratman's approach to shared intention.

In order to block a dangerous circularity in his analysis, Bratman justifies the content of intentions that we do a joint action J[13] in the following way: "we seek basic cases in which each participant intends a joint activity understood in a way that is neutral with respect to shared intentionality" (1999: 147) which amounts to a concept of the joint action in the weakest sense of avoiding to interfere with each other intentional actions without that such avoidance be intended by the participants.

Adopting the same line of thought, the standard preferences of the individuals engaging in the joint action are, in the most basic case, merely their usual self-regarding preferences. Granted that such preferences can conflict, for example "about the colour paint to use (…) such conflicts call for bargaining (…) about how we are to paint together" (1999: 112), and sharing an intention should ensure this. However, as I have shown in the previous sections, a shared intention *is not* able to ensure coordination, even in very simple cases where the self-regarding preferences of the participants are compatible already from the beginning.

In our working example, I have assumed that they both have an identical *independent* preference ordering under common knowledge, and in a situation in which the outcome depends on the choices of both (none of them having control over the outcome), the agents only have interdependent reasons to choose the blue-paint. And we have seen that the problem lies precisely in such interdependency in reasons for action.

A possible move, at this point, is to argue that a commonly known compatible preference ranking is too weak as a background when the agents share an intention. What would happen if they had *interdependent* preferences? Would dependency at the level of preferences help in achieving coordination?

To evaluate this possibility we first need to distinguish 'independent' from 'dependent' preference orderings.

[13] In what follows, I will use 'we do *J'* as a notational shortcut for "we do a joint action" or "we do an action together" that will be considered as equivalent propositions.

Adopting the standard economic approach, preferences are defined as binary relations over outcomes, and are considered as mental attitudes towards alternatives independently of any actual choice. But under the assumption that an agent is instrumentally rational, when he prefers an outcome to the others, he chooses the action (a strategy in the context of game theory) that realizes that outcome[14].

Independent preferences may be naturally viewed as *preferences that do not depend on other preferences*. In other words, an independent preference is unconditional in the sense that it is an all-things-considered judgment[15]. A *preference ordering* is *independent* from other preferences when, ceteris paribus, *changing other preferences does not change it*.

A *preference is dependent when one prefers something to something else conditionally on the belief that a state of affairs holds*. This fact entails that when such condition changes, the ordering will change as well[16]. The special case that interests us here is *when such condition is that of another agent having a certain preference*.

Within Bratman's formulation, behind an intention that is shared, there is an interpersonal dependency (inter-dependency) between the intentions of the two agents: R2D2 *intends* that we *J* if and only if C3PO *intends* that we *J*. Analogously, a condition at the level of preferences would specify that R2D2 *prefers* a blue paint if and only if C3PO *prefers* the blue paint too.

Notice that this situation is not the one we have discussed so far. When R2D2 and C3PO have a common preference ranking, they both *independently* prefer the Paint_Blue-Paint_Blue outcome to the others. *Although their preferences are independent, their strategic interaction*

[14] There is no unanimous consensus in the economic literature on what is the best interpretation of the 'preference' construct. I endorse in what follows the interpretation offered by Osborne and Rubinstein (1994).

[15] This is not to say that unconditional preferences are *absolute* preferences. Unconditional preferences are, actually, better understood as *frame-relative*, which means that *preferences are defined relative to particular descriptions of the choice problem*. Different descriptions of a choice problem can elicit different preference orderings. Notwithstanding so, the preferences are stable and consistent when the framing is held constant. On the relation between frames and preferences see Sugden (1995) and more extensively Bacharach (2006).

[16] For a model of preference change as a consequence of the agent coming to believe that a condition still holds or does not hold anymore see Jeffrey (1977).

provides them with interdependency at the level of choices. Remember that, up to this point, we have assumed that R2D2 and C3PO are strategic reasoners that are interested in the preferences of each other only to predict their mutual choices.

However if there were dependency between the preferences of the two robots, each of them would prefer an outcome (at least partially) *because* the other has the same preference. This notion of dependency seems fit to capture an aspect involved in 'sharedness' since *each of them endorses a judgment in virtue of the other endorsement, and stops endorsing it when the other stops*.

6. Common 'social' preferences are interdependent

Dependency between the preferences of different agents is related to the notion introduced by economists with the label of social preferences. An agent exhibits social preferences when she "not only cares about the material resources allocated to her but also cares about the material resources allocated to relevant reference agents" (Fehr and Fischbacher 2002: C2) and, more precisely, the key-aspect of a social preference is that "one's evaluation of a state *depends* on how it is experienced by others" (Bowles 2004: 109).

A standard example of a social preference is the preference for altruism. When an agent is an altruist, she tends to promote the welfare of another one as her own, and she is disposed to sustain even personal costs to this aim. Such pro-social attitude corresponds to the idea that an agent is *adopting* the goal of another one, and she comes to pursue this goal *since* and *until* it is pursued by the other one (Conte and Castelfranchi 1995). However the real altruist *adopts* the goal of another one due to a *terminal* motive (benevolence) and not as means to promote one's own self-regarding ends.

But, the cognitive process of goal adoption can be based on a more general spectrum of motives. Adopting the goal of another can be instrumentally useful to realize one's own selfish goals or it can be based on a common goal between two agents (Castelfranchi 1997 and 2003) as is the case in joint actions[17]. When we realize to have a goal in common (i.e. to have the house

[17] Castelfranchi (1997) distinguishes benevolent (terminal), instrumental and cooperative motivations behind goal adoption.

painted), we can be motivated to adopt the instrumental goals of each other as a way to achieve this common goal. Goal adoption then is an other-regarding cognitive process in the sense of being aimed to further the goal of another, and this might happen not necessarily on the basis of altruistic motives. A pro-social attitude is not necessarily an altruistic one.

In any case, the process of goal adoption originates only conditional goals: if I adopt your goal, in case you revised your own I would revise mine too. It is, then, clear that *such a conditional goal grounds a social preference*: *a dependent or conditional preference*.

Consider for example a pro-social husband that prefers going to the movie (instead of staying at home) with his wife *because* she prefers going to the movie (he would not prefer it but for her preference). Given the fact that the wife has been satisfied, she might decide to reciprocate. She could start to care for her husband by suggesting a movie that he happens to like (between two that she may be indifferent on). Two agents with social preferences in this dynamics could 'mesh' their choices better than purely selfish ones.

A situation in which the agents have *interdependent preferences* occurs when at the same time *both prefer something conditionally on the other preferring the same thing* (both have pro-social preferences).

The prototypical situation used to illustrate the structure of the matrimonial dilemma just discussed is usually represented within the following matrix:

		Wife	
		Staying at home	Going to the movies
Husband	Staying at home	2,1	0,0
	Going to the movies	0,0	1,2

Figure 3: The battle of the sexes

This game is known as the 'battle of the sexes', and is used to model interactions in which the two agents have a *common goal* (i.e. an interest in coordination) but *conflicting preferences about what to do together*. The interest in coordination derives from the fact that they both independently prefer to stay together above staying apart. Unilateral sacrifices in both senses (the husband following the wife or vice-versa) are Nash equilibria, and then, again, even in this game,

there is a coordination problem that is a problem of choosing the same profile of strategies at the same time.

The payoffs of this game however individuate a structure of interaction that is only superficially similar to my matrimonial scenario. Actually, in the standard battle of the sexes, the husband *independently* prefers to stay with the wife and to stay at home, and he reasons strategically in order to achieve his *individual* intention that they stay together. Moreover, we could say that the husband and the wife fighting in this battle do not actually *share* an intention to stay together, but simply have common knowledge of their reciprocal intentions 'that we stay together'.

To put it differently, the standard way of presenting this problem rules out the possibility of a *compromise* between the married couple.

However, very often, when the independent preferences of the parties don't align perfectly, this initial conflict happily ends in a compromise. Changing the example a little, if the husband and the wife *cared for each other* at the same time they could settle to go somewhere else, which is neither the movie nor staying at home. On the basis of their common knowledge that *they both also prefer what the other prefers*, they might find an agreement on a third option (e.g. on having dinner outside). The third option, the one on which they could converge, is preferred by them only *conditionally*, and neither of them has a preference independent from the other.

More generally, *finding a compromise necessarily implies having some interdependency between the preferences*, and so some form of *mutual goal adoption*. If a compromise is an agreement by mutual concessions of the parties, both parties ought to have social preferences to reach it. To 'concede' is in fact to adopt the goal of another agent when this adoption is costly (some other goal is frustrated), in order to satisfy a more important goal. Consequently, in concessions, an agent prefers something also because the other prefers it, and so mutual concession needs mutual social preferences to ensure that both parties pay some costs in order that both can obtain some benefits[18].

[18] In the standard battle of the sexes there is no actual *concession* but only unilateral *sacrifices*. Sacrificing oneself means, roughly, to be willing to pay some additional costs by giving up something valued in order to achieve another goal considered more important. Differently, in concessions, one actively promotes something that another agent wants.

Since structuring the relevant bargaining between the parties is one of the roles played by shared intentions, then common social preferences, which underlie the possibility to reach compromises, offer themselves as good candidates for our original quest. Common social preferences are interdependent preferences and fit nicely in the theory of shared intention as a support for one of the defining roles Bratman has identified as necessary.

7. Interdependent preferences are not enough

Consider, however, more carefully the example of the married couple. The structure of the payoffs illustrated in the standard battle of the sexes is better understood if we attribute to the players only independent (self-regarding) preferences. It is true that the husband intends to stay with the wife, but he does so in a self-regarding manner since, as I have alluded, he does not actually care 'for' his wife. If we assume that a functional couple gives some additional weight to what the other cares, then we are attributing social preferences to them.

More precisely, while with common self-interest preferences, each cares *about* the other (and this is common knowledge) with common social preferences each cares *for* the other (and this is common knowledge): when caring about the other, one is concerned with what the other can do, while *when caring for the other, one is also concerned with what the other cares* (i.e. what the other prefers).

If they have social preferences (each actually cares for the other), then the husband and the wife are actually playing a different game that can be illustrated as follows:

		Wife		
		Staying at home	Going to the movies	Going to the restaurant
Husband	Staying at home	2,2	1,1	0,-1
	Going to the movies	1,1	2,2	0,-1
	Going to the restaurant	-1,0	-1,0	3,3

Figure 4: The caring-for-each-other game

Assuming that the husband cares for his wife (cares about the satisfaction of her preferences, and not only about the impact of her choice on his own), he is motivated to go the movies also by the fact of contributing to her preference satisfaction: we can represent this additional motivation with an increase in utility, say of 1 unit. The same is true for his wife's payoff of staying at home. The standard conflict of interests is not part of this different game because, when they do care for each other, it is now indifferent for both of them to stay at home or to go to the movies.

However, we have supposed that a new third option exists (the compromise) in which they both would receive additional satisfaction (gaining each 1 unit of utility if they both made the same pro-social choice because they would care for each other at the same time (their pro-social motives or adopted goals are satisfied at the same time). Suppose, moreover, that they both might even like this option for some additional self-regarding reason. The husband, for example, prefers eating outside to watching a movie while the wife prefers going outside to staying at home as usual. But, under these additional assumptions, a new Nash equilibrium is now available where going to the restaurant is the best response for both and it is also a Pareto superior outcome. *Interdependent preferences can limit possible conflicts turning the need for mutual sacrifices into a more positive need for mutual concessions*.

However, it should also be clear at this point, that the married couple still have some difficulties because they are now facing the same Hi-Lo problem where R2D2 and C3PO are still stuck. Can they do better than the robots? *Do they have now an independent reason to choose Hi?*

Having interdependent preferences (a common social preference) is a step toward solving the problem but it is still not enough.

What the husband (the wife) is valuing for example is his (her) going to the restaurant with his wife (her husband), and that the same fact is valued by her (him) too. Because they are still in a strategic setting, however intertwined their preferences might be, the agents are still supposed to reason in a strategic way.

The husband could wonder whether his dependent preference to go the restaurant provides him an independent reason to go there. Unfortunately in a strategic setting the answer is still negative. The *interdependent preferences* still provide the agents with *interdependent reasons* for

doing an action. Actually, in case the wife went to the movie, he would again do better by going to the movie with her[19]: the interdependent reason provided by the interdependent preferences is again preventing a resolute choice.

Consider, however, this next move. *What if part of the problem laid in the way the agents are reasoning*? Is strategic reasoning really necessary to model this problem or is it part of the problem itself?

If the interdependency of their reasons is their ultimate problem, it may also be connected to the way the agents are *reasoning* about this situation. My strategy is, then, to justify why agents with a shared intention should not be modelled as strategic reasoners when confronted with a Hi-Lo kind of interaction.

8. This is not a game!

To evaluate this possibility we need to understand what strategic reasoning is, and which are the conditions that justify its attribution to the agents.

When an agent engages in strategic reasoning, she is aware of the fact that her desired outcome is partially dependent on the autonomous action of another agent. Since she treats the other as an intentional agent similar to her, she needs also to predict what the other will do by attributing beliefs, goals and preferences in order to choose her best action, knowing that the other is doing the same. If there is not a unique best reply for both, then the interdependency of the reasons for action creates the fatal ambiguities that originate the kind of coordination problems that are puzzling us. So, mutual strategic reasoning with a mutual "intentional stance" creates the problem (Dennett 1987).

Let's then unravel when an agent should endorse strategic reasoning to deal with a social interaction. By understanding this we can also identify the corresponding conditions that justify the

[19] In the matrix this fact is illustrated by assigning a negative payoff to the situation in which one goes to the option that is only interdependently preferred (the restaurant) while the other sticks to the private independent preferences. Going to the restaurant without the wife is more frustrating because it compromises two goals of his: being with his wife and staying at home.

modeller in attributing this kind of reasoning to the agents (i.e. the cognitive mechanisms behind overt behaviours).

Strategic decisions arise when at least two agents *interfere* with each other, i.e. the fulfilment of their individual goals is partially dependent on what the other does. If their interfering behaviours are not adequately taken into account, the agents will most probably fail to reach their goals. This objective fact poses a problem that can be solved at different levels of cognitive complexity, and not necessarily by reasoning. A learned or evolved set of heuristics, for example, can be a solution to iterated interferences. Quite paradoxically, were the others to have no reason to do an action but being at the same time reliably predictable by us, heuristics could exploit this fact making coordination simple and immediate (the best response would be more easily identifiable; Hurley 2005). If we considered the others as natural events and not as cognitive agents, we can give for granted their behaviours, and adapt to them just like we do with moving objects that we try to avoid[20].

Although this is possible, very often we need to appeal at least to a primitive form of 'mind-reading'. In traffic, for example, this happens when we treat the others as 'limited' cognitive agents, predicting their intentions without necessarily noticing that the others are doing the same with us.

However to understand the role of shared intentions in coordination problems, we need to focus on solutions at the deliberative level of action control, when the agent engages in *practical reasoning*, and they are aware that the others are doing the same. Strategic reasoning is, from this perspective, only a peculiar kind of practical reasoning when there is the practical necessity to care *about* another agent's choices.

But caring about another agent's choices can serve two distinct aims: the former is to *adapt* one's own action, the latter is to possibly *influence* the choices made by others (Castelfranchi 1997).

Those concerned with understanding strategic interaction have mainly focused on *adaptation*. Knowledge of each other instrumental rationality and preferences is everything an agent

[20] Game theorists distinguish two kinds of decisions: 'parametrical' and 'non-parametrical' decisions. When deciding parametrically, one is only concerned with variables that are independent from one's own intentions (the world is passive). Differently, a non-parametrical choice is one in which the "world" anticipates one's own course of actions, and so the agent should react also to this possibility in order to be successful.

is given to choose an option. More specifically, even if the agents have both beliefs and goals about others' behaviours, only their mutual social beliefs are relevant to choose what to do (i.e. the goals are simply ineffective). *Strategic reasoning then is necessary because it is the only way in which an agent can identify the best option to pursue*.

Influence, differently, aims to modify the choices of another agent by *changing* his beliefs and goals. If some pre-play interaction is allowed, agents can communicate or can change the environment such that, in the end, the best strategy for the others in the game might be different. One standard way to do so is by providing the others with different incentives to make them choose on a different payoff structure.

In any case, *when the agents are allowed to influence each other also the social goals are relevant*. Such goals mention the behaviour of others as goal-driven behaviour. When also the preferences are known, an agent can plan to influence the other knowing what the other is concerned in. In this case, *social motives (adopted goals) can be derived instrumentally in order to help the other satisfy her goals as means for one's own goals*. However, even when both instrumentally adopt the goals of each other, if unaware of both acting for the other at the same time, when they need to choose a strategy, the agents can only adopt strategic reasoning to select their own strategy[21].

But is strategic reasoning justified when the agents are disposed to care *for* each other and this is mutually known? When the agents are so disposed and they mutually know it, they are reasoning for themselves and for the others at the same time. They are still concerned in reaching their own goals but they are also caring for the others (maybe for instrumental reasons). They both are actually concerned that they both reach their own goals.

In this frame of mind, when an agent has to choose a course of action she will reasonably take into account the fact the other agent is reasoning also *for* her, and on this basis, she can decide to *collaborate*. A *collaborative* choice then is not seen by the agents as a strategic one. In other

[21] This is also true even if one of the agents is a real altruist (so that her pro-social motives are terminal) when thinking the other is not. The agent then is acting not to influence the other but to support her. Nonetheless she cannot do better than endorsing strategic reasoning even if it's done for the other.

words, *an interference problem in presence of mutually known pro-social* (either instrumental or terminal) *motives can be framed as a collaborative choice* (see Figure 5).

	Adaptation	Influence	Collaboration
Strategic Reasoning	Social beliefs	Social goals or unilateral adopted goals	-
Collaborative Reasoning	-	-	Mutual Adopted goals

Figure 5: Relevant attitudes in strategic vs collaborative interaction

But what is then the individual practical reasoning behind collaboration?

If strategic reasoning is individual instrumental reasoning given a constraint (the best available prediction of another intentional agent's action), *collaborative reasoning is individual instrumental reasoning based on an assumption of common pro-social motives* (adoptive goals). In collaborative reasoning one does not predict the best action of the other, and then selects one's own given this fact. Rather, a collaborating agent has acquired the evidence that the other is reasoning also for her, and on this basis chooses the best action (that is the best also for the other).

Hence, my suggestion is that on the background of a shared intention, *the agents have evidence that they are mutually caring for each other* (i.e. they mutually know that they have interdependent preferences), and so *they frame their interaction as a collaborative choice and not as a strategic one*. Differently put, when the agents are in this frame of mind, even if their interference with a Hi-Lo payoff structure is a game, it is not seen as a game by them[22].

9. The solution

Going back to the original example, we have assumed that both R2D2 and C3PO have common knowledge of their reciprocal independent preferences for a *blue* house.

[22] It may be suggested, at this point, that this kind of interaction is best modelled within the formal framework of 'cooperative' game theory. However, cooperative game theory abstracts from "how coalitions form and how their members choose joint actions" (Osborne and Rubinstein 1994: 256). However, these are precisely the problems we are dealing with in this paper.

Suppose however, that R2D2 and C3PO are indeed disposed to care for each other, i.e. they both have the goal to adopt each other's goals as far as *J* is concerned. Hence, they have interdependent social preferences towards the other as described earlier[23].

Remember that what keeps Bratman's approach to shared intentionality apart from other theories, is the focus on the *content* of intentions, and not on the mode or form in which such intentions are held by individuals. In order to share an intention between you and me, we don't need to think to ourselves as members of a group from *whose* perspective we intend something (versions of this idea underlie the theories of Searle, Tuomela, Gilbert and Bacharach). It is sufficient to be able to formulate normal individual intentions with a specific and interlocking content.

If non-standard preferences are allowed, then the interlocking intentions (condition 2) and the interdependent intentions to paint the house (condition 6) can be used to justify the assumption that the two robots are disposed to care for each other. Being dependent on the other agent, in the way conditions 2 and 6 specify, signals an underlying (at least instrumental) pro-sociality.

Given our previous analysis, when R2D2 cares for C3PO when aiming to paint the house together, the former robot has social preferences toward the latter, and this entails that some of R2D2's preferences depend on C3PO's preferences. The same is true for C3PO.

This disposition to adopt each other's goals to achieve the outcome is then the evidence the robots need to engage in collaborative reasoning. Even if their preferences seem independent (they like the blue paint for independent reasons), they know that each of them would be prepared to reconsider such preference in case of disagreement. Such counterfactual possibility makes their preferences actually interdependent.

We have now all the conceptual resources to dissolve the paradox of Hi-Lo games in the context of already established shared intentions.

Recall that both R2D2 and C3PO mutually know that, given their mutual disposition to care for the other in achieving the joint outcome, their reciprocal ranking is:

Outcomes	R2D2	C3PO
Blue House	2	2

[23] This is one possible way to analyse what the condition that the agents have an intention to help each other (condition 4) actually imply. However, as I will argue, this condition is implicit in both conditions 2 and 5.

Yellow House	1	1
Mixed-color House	0	0

Hence, on the basis of a shared intention and of the disposition to care for each other, each of them is now *independently* justified in the following piece of individual practical (but *collaborative*) reasoning:

I, R2D2, intend that we paint a house together

I, R2D2, care for C3PO in painting the house

I, R2D2, believe that the blue paint is the best colour for me *and also for C3PO*

I, R2D2, intend that we paint a *blue* house

When they both know that both are disposed to care for each other, they can safely derive the corresponding intention, and proceed in their parallel planning.

Although their interdependent preferences only provide an interdependent reason to choose the *blue* paint over the *yellow*, this is a real problem *only if they take themselves to be in a strategic interaction*. However, *their disposition to care for each other is evidence that each would change their preference, if the other had a different one*. Hence, this counterfactual possibility justifies their collaborative choice.

If one requires additional justification for endorsing collaborative reasoning, one also needs some justification for appealing to strategic reasoning which in this setting is not indeed intuitive.

Moreover, even if social in their scope (i.e. these pieces of reasoning mention also what is the best option for the other), they are conducted *from their individual perspectives* without appealing to any identification with the group they are forming. Psychologically, *both of them are individually reasoning also for the other* without the need of a we-mode kind of thinking from the group point of view.

Eventually, from a functionalist perspective, their parallel reasoning 'realize', at a different level, *a form of shared practical reasoning* so that the following schema of practical reasoning can be attributed to the group they are forming:

Shared practical reasoning (emergent):

We intend to J

M is the best means for us to J

We intend to M

Even if each of them can derive, from the individual intention that we paint a house, the individual intention that we paint a *blue* house, they are at the same time *collectively* deriving the *shared* intention to paint a *blue* house from the shared intention they were committed to before[24].

10. How much ice has been cut?

It seems that what looked like a relevant and difficult problem has been now solved in a simple way. Recall that "paradox" of Hi-Lo games is that the standard theories of interactive choice are unable to suggest Hi-Hi as the solution for this kind of interactions even if choosing Hi is strongly intuitive. But an adequate explanation of the choice of Hi must explain both why it is *obvious* for us, and why it is also the *correct* (and so rational) choice to make.

Within my perspective, choosing Hi is an *obvious* choice because it is natural for us to be disposed to have adoptive social goals and, hence, social preferences. We humans are the only species in the biological realm to be disposed to care for each other, even for unrelated others (Tomasello et al. 2005)[25]. Being disposed to take the other into account in one's own planning is what makes us unique in the biological realm. Nowadays, we start to have explanations for this uniqueness (Gintis et al. 2005) that is puzzling only from an evolutionary perspective, where understanding the mechanism by which natural selection might give origin to pro-social psychologies is the problem. Our common sense tells us that *we simply are collaborative* (at least quite often) in the sense of caring one for the other. What is less obvious is that *by being other-regarding we are capable of solving many problems that are unsolvable for merely selfish agents*.

This explanation also takes into account that choosing Hi (in a Hi-Lo coordination problem) is more immediate (and obvious) then finding the compromise in the matrimonial dilemma sketched above. While in both situations interdependent preferences have a role, the compromise is

[24] *Commitment* at the group level here is used in the same sense in which an agent, when intends to do something, is *committed* to act as intended. This notion of commitment does not entail any kind of deontic duty or social obligations towards the others (Bratman 1987 and 1999).

[25] Tomasello and colleagues hypothesize that only humans have a "motivation to share psychological states with others" (2005: 675).

cognitively more demanding because the agents need to 'find' something in common which is not already there in front of their eyes.

The second criterion of adequateness requires a justification of the *correctness* of the underlying reasoning. As we have seen, the reasoning that the agents are endorsing, *individual collaborative reasoning*, *is just standard individual practical reasoning*, and hence as far as practical reasoning is valid, collaborative reasoning is valid too.

On the basis of these reasons, I think it is justified to claim that, on the background of a shared intention and of interdependent preferences, *the agents do not frame their choice as a strategic one*. A strategic choice is one in which an outcome I favour depends on your choice. Strategic choices are choices with common knowledge of reciprocal payoffs and mutual dependence as for the realization of an outcome. Strategic agents are agents that simply interfere one with the other, and need to predict each other actions to achieve their individual goals (strategic agents merely care about each other choices). Common knowledge of all this makes their reasons for choosing a course of action interdependent. When agents care for each other at the same time, however, they do not simply take the action of the other as a given constraint, but they are disposed to proactively adopt the goals of the other as one's own. Agents in this context frame the problem as a collaborative one and not simply as strategic.

As we have seen, if the agents have an interdependent preference for an outcome (if it is an outcome that I prefer and you prefer, and we realize that we would not prefer it if the other didn't prefer it) then that outcome is up to our collaborative choice, and not only to my-choice-given-yours. By knowing each other preferences, and that we are disposed to adopt each other goals, each of us can infer in parallel a new more specified intention, and can safely act on it (even when evaluated from the point of view of rationality).

In the end, on the background of a shared intention, the choice of Hi is both a preference and a reasoning effect in the sense that it presupposes both non-standard preferences (social preferences

based on pro-social motives[26]) and a framing of the situation as a collaborative choice calling for usual individual practical reasoning (although social in its scope).

Hence, contrary to Bacharach's claim that assuming social preferences "cuts little ice in any pure coordination problem" (2006: 111), I have shown that commonly known pro-social motives may be helpful to "melt" the slippery surface of a strategic game with interdependent reasons for action.

11. How to be together in intending a joint action

Let's take stock. The Hi-Lo problem has been introduced to cast doubts that Bratman's analysis of a shared intention was sufficient. Within standard assumptions about preferences and reasoning, I have argued that it is not, because two of the defining roles putatively fulfilled by a shared intention are not actually satisfied. More precisely, if one assumes standard self-regarding preferences and standard strategic reasoning, an intention that is shared by two agents in Bratman's sense does not ensure that their actions will be in coordination (coordinating the agents in choosing Hi). Moreover, sharing an intention cannot effectively structure relevant bargaining between them (coordinating them on finding a compromise when this is required). Facing these limitations a functionalist is forced to admit that a shared intention is not in fact realized. However, I have also suggested that, by making explicit that an individualistic analysis à la Bratman presupposes interdependent preferences and a capacity for goal adoption, and by offering an alternative to strategic reasoning (collaborative reasoning), it is possible to cope with this impasse and, being coherent with the original account, it is also a means for rescuing it.

In the introductory remarks I have mentioned two distinct worries. Having tackled the first (i.e. it is possible to derive a more specific shared intention from a general one), it is now necessary to meet the other one. The second worry was that forming of a shared intention is itself solving a coordination problem (and very often a Hi-Lo one). This objection is still to be addressed given

[26] See, for instance, Gintis (2003: 161). The literature of social preference is, however, only focused on *terminal* motives: i.e. altruistic motives. I don't see any special reason why one cannot be collaborative when aiming to a joint outcome, if this is the best way to achieve this outcome.

that, for the time being, I have assumed that the agents already shared a certain intention. Moreover, it is such knowledge of their reciprocal intentions that was also used as evidence by the agents that they are caring for each other, something that also justifies the endorsement of collaborative reasoning.

Are the agents similarly justified when an intention to do a joint action is not already shared?

Such objection questions the interdependency of the intentions that we do *J* (condition 6 in the definition of shared intention) because it reveals a problematic interdependency between the reasons for choosing to pursue the joint action[27]. If in order to form a shared intention to *J*, I intend that we *J if and only if* you intend that we *J*, what is the independent reason in the end to choose *J*, and not, say, the alternative joint option *M*? How can I independently intend that we *J* if my reason for choosing *J* is interdependent with yours?

Consider again R2D2 and C3PO, and assume now that they are not already sharing the intention to paint the house together. If they face the choice between painting a house or going to the movies (knowing that they both prefer to have a painted house), they are again facing the problem of coordinating their choices of what they already know to be the best option for both. Without communicating, how can they resolve to choose the outcome preferred by both?

Since they both know their reciprocal preference, they know that both have the *goal* that they paint the house. Their reasons for having this goal need not be interdependent, because each of them may well have independent and self-regarding reasons for it. At this point, the only further assumption that is needed is that they both know that they will adopt each other relevant goals in doing *J*. With this assumption, and on the basis of the arguments developed in the previous section, I claim, analogously, that *it is natural for them to frame this situation as a collaborative choice and not as a strategic one*. And, although such dispositions are not a reason to choose a specific joint option, *they explain why they see the problem as a collaborative choice*. If they do not reason in a strategic manner, having only interdependent reasons for their choices is not a problem. They can

[27] Bacharach raises this objection precisely to conclude that Bratman's account of shared intention is flawed because it is impossible to infer these interdependent intentions from an individual perspective (2006: 138). In a similar vein, David Velleman (1997: 35) also asks: "how can I frame that 'we' are going to act, if I simultaneously regard the matter partly up to you?".

choose the option of 'collaborating in painting the house', if they are confident that the each of them is reasoning also for the other one, and know that that joint option is the *instrumentally best* for both.

Even in the absence of a previous shared intention, they are justified in forming the intention that we paint a house on the basis of the interdependent preference for that outcome (provided that they take each other to be 'collaborative').

Although a Pareto-superior equilibrium in a coordination problem cannot be selected from an individual perspective if one reasons strategically, when the agents strive for collaboration, at least mentally, they are not making a strategic choice. They can easily choose the option that is best for both by reasoning collaboratively. Dispositions to care for each other and the social preferences they ground make this possible.

12. Shared intention revisited

One last issue to be explored is whether the pro-social activity of caring for each other should be seen as a 'constitutive' condition in the analysis of shared intention.

Above I have discussed the inadequacy of condition 3 in Bratman's analysis (intending that we *J* in part by way of sub-plans that mesh with the sub-plans of the other) for enabling the robots to select a more specific course of action when choosing between two (although ranked) meshing sub-plans. If interpreted within standard assumptions, Bratman's shared intention cannot solve the Hi-Lo problem directly because a shared intention is mute in these situations. The meshing-subplans condition has been introduced precisely to establish a background against which relevant bargaining can take place (Bratman 1999: 121) in order to exclude situations in which each of us intends that we paint the house, but you intend to paint it blue all over and I intend to paint it yellow. When we share an intention we should be "prepared to compromise" (120), otherwise we would not achieve coordination. I have also argued that only common social preferences enable a compromise, and that a preference in favour of the compromise is an interdependent preference. Hence to be "prepared" to compromise, one needs to attribute to the agents *social preferences*.

Finally I have contended that having these preferences, presupposes the disposition to adopt each other relevant goals in *J*.

Such disposition, however, is not enough: when two agents *share* an intention, they are not only reciprocally disposed to adopt each other goals relevant for *J, they are committed to do it.* That is, in order to share an intention, all the participants must also have an *intention that we adopt each other goals relevant for J-ing.* And this is so, because *an intention in favour of this joint cognitive process – mutual goal adoption - has to play the role that every intention plays*: it should filter new inconsistent options out of deliberation (e.g. a specific sub-plan that does not mesh with the plan chosen by the other agent), it is input for additional means-end reasoning (e.g. for means or sub-plans that satisfy the goals of both), and it commits each of them to identify sub-plans in the future in a way that resist constant revision of previously plans satisfying the goals of both.

Granted all this, I propose a revisited definition of a shared intention (SI-Rev):

We intend to *J* if:

(1)	I intend that WE *J*	You intend that WE *J*
(2)	I intend that WE *J* IN PART BY WAY OF YOUR INTENTION THAT WE *J*	You intend that WE *J* IN PART BY WAY OF MY INTENTION THAT WE *J*
(3*)	I intend that WE ADOPT EACH OTHER RELEVANT GOALS FOR *J-ING*	You intend that WE ADOPT EACH OTHER RELEVANT GOALS FOR *J-ING*
(4)	I believe that IF WE CONTINUE TO INTEND THAT WE *J* AS A RESULT WE WILL *J*	You believe that IF WE CONTINUE TO INTEND THAT WE *J* AS A RESULT WE WILL *J*
(5)	I believe that I INTEND THAT WE *J* IF AND ONLY IF YOU INTEND THAT WE *J*	You believe that YOU INTEND THAT WE *J* IF AND ONLY IF I INTEND THAT WE *J*
(6)	All the conditions above (including this one) are common knowledge between me and you	

SI-rev is close in spirit to Bratman's analysis because the crucial condition 3* (caring-for condition) is interlocking in its content too. The caring-for condition is clearly also a generalization of the original disposition to help the other in fulfilling his role, though is aimed to do much more than this.

It is important, once again, to warn that being pro-socially motivated does not amount to being altruists. It is sufficient that *agents that share an intention to J, are pro-socially motivated as far as the pursuit of J is concerned.* Both of them can have very different (and even conflicting) reasons behind the decision to take into account the goals of the other. But once they come to this decision, they can have instrumental social preferences. But, being prepared to reason for the other, and knowing that the other is prepared likewise, I think may also explain the "sense of collectivity" that is felt by those who jointly act together (Searle 1990).

13. Conclusion

With the aim of exploring how to best understand core cases of joint intentional action, I have begun this paper by questioning Bratman's theory of shared intention. To this aim I have employed a problem (the Hi-Lo paradox) emerged in the context of game theory. I consider the Hi-Lo problem as a very elegant way to identify the limits of strategic reasoning (one of the basic assumptions behind game theory[28]), and not of individualistic accounts of agency (as it is usually contended). Team reasoning and strategic reasoning in interactive decision problems, as they are usually interpreted, are necessary to understand group-regarding and self-regarding behaviour. Part of the aim of this paper was precisely to defend the view that, *even at the level of agency of the individual, there can be other-regarding attitudes that are necessary to understand relations between the agents.* Pro-social motives and interdependent preferences are means by which we glue each to the other, in a form of push-me/pull-you dynamics sustained by collaborative reasoning.

[28] Quoting Osborne and Rubinstein: "The basic assumptions that underlie [game] theory are that decision makers pursue well-defined exogenous objectives (they are rational) and take into account their knowledge or expectations of other decision makers' behavior (they reason strategically)" (1994: 1).

An individualistic approach to joint intentional action is from this perspective necessary to ground more complex forms of group action (Castelfranchi 2003). Without detailed models of these more basic phenomena, we are not able to understand 'helping' or 'punishing' behaviours that participants pro-actively take one towards the other. Individuals that group-identify are willing to sacrifice themselves for the group, while individuals that care each for the other are also disposed to support the others in pursuing a joint endeavour. To paraphrase a statement by Susan Hurley (2003) then, I hope to have shown that *the limits of egoism are not the limits of individualism.* It is the individual that intends that more than one agent does a joint option, and it do so from his individual perspective. What is special is the way such joint option is conducted. *Individuals are together in intending a joint option when they reason collaboratively during action execution.*

References

Aumann, R. J. & Sorin, S. (1989) Cooperation and bounded recall, Games and Economic Behavior, 1: 5-39.

Bacharach, M. (2006) Beyond Individual Choice. Teams and Frames in Game Theory, Ed. by N. Gold and R. Sugden, Princeton, NJ: Princeton University Press.

Bowles, S. (2004) Microeconomics. Behavior, Institutions and Evolution, Princeton, NJ: Princeton University Press.

Bratman, M. (1987) Intentions, Plans and Practical Reason, Cambrigde, MA: Harvard University Press.

Bratman, M. (1999) Faces of Intention, Cambridge: Cambridge University Press.

Bratman, M. (2007) Structures of Agency, Oxford: Oxford University Press.

Castelfranchi, C. (1997). Individual social action. In G. Holmstrom-Hintikka and R. Tuomela (eds.), Contemporary Theory of Action, Vol. II, Kluwer, Dordrecht, pp. 163-192.

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, Kluwer Academic Publishers.

Castelfranchi C. (2006) From conversation to interaction via behavioral communication. In S. Bagnara & G. Crampton Smith (Eds) Theories and Practice in Interaction Design, New Jersey (USA): Erlbaum, pp 157-179.

Conte, R. & Castelfranchi, C. (1995) Cognitive and Social Action, London: UCL Press.

Dennett, D. (1987) The Intentional Stance, Cambridge, MA: MIT Press.

Farrell, J. (1988) Communication, coordination and Nash equilibrium, Economics Letters, 27, 209-214.

Fehr, E. & Fischbacher U. (2002) Why Social Preferences Matter. The Impact of Nonselfish Motives on Competition, Cooperation, and Incentives, Economic Journal, 112, C1-C33.

Gauthier, D. (1975) Coordination, Dialogue, 14: 195-221.

Gilbert, M. (1989) On Social Facts. Princeton, NJ: Princeton University Press.

Gintis, H. (2003) A critique of team and Stacklberg reasoning, Behavioral and Brain Sciences, 26(2): 160-161.

Gintis H., Bowles S., Boyd R., and Fehr E. (2005) Moral Sentiments and Material Interests: On the Foundations of Cooperation in Economic Life, Cambridge: The MIT Press.

Hodgson, D. H. (1967) Consequences of Utilitarianism. Oxford: Clarendon Press.

Hollis, M (1998) Trust within Reason, Cambridge: Cambridge University Press.

Hurley, s. (2003) The limits of individualism are not the limits of rationality, Behavioral and Brain Sciences, 26(2): 164-165.

Hurley, S. (2005) Social heuristics that make us smarter, Philosophical Psychology, 18(5): 585-611.

Jeffrey, R. (1977) A note on the kynematics of preference, Erkenntnis, 11: 135-141.

Osborne, M.J. & Rubinstein, A. (1994) A Course in Game Theory. Cambridge, MA: MIT Press.

Roy, O. (2009) Intentions and interactive transformations of decision problems, Synthese, 169, 335-49.

Searle, J. (1990) Collective intentions and actions. In P. Cohen et al. (Eds) Intentions in Communication, Cambridge, MA: MIT Press, pp. 401-415

Schelling, T. (1960) The Strategy of Conflict, Cambrigde, MA: Harvard University Press.

Sugden, R. (1993) Thinking as a team: toward an explanation of nonselfish behavior, Social Philosophy and Policy, 10: 69-89.

Sugden, R. (1995) A theory of focal points. Economic Journal, 105: 1269-302.

Sugden, R. (2000) Team Preferences, Economics and Philosophy, 16: 175-204.

Tomasello M., Carpenter M., Call J., Behne T. and Moll H. (2005) Understanding and sharing intentions: the origins of cultural cognition, Behavioral and Brain Sciences, 28: 675-735.

Tuomela, R. (1995) The Importance of Us: A Philosophical Study of Basic Social Notions, Stanford, CA: Stanford University Press.

Velleman, J. D. (1997) How to share an intention, Philosophy and Phenomenological Research, 57: 29-50.