Forthcoming in *Journal of Social Ontology*

Team reasoning and intentional cooperation for mutual benefit

Robert Sugden

School of Economics, University of East Anglia
Norwich NR4 7TJ, United Kingdom
r.sugden@uea.ac.uk

5 October 2014

Acknowledgements I thank an anonymous referee for helpful comments. My work was supported by the Economic and Social Research Council (award no. ES/K002201/1.

Abstract This paper proposes a concept of intentional cooperation for mutual benefit. This concept uses a form of team reasoning in which team members aim to achieve common interests, rather than maximising a common utility function, and in which team reasoners can coordinate their behaviour by following pre-existing practices. I argue that a market transaction can express intentions for mutually beneficial cooperation even if, extensionally, participation in the transaction promotes each party's self-interest.

In one of the most famous passages in *The Wealth of Nations*, Adam Smith (1776/ 1976, pp. 26–27) asks us to think about our relationships with our butchers, brewers and bakers when we provide ourselves with our dinners. According to Smith, we do not appeal to the butcher's humanity or benevolence as a reason for him to supply us with meat. Instead, we talk to him about the advantages that *he* will gain by trading with us. The implication of this remark is that each party to a market transaction views it as a means of promoting his individual interests. Of course, it is essential to Smith's account that both parties *in fact* benefit from the transaction, but their mutual benefit is not *intended* by either of them. Each party's intention is his own benefit; from his point of view, the other's benefit is an unintended consequence. This understanding of market relationships is deeply embedded in modern economics. Economists have usually agreed with Smith that the role of self-interest in the workings of the market is not a matter for regret. However, virtue ethicists such as Anderson (1993) and Sandel (2012) invoke the same understanding of market relationships to argue that the market – however useful it may be when confined to its proper place – is a morally impoverished domain whose values and motivations are liable to corrode the virtues of other spheres of social life.

Bruni and I have argued that the intentions of market participants need not be construed as self-interested (Bruni and Sugden, 2008, 2013). To the contrary, we argue that it is possible for trading partners to intend that their transactions are mutually beneficial; and we suggest that many people *do* approach market transactions with intentions of this kind. Such people do not have to show the sort of benevolence that Smith thought was uncharacteristic of shopkeepers' attitudes to their customers. Nor, in normal circumstances, do they have to behave in ways that deviate from received economic theories. They simply intend to play their parts in bringing about the mutually beneficial outcomes that are the normal consequences of market transactions. Nevertheless, it is a morally important question whether the market should be thought of as a domain of self-interest or as one of intentional cooperation for mutual benefit. On the latter view, the market is not a reservoir of amoral attitudes that are in danger of spilling over and corrupting civil society: it is an integral part of that society.

These arguments, which I will not rehearse any further, provide the context for the current paper. In this paper, my objective is to propose a general concept of *intentional*

cooperation for mutual benefit that can encompass ordinary market transactions carried out with everyday goodwill on both sides. I develop this idea through a new formulation of team reasoning.

1. Two games

In presenting my analysis, I focus on two simple games, defined in terms of the strategies or moves available to the players and the resulting payoffs. Before presenting the games, I need to explain what I mean by 'payoff'.

Each player's payoffs are to be interpreted as normalised measures of the values of the relevant outcomes to her, *in terms of her own interests, as judged by her*. I will follow the conventions of classical game theory in not attaching any formal significance to interpersonal comparisons of payoffs.[1] However, in thinking intuitively about particular games, it will often help to think of differences in payoffs as corresponding with differences in individuals' holdings of some commodity (for example, money) that is universally valued.

My interpretation of payoffs in terms of interests differs from the one that has traditionally been used by game theorists. On the latter interpretation, each player's payoffs are utility indices in the sense of von Neumann and Morgenstern (1947), representing all-things-considered preferences that are assumed to satisfy the axioms of expected utility theory. 'All things considered' is often taken to imply that each individual's choices are determined by her preferences, or that her preferences are revealed in her choices. But this approach requires, as a matter of conceptual necessity, that an individual's pro-social motivations can always be represented as payoffs to her as an individual. One of the fundamental intuitions of theories of team reasoning is that this way of thinking about motivation is too restrictive. In interpreting payoffs in terms of interests, I do not presuppose that each player acts in the way that maximises her (expected) payoff. A player who does act in this way (given my interpretation of 'payoff') will be called *self-interested*. A player who does not will be called *non-self-interested* or, for short, *non-selfish*.

My first game, the *Trust Game*, is now one of the paradigm games of the literature on social preferences. In its modern manifestation, it is due to Berg, Dickhaut and McCabe

[1] Since my analysis will not make use of mixed strategies, it is sufficient to interpret payoffs as ordinal representations of each player's interests. But, on the analogy of another convention of classical game theory, one might wish to interpret the cardinal properties of payoffs as representing players' attitudes to risk *in the context of judgements about their interests*.

(1995), but it has a far longer history: versions of the game are analysed in Hobbes's *Leviathan* (1651/ 1962, Chs 14–15) and in Hume's *Treatise of Human Nature* (1740/ 1978, pp. 520–521). The version I will use is shown in Figure 1.

[Figure 1 near here]

The numbers shown in the figure represent the possible payoffs of the game to the two players, A (listed first) and B (listed second). A moves first, choosing between *hold* and *send*. If he chooses *hold*, the game ends, with a baseline payoff of zero for each player. Intuitively, A's choice of *send* can be visualised as the action of investing one unit of money in an activity which will generate a net surplus of four units. If A chooses *send*, B then chooses between two alternative distributions of the costs and benefits of this activity. If she chooses *keep*, A loses his investment to B and, in addition, B gains the whole of the net surplus. If B chooses *return*, A's investment is returned and the net surplus is divided equally between the two players.

If both players act on self-interest and if each knows that this is true of the other, the outcome is (0, 0). (If A were to *send*, B would *keep*; knowing this, A chooses *hold*.) However, it is a matter of common experience (amply confirmed by experimental evidence) that in situations of this general kind, individuals in A's position sometimes choose *send*, and individuals in B's position sometimes respond by choosing *return*. Intuitively, it seems natural to say that the strategy combination (*send*, *return*) is a practice of *trust*. In choosing to *send*, A trusts B to *return*; in choosing to *return*, B reveals herself as trustworthy by repaying A's trust. For my present purposes, the problem is to firm up this intuition. What exactly does it mean to say that *send* is an act of trust, and how does its being such an act motivate A to choose it? And what does it mean to say that *return* is a repayment of trust, and how does that motivate B to choose it?

Before trying to answer these questions, I present my second game, the *Market Game*. This is shown in Figure 2. The difference between this and the Trust Game is that B's payoff from (*send, keep*) is –1 rather than 5. Thus, if A chooses *send*, it is in B's interest as well as A's that B chooses *return*. So if both players act on self-interest and if each knows that this is true of the other, A will choose *send* and B will choose *return*. This combination of actions is mutually beneficial, just as it is in the Trust Game, but one might not want to call it a practice of trust.

[Figure 2 near here]

Why do I call this the 'Market Game'? Suppose that A is Smith's baker and that B is his customer, wanting bread for her dinner. The baker has displayed various loaves of bread, with labels showing their prices. The customer asks for a particular loaf. The baker wraps it and hands it over the counter to the customer. She takes it and then hands over coins equal in value to the price. We might model the final stages of this interaction as a game in which A chooses whether to hand over the bread (*send*) or not (*hold*), and if A chooses the former, B chooses whether to hand over the money (*return*) or to run out of the shop without paying (*keep*). In normal circumstances, the rankings of payoffs for each player will be as in the Market Game. Relative to the baseline of not trading, the exchange of the bread for the money is mutually beneficial. If the customer tries to avoid paying, the baker will certainly be inconvenienced, but it is very unlikely that the expected benefits to the customer will exceed the expected costs. (She might be caught and punished; the baker will probably refuse to deal with her again; her action might be observed by third parties whose trust she may later want to rely on.) The point of this story is that everyday market transactions often have the structure of the Market Game.

Of course, one can imagine variants of this story in which the interaction between potential trading partners is better modelled by the Trust Game. For example, Akerlof (1982) argues that this is sometimes true of interactions between employers and workers. By paying more than the workers' reservation wage, the employer signals her expectation that they will exert more than the minimum level of effort necessary to keep their jobs; the workers respond by behaving as the employer expects. Economics needs to be able to explain the prevalence of mutually beneficial behaviour in interactions like the Market Game *and* the fact that such behaviour is at least sometimes found in interactions like the Trust Game. I will argue that both kinds of practice can express intentions for mutually beneficial cooperation.

2. Trust and social preferences

In economics and game theory, two default assumptions are often treated as unproblematic. The first of these is that each individual is *individually rational* – that is, has a preference ordering over all relevant outcomes and seeks to maximise the utility function that represents those preferences. The second assumption is that each individual's preferences are *interest-based* – that is, correspond with her interests, as she judges them. An observation of (*send*, *return*) in the Market Game would normally be explained in terms of these two assumptions,

combined with some additional assumption about individuals' knowledge or beliefs, sufficient to imply that player A expects player B, when choosing between *keep* and *return*, to be individually rational and to have interest-based preferences.

When it is necessary to explain non-selfish behaviour, the standard practice is to retain the assumption of individual rationality and to give up that of interest-based preferences. Individuals are instead assumed to act on *social preferences* – that is, preferences that take some direct account of other individuals' payoffs, beliefs or intentions. In the literature of social preferences, it is a common practice to model the outcomes of games in terms of the players' *material payoffs* – that is, increments of some universally valued commodity. Theories of social preference are based on hypotheses about how players' all-things-considered preferences relate to their own and others' material payoffs, and to other relevant features of the game. A common feature of these hypotheses is that an individual who acts on a social preference is willing to incur some material loss to achieve a socially-oriented end, such as reducing inequality between herself and others, rewarding others for acting on good intentions, punishing others for acting on bad intentions, or avoiding violating social norms. The *utility payoffs* of a game can then be defined as a representation of players' all-things-considered preferences. Standard game-theoretic modes of analysis are applied to the game, defined in terms of its utility payoffs.

However, it is very difficult to find a psychologically plausible and non-trivial specification of social preferences that can explain practices of trust, such as (*send*, *return*) in the Trust Game. I say 'non-trivial' because, given any observed behaviour in any specific game, it is always possible to assume that the relevant player has a preference for behaving in exactly that way. Or, equally trivially, it is always possible to assume that behaving in that way is prescribed by a game-specific social norm, and that the player has a preference for conforming to this norm or for avoiding the sense of guilt associated with violating it. Recall that I have set myself the problem of explaining what it means to say that (*send*, *return*) is a practice of trust, and how this fact can motivate players to choose these strategies. This problem is not resolved merely by asserting that players choose *send* and *return* because they prefer to do so, or that these choices are prescribed by a social norm.

Consider a Trust Game in which A chooses *send* and B chooses *return*. Suppose we want to explain this observation in terms of individual rationality and social preferences. At first sight, it might seem that the only real problem is to explain why B chooses *return*, since if A expects this, it is in his self-interest to choose *send*. One possibility is to invoke a theory

of social preferences in which each player's utility is a function of the profile of material payoffs to the two players. For simplicity, assume that the payoffs shown in Figure 1 are material payoffs as well as measures of individual interest. Then *return* would be individually rational for B if her utility from (2, 2) was greater than her utility from (–1, 5), which would be the case if she were sufficiently altruistic or if, as in the models of social preferences proposed by Fehr and Schmidt (1999) and Bolton and Ockenfels (2000), she were sufficiently averse to being on the advantageous side of inequality. The problem with this explanation is that it makes no reference to the action by A that preceded B's decision, and so cannot represent the intuition that B is *repaying* a previous act of trust. It is well-established experimentally that, in two-player sequential games in which each player moves only once (if at all), the behaviour of second movers is influenced by the payoff profiles that have been made infeasible by the first mover's decision (e.g. Falk, Fehr and Fischbacher, 2003).

So if a satisfactory social-preference explanation of B's choice of *return* is to represent this as the repayment of truest, it has to make B's preferences over (–1, 5) and (2, 2) conditional on some factor that can be activated by A's choice of *send*. There are two obvious possibilities – that A's choice reveals something about his beliefs, and that it reveals something about his intentions.

As far as beliefs are concerned, it is natural to say that A's choice of *send* is evidence of his belief that B will choose *return*. (It is not conclusive evidence, because a sufficiently altruistic A might prefer (–1, 5) to (0, 0), but let us leave that possibility aside.) We might hypothesise, following Pelligra (2005), Bacharach, Guerra and Zizzo (2007) and Battigalli and Dufwenberg (2007), that if (in B's belief) A believes that B will act in a way that will benefit A, B has a preference for confirming that expectation – or, which comes to the same thing, has a preference for avoiding the sense of guilt associated with disconfirming it. This would allow us to explain B's *return* as a response to A's *send*, rather than as an unconditional act of altruism. But consider the Confidence Game, shown in Figure 3. In this game, too, A's choice of *send* is naturally interpreted as signalling his belief that B will choose *return*; and B's choice of *return* would clearly benefit A. But there is a fundamental difference between the two games: in the Trust Game, (*send*, *return*) is mutually beneficial, but in the Confidence Game it benefits A at B's expense. B might reasonably think that A's expectation of *return* in the Confidence Game is gratuitous, and that to confirm that expectation would be to reveal her susceptibility to a confidence trick rather than her

trustworthiness. The implication is that trustworthiness is something more than conforming to other people's expectations.

[Figure 3 near here]

So perhaps the crucial feature of A's choice of *send* in the Trust Game is what it signals about his intentions. The idea that people care about other people's intentions is a common theme in the literature of social preferences. In this literature, it is a standard modelling strategy to follow Rabin (1993) in characterising intentions as *kind* or *unkind*. Each player's intentions are defined in terms of the payoff profiles that his actions can be expected to induce, given his beliefs about the other player's actions.

To get a feel for the underlying idea, consider the simultaneous-move Dilemma Game shown in Table 1. (Since I want to leave open the possibility that an individually-rational player would choose *cooperate*, I have resisted the temptation to call the game a *Prisoner's* Dilemma.) As before, assume that the payoffs in this game are material payoffs as well as measures of individual interest. Suppose that Row expects Column to choose *defect*. Given this belief, Row's choice is between the payoff profiles (–1, 2) and (0, 0). Since (–1, 2) is better for Column and worse for Row than (0, 0), a choice of *cooperate* by Row would reveal Row's kind intentions. (He has been kind because he has chosen to take a smaller payoff than he could have done, in a context in which this choice benefits Column. In the language of economics, he has had the opportunity to make a trade-off between his payoffs and Column's, and has chosen a point on the trade-off frontier that is relatively favourable to Column.) Conversely, a choice of *defect* by Row would reveal unkind intentions. (He has chosen to take a larger payoff than he could have done, in a context in which this choice harms Column.) Now suppose instead that Row expects Column to choose *cooperate*. A similar argument shows that, in this case too, Row would reveal kind intentions by choosing *cooperate* and unkind ones by choosing *defect*. Because the game is symmetrical with respect to the players, Column's intentions have the same properties. Rabin's crucial assumption is that each player derives utility not only from her own material payoffs, but also from what I shall call *emotional reciprocity* – being kind to co-players whose intentions are kind, and unkind to those whose intentions are unkind. It is easy to see that in the Dilemma Game, (*defect*, *defect*) is always a Nash equilibrium, but that if Row and Column have sufficiently strong preferences for emotional reciprocity, (*cooperate*,

cooperate) is a Nash equilibrium too.[2] In this game, therefore, intention-based social preferences can support mutually beneficial non-selfish behaviour.

Table 1: The Dilemma Game

		Column's strategy	
		cooperate	*defect*
Row's strategy	*cooperate*	1, 1	–1, 2
	defect	2, –1	0, 0

But now consider the implications of applying the same specification of social preferences to the Trust Game. Can there be a Nash equilibrium in which A is certain to choose *send* and B is certain to choose *return*? To see that the answer is 'No', suppose that A knows that B will choose *return*, and that B knows this. A's choice is then between (0, 0), which would result from *hold*, and (2, 2), which would result from *send*. According to Rabin's definitions, choosing (2, 2) rather than (0, 0) is neither kind nor unkind. Kindness and unkindness are revealed in the trade-offs that a player makes between his payoff and that of his co-player; to show kindness he has to incur some loss of material payoff – that is, to act contrary to self-interest – in a context in which this benefits his co-player. Since (2, 2) is better for both players than (0, 0), questions about kindness and unkindness do not arise. So, were A to choose *send*, that choice would not induce in B any positive or negative emotional reciprocity. Thus B would act on self-interest and choose *keep*, contrary to the initial supposition.[3]

This conclusion may seem paradoxical, but it reflects the fundamental logic of a modelling strategy in which socially-oriented motivations are represented as non-selfish preferences (that is, preferences that are not interest-based) acted on by individually-rational players. It is an essential feature of (*send*, *return*), understood as a practice of trust, that both

[2] A strategy profile is a Nash equilibrium if each player's strategy is optimal for her, given that the other strategies in the profile are chosen by the other players.

[3] If one takes account of mixed strategies, it is possible for there to be a Nash equilibrium in which A plays *send* with certainty and in which B plays *return* with some probability that is positive but less than 1/3. In such an equilibrium (if it exists), A's choice of *send* is kind, and so B derives utility from reciprocating this kindness. But it is still paradoxical that the certainty of trust and trustworthiness cannot be common knowledge.

players benefit from both players' adherence to the practice. If A plays his part in the practice, expecting B to play hers, he must believe and intend that his action will lead to an outcome that will in fact benefit both of them. Thus, if self-interest has the status of a default assumption and if A is known to be individually rational, his choice of *send* cannot signal that his preferences are non-selfish. Intuitively, however, it seems that that choice *can* signal a socially-oriented intention and an expectation that B will reciprocate this intention. If we are to make sense of this intuition, we need to give up the assumption of individual rationality.

3. Trust and team reasoning

A better way of understanding trust, I suggest, is to treat (*send*, *return*) as a joint action that the two players take part in *together*, and which benefits them both. Viewed in this perspective, A's choice of *send* can be interpreted as his part of that joint action. In making this choice, he signals his expectation that B will play her part too. Expecting this, he chooses *send* with the intention that the joint action (*send*, *return*) will be realised. B's choice of *return* confirms A's belief and reciprocates A's intention. I will argue that this structure of belief, intention and action can be represented by using a model of *team reasoning*.

The idea of team reasoning was first proposed by Hodgson (1967) as part of a demonstration that rule and act utilitarianism are based on fundamentally different modes of reasoning. This argument was developed more fully by Regan (1980) in his theory of 'cooperative utilitarianism'. The significance of team reasoning for game theory was, I think, first pointed out by me (Sugden, 1991, 1993). There are close connections between team reasoning and other 'we' notions used in the literature of social ontology, particularly the concepts of plural subjects (Gilbert, 1989), group agency (List and Pettit, 2011) and collective intentionality (Tuomela and Miller, 1988; Searle, 1990; Bratman, 1993; Bardsley, 2007). As argued by Gold and Sugden (2007), the theory of team reasoning can be interpreted as an alternative way of treating the subject matter of these other analyses of 'we'. For example, collective intentions can be characterised as intentions that are supported by team reasoning. When I use terms such as 'joint action' and 'joint intention' in the context of team reasoning, it should be understood that I am not importing specific properties defined in other contributions to social ontology; I am merely using these terms in their everyday senses in interpreting the formal structure of the theory of team reasoning.

The core idea in this theory is that when two or more individuals engage in team reasoning, each of them asks 'What should *we* do?', and not (as in conventional game theory) 'What should *I* do, given my beliefs about what others will do?' Notice that these two questions remain distinct even if the person who asks 'What should I do?' has preferences that take account of others' payoffs. Thus, team reasoning cannot be reduced to standard game-theoretic reasoning by re-defining payoffs. A team reasoner considers the possible *profiles* of strategies that can be chosen by the players in combination. She assesses these profiles in terms of their consequences for the players *together*, finds the profile that is in the common or collective interest of the players, and then chooses her component of that profile.

This core idea can be developed in different ways. To date, the fullest game-theoretic development is that by Bacharach (1999, 2006). However, I will argue that Bacharach's approach does not adequately represent the intuitive idea that players of the Trust and Market Games can act on joint intentions to achieve mutual benefit.

Any theory of team reasoning needs to explain which sets of individuals, under which circumstances, come to perceive themselves as teams. Bacharach treats this as a question about *group identification*: an individual engages in team reasoning with respect to a particular group if and only if he identifies with that group (that is, thinks of himself as part of that group's agency). For Bacharach, group identification is ultimately a psychological phenomenon, not a matter of rational choice. The underlying thought is that the question of whether a particular action is rational is ill-formed unless the unit of agency has been specified: an action is rational *for* an agent to the extent that it can be expected to achieve that agent's objectives. Thus, the question 'Who am I?' (or 'Who are we?') is logically prior to rational choice. I will say more later about Bacharach's psychological theory of group identification.

Any theory of team reasoning also needs a representation of the collective or common interests of the group or *team* of individuals who reason collectively. In Bacharach's theory, the team's objectives are represented by a *team utility function* that assigns a utility value to every strategy profile (Bacharach, 1999, p. 120; 2006, pp. 87–88). The question 'What should we do?' is construed as 'How can we maximize team utility?' Thus, Bacharach represents team reasoning as *instrumentally* rational, on the model of individual reasoning in conventional decision theory; the difference is that the reasoning described by Bacharach is instrumentally rational *for the team*. Bacharach (2006, pp. 87–88)

argues that is reasonable to assume that team utility is an increasing function of individual payoffs, and suggests that additional properties of this function might include the 'utilitarian' addition of individual payoffs or 'principles of fairness such as those of Nash's axiomatic bargaining theory'. Notice that, although Bacharach does not make any firm proposals about how interpersonal comparisons should be made, any function that assigns a utility value to every strategy profile (and whose application is not restricted to a very narrow class of games) must incorporate interpersonal comparisons between the payoffs of different team members. Thus, team reasoning as modelled by Bacharach can involve trade-offs between members' interests: achieving the best outcome for the team may require that some members bear losses so that others achieve greater gains.

This way of thinking about the good of the team does not fit well with the idea of intentional cooperation for mutual benefit that I have suggested is at the heart of practices of trust. Of course, given the assumption that team utility is increasing in individual payoffs, any joint action that is mutually beneficial to the players of a game (relative to some given benchmark) will also increase the utility of the team that comprises those players. Nevertheless, intending that each player benefits is not the same thing as intending the benefit of the team of players, considered as a single entity. To put this another way, intending to promote the *common* interests of team members is not the same thing as intending to promote the *collective* interests of the team. The former intention is cooperative in a sense that the latter is not.

In Bacharach's theory, once an individual has identified with a team, his willingness to act on team reasoning is not conditional on any assurance that other team members will do the same. When engaging in team reasoning, each player takes account of any probability that other players may fail to identify with the team, but his own reasoning considers only what is best for the team (Bacharach, 2006, pp. 130–135). For example, consider player Row in the Dilemma Game. Suppose that he has identified with the team {Row, Column}, and suppose that team utility is given by the sum of the payoffs to the two players. So, viewed from the perspective of the team, *cooperate* is a strictly dominant strategy. Team reasoning must therefore prescribe that Row chooses *cooperate*, whatever his beliefs about the probability that Column identifies with the team. In particular, it prescribes this choice for Row even if, with probability close to one, Column will use individual reasoning and so choose *defect*. This feature of Bacharach's theory excludes the potential role of reciprocity in motivating cooperative behaviour. It is another instance of Bacharach's focus on the

pursuit of collective rather than common interests. If one is trying (as I am, but Bacharach perhaps was not) to construct a team-reasoning theory of intentional cooperation for mutual benefit, reciprocity must surely be given a role. In such a theory, I suggest, a person who is motivated to seek cooperation need not be committed to act on the prescriptions of team reasoning unless she has adequate assurance that other members of the team will do so too.

In sketching a psychological theory of group identification, Bacharach (2006, pp. 84–86) proposes the hypothesis that group identification is more likely in games with the property of *interdependence*. Roughly, a game has interdependence if there is some strategy profile for which the outcome is strictly Pareto-superior to (that is, has a strictly greater payoff for every player than) at least one Nash equilibrium of the game. The intuition seems to be that players are more likely to think of a game as posing a decision problem 'for us' if they can expect team reasoning to secure mutual benefit *relative to a possible outcome of individually rational choice*. In the Trust Game, for example, (*hold*, *keep*) is the unique subgame-perfect Nash equilibrium.[4] Since the outcome of this individually-rational strategy profile is strictly Pareto-inferior to that of (*send*, *return*), the interdependence property holds.

To use the outcome of individually rational choice as a benchmark in this way is to treat individual rationality as an unproblematic norm, and to treat team reasoning as a kind of add-on reasoning module that is activated only when individual rationality might lead to collectively undesired consequences. But why should individual rationality be privileged in this way? Consider the Market Game. In this game, (*send*, *return*) is the unique subgame-perfect Nash equilibrium. The outcome of this strategy profile is strictly Pareto-superior to both of the other possible outcomes. So there is good reason to expect that if both players were individually rational, and if each knew that this was true of the other, they would arrive at the unique Pareto-optimal outcome. But that does not mean that the players cannot understand (*send*, *return*) as a mutually beneficial joint action. Intuitively, it seems that they *could* understand it in this way, each choosing his or her component of the joint action with the intention of achieving mutual benefit. To do this, however, they would have to use a concept of mutual benefit that was not defined relative to the benchmark of individually rational choice.

[4] In a game in which the players move sequentially, it is possible to define 'subgames' that are reached after particular moves have been played. A strategy profile for the whole game is a subgame-perfect Nash equilibrium if it is a Nash equilibrium, not only in the whole game, but also in every subgame. In the Trust Game, A's choice of *send* leads to a one-player subgame in which B chooses between *keep* and *return*; in this subgame, *keep* is the unique Nash equilibrium.

A further feature of Bacharach's theory of team reasoning (shared by the representations of team reasoning in my 1991 and 1993 papers) is that it attributes a high degree of collective rationality to teams. In Bacharach's theory, once it is common knowledge that each member of a group of individuals has identified with that group as a unit of agency, each of them recognises *the same* team utility function as their common objective. Provided there is a unique strategy profile that maximises that function, each member of the team can discover that profile by independent reasoning. Thus, in many games which would present coordination problems to individually rational agents, team reasoners can resolve those problems by rationality alone. (The qualification 'many' is necessary because this method of coordination fails if two or more distinct strategy profiles induce exactly the same optimal level of team utility.)

In some games, there is so little room for disagreement about the relevant properties of team utility that Bacharach's explanation of coordination works well. This is particularly true of the Hi-Lo Game, which figures prominently in Bacharach's arguments (as in those of Hodgson [1967] and Sugden [1991, 1993]). A version of this game is shown in Table 2. Here, it seems indisputable that team utility is uniquely maximised by the strategy profile (*high*, *high*), and that this is therefore the uniquely rational choice for team reasoners. This argument can be developed to offer explanations of how players coordinate on saliently-labelled strategy profiles in pure coordination games of the kind discussed by Schelling (1960) – for example the game in which two players who are unable to communicate with one another are rewarded if and only if they both give the same answer to some question such as 'Name a place to meet the other player in New York City'. (Roughly, these explanations work by building strategy labels into the formal structure of the relevant game so as to transform it into a Hi–Lo game. See, for example, Bacharach [1993], Sugden [1995], Janssen [2001], and Casajus [2001].)

Table 2: The Hi–Lo Game

		Column's strategy	
		high	*low*
Row's strategy	*high*	2, 2	0, 0
	low	0, 0	1, 1

In many games, however, it is implausible to assume that, merely by virtue of group identification, individuals can identify a uniquely team-optimal strategy profile. Experimental evidence seems to show that explanations of coordination which assume that players reason independently to team-optimal solutions work well in some games but not in others (e.g. Crawford, Gneezy and Rottenstreich [2008]; Bardsley et al. [2010]). In many real-world situations, mutually beneficial cooperation consists in conforming to complex and sometimes arbitrary conventions that could not be reconstructed by abstract rational analysis. For example, consider the many informal conventions governing who gives way to whom on the roads. Having understood what these conventions are, a road user who conforms to them can readily think of herself as participating in mutually beneficial practices; but she would not be able to discover these conventions by reasoning about optimal solutions to traffic management problems. Quite apart from the technical difficulty of specifying and solving those optimisation problems, there is no guarantee that the conventions that are in operation are the optimal ones. If individuals are to cooperate effectively, they need to be ready to play their parts in mutually beneficial practices that seem to them to be – and perhaps really are – less than ideal.

To sum up the argument so far: if intentional cooperation for mutual benefit is to be represented as team reasoning, we need a theory of team reasoning that differs from that proposed by Bacharach. We need a theory: in which team members aim to achieve their common interests, not to maximise a common utility function; in which individuals act on team reasoning only if they have assurance that sufficient other members of the team will do so too; in which individually rational choice is not used as a benchmark for defining mutual benefit; and in which team reasoners can coordinate their behaviour by following pre-existing practices that are less than optimal. I will now outline such a theory.[5]

4. A new representation of team reasoning

As a first step, I propose a definition of a 'mutually beneficial practice'.

Consider any game for n players (where $n \geq 2$), defined in terms of the strategies available to the players and the payoffs that result from the possible combinations of strategies. Payoffs are interpreted as in Sections 2 and 3. The players may move simultaneously, as in the Dilemma and Hi–Lo Games, or sequentially, as in the Trust,

[5] This outline develops, and in some respects corrects, ideas first sketched out in Sugden (2011).

Market and Confidence Games. Simultaneous-move games are described in 'normal form', as in Tables 1 and 2; sequential-move games are described in 'extensive form', as in Figures 1, 2 and 3. A strategy in a sequential-move game determines which choice the relevant player will make at every contingency that is possible, given the rules of the game.

For each player $i = 1, \ldots, n$, there is a set S_i of strategies, from which she must choose one; a typical strategy for player i is written as s_i. For each strategy profile $(s_1, \ldots, s_n)$, there is a payoff to each player i, written as $u_i(s_1, \ldots, s_n)$. For each player i, let $\bar{u}_i$ be her *maximin* payoff – that is, the highest payoff that she can guarantee herself, independently of the other players' strategy choices. (Formally: for each strategy in S_i, we find the minimum payoff that i can receive, given that this strategy is chosen; then we find the strategy for which this minimum payoff is maximised. This strategy's minimum payoff is i's maximin payoff.) I shall treat each player's maximin payoff as the benchmark for defining the benefits of cooperation. The intuitive idea is that a player can guarantee that she receives at least this payoff without engaging in any intentional interaction with the other players.

This benchmark might be interpreted in the spirit of Hobbes's (1651/ 1962) state of nature. A Hobbesian might say that whatever an individual can be sure of getting for herself by whatever means, irrespective of what others do, cannot be a product of cooperation, and so each player's maximin payoff sets a lower bound to the value that she can achieve from the game without cooperating with others. Alternatively, one might take a more moralised approach, in which the rules of the game are interpreted as specifying what individuals can *legitimately* or *rightfully* do, rather than what they can *in fact* do.[6] For example, in a model of an exchange economy, one might postulate an initial distribution of endowments and a system of rules that allows each individual to keep her own endowments if she so chooses and allows any group of individuals to trade endowments by mutual consent. In such a model, each player's maximin payoff would be the value to her of keeping her endowments.

I begin with the case of a two-player game, for which the concept of mutual advantage is relatively easy to define. I shall say that a strategy profile (s_1^*, s_2^*) is a *mutually beneficial practice* in a two-player game if and only if, for each player i, $u_i(s_1^*, s_2^*) > \bar{u}_i$. In other words: (s_1^*, s_2^*) is a mutually beneficial practice if and only if each player benefits, relative to her maximin benchmark, from both players' participation in the practice.

[6] This way of thinking about games is developed in Sugden (1985) in an analysis of liberty and rights.

In each of the games that I have presented so far, I have deliberately calibrated payoffs so that each player's maximin payoff is zero. For example, in the Trust Game, player A can guarantee a payoff of zero by choosing *hold*, but incurs the risk of a negative payoff if he chooses *send*. B can ensure a positive payoff if A chooses *send*, but she cannot prevent him from choosing *hold*, which would give her payoff of zero.

In the Trust Game, one and only one (pure) strategy profile, namely (*send*, *return*), is a mutually beneficial practice. Exactly the same is true of the Market Game, consistently with my argument about the parallelism between the two games. In contrast, but in line with my discussion of that game, there is no mutually beneficial practice in the Confidence Game. For completeness, I add that (*cooperate*, *cooperate*) is the unique mutually beneficial practice in the Dilemma Game, and that in the Hi–Lo game, (*high*, *high*) and (*low*, *low*) are both mutually beneficial practices.

Generalising the definition of 'mutually beneficial practice' to games with any number of players is not completely straightforward. Consider the three-player Snowdrift Game, shown in Table 3. The story behind the game is that A, B and C are the drivers of three cars stuck in the same snowdrift, each equipped with a shovel. If a way out is dug for any one car, the others can use it. Each driver chooses whether to *dig* or to *wait* (hoping either that someone else will dig, or that a snowplough will arrive on the scene). Digging has a cost of 6, divided equally between those who do the work; provided there is at least one digger, each player gets a benefit of 4 from the work that is done. Each player gets his maximin payoff of zero by choosing *wait*. However, if any two players *dig*, all three get positive payoffs.

It seems obviously right to say that (*dig*, *dig*, *dig*), which gives the payoff profile (2, 2, 2), is a mutually beneficial practice. But what about (*dig*, *dig*, *wait*), which gives (1, 1, 4)? Relative to their maximin payoffs, all three players benefit from this practice; but is the benefit *mutual*? Surely not: C benefits from A's and B's participation in the practice, but that benefit is not reciprocated. One way of putting this is to say that, irrespective of C's strategy choice, A and B can each be sure of getting a payoff of at least 1 if they both choose *their components of* the practice (*dig*, *dig*, *wait*). Thus, neither of them benefits from C's choosing her component.[7]

[7] However, if A and B were to treat C's choice of *wait* as given, they would effectively be playing a two-player game between themselves – the game represented by the matrix in the top part of Table 3,

Table 3: The Snowdrift Game

If C chooses *wait*:

		B's strategy	
		wait	*dig*
A's strategy	*wait*	0, 0, 0	4, –2, 4
	dig	–2, 4, 4	1, 1, 4

If C chooses *dig*:

		B's strategy	
		wait	*dig*
A's strategy	*wait*	4, 4, –2	4, 1, 1
	dig	1, 4, 1	2, 2, 2

Generalising this argument, I propose the following definition. In any game for n players (where $n \geq 2$), a strategy profile $\mathbf{s}^* = (s_1^*, \ldots, s_n^*)$ is a *mutually beneficial practice* if and only if two conditions are satisfied. *Condition 1* is that, for each player $i = 1, \ldots, n$, $u_i(\mathbf{s}^*) > \bar{u}_i$: relative to her maximin benchmark, each player benefits from the practice. To formulate the second condition, let N be the set of players $\{1, \ldots, n\}$, and consider any *subgroup G*, where G is a subset of N that contains at least one and fewer than n players. Let G' be the complement of G. For each player j in G, let $v_j(G, \mathbf{s}^*)$ be the minimum payoff that j can receive, given that each member of G chooses his component of $\mathbf{s}^*$. I will say that G *benefits from the participation of* G' *in* $\mathbf{s}^*$ if and only if $u_i(\mathbf{s}^*) \geq v_j(G, \mathbf{s}^*)$ for all j in G, with a strict inequality for at least one j. *Condition 2* is that, for *every* subgroup G that contains at least one and fewer than n players, G benefits from the participation of G'.

---

with C's payoffs removed. In that game, the choice of *dig* by both A and B would be a mutually beneficial practice *for A and B*. Viewing their situation in this way, A and B might each choose *dig* as their parts of this two-person practice, while being aware that C was taking a free ride. I used this idea in an early theory of reciprocity, which I now see as a precursor of the theory of team reasoning (Sugden, 1984).

In a two-player game, Condition 2 is redundant. (Consider any two-player game and any strategy profile (s_1*, s_2*) which satisfies Condition 1. Thus $u_1(s_1^*, s_2^*) > \bar{u}_1$. By the definition of 'maximin payoff', $\bar{u}_1$ is as least as great as player 1's minimum payoff, conditional on his having chosen s_1*. So $u_1(s_1^*, s_2^*)$ is strictly greater than player 1's minimum payoff, given his choice of s_1*. This implies that the subgroup {1} benefits from the participation of its complement {2} in (s_1*, s_2*). By the same reasoning, {2} benefits from the participation of {1}. So Condition 2 is satisfied.) But when $n > 2$, neither condition implies the other.

Notice that Condition 2 does not require that *every* player benefits from *every other* player's participation in the practice **s***. For example, consider a variant of the Snowdrift Game in which A's choice of *dig* benefits only A and B, B's choice of *dig* benefits only B and C, and C's choice of *dig* benefits only C and A. C does not benefit from A's participation in the practice (*dig*, *dig*, *dig*), A does not benefit from B's participation, and B does not benefit from C's. Still, each subgroup benefits from the participation of its complement, and so Condition 2 is satisfied.

Notice also that, in defining the benefit that G receives from the participation of G' in the practice **s***, Condition 2 takes G's participation in that practice as given. It does not ask what payoff profiles G could have guaranteed itself by concerted action. Recall that I want to be able to say that an ongoing practice is mutually beneficial even if it is less than optimal. For example, suppose that **s*** and **s**** are two different priority rules that could be followed by the one million users of a national road network. In fact, everyone follows **s***, and this works well; relative to maximin benchmarks, everyone benefits greatly. However, traffic engineers can show that there would be a small but positive benefit to everyone if everyone switched to **s****. It is possible that a subgroup of 999,999 road users could guarantee that each of them would be better off if they all switched to **s****, irrespective of the behaviour of the one remaining individual. But it still seems right to say that this subgroup benefits from its complement's participation in the ongoing practice **s***, and hence that this practice is mutually beneficial.

My definition of a mutually beneficial practice does not impose any restrictions on how the benefits of a practice are distributed between the participants, beyond the condition that every participant gains *some* benefit. One might argue that an account of cooperation needs to take account of the distribution of benefits, and that for a practice to be genuinely cooperative, benefits must be distributed in a reasonably fair way. I say 'reasonably'

because my analysis is intended to apply to ongoing practices, without assuming that individuals can solve coordination problems by abstract team reasoning. It would be inappropriate to require that, in order for individuals to be led by team reasoning to participate in cooperative practices, those practices must be *perfectly* fair according to some well-defined criterion that everyone endorses. Still, by adding some minimum standards of fairness, it might be possible to construct a satisfactory definition of a *fair* mutually beneficial practice. For the purposes of this paper, however, I leave this issue aside.

As a preliminary to presenting a schema of team reasoning, I need to state some definitions. I shall say of any proposition *p* and any set of players *N* that *in N, there is common reason to believe p* if and only if (i) each player *i* in *N* has reason to believe *p*, (ii) each player *i* in *N* has reason to believe that each player *j* in *N* has reason to believe *p*, and so on.[8] For any property *q*, I shall say that *in N, there is reciprocal reason to believe that q holds for members of N* if and only if (i) each player *i* in *N* has reason to believe that *q* holds for each player $j \neq i$ in *N*, (ii) each player *i* in *N* has reason to believe that each player $j \neq i$ in *N* has reason to believe that *q* holds for each player $k \neq j$ in *N*, and so on.

Notice that the definition of 'reciprocal reason to believe' makes no reference to what any player has reason to believe *about himself*. This omission is significant when the property *q* refers to choices made by the players themselves. For example, take the Dilemma Game and consider what is implied by the proposition that, in the set of players {Row, Column}, there is reciprocal reason to believe that 'will choose *cooperate*' holds for members of that set. Among these implications are: that Row has reason to believe that Column will choose *cooperate*; that Row has reason to believe that Column has reason to believe that Row will choose *cooperate*; and so on. But nothing is said about whether Row has reason to believe that *Row* will choose *cooperate*. Nor (since one can have reason to believe a proposition that is in fact false) has anything been said about whether *in fact* Row will choose *cooperate*. For example, suppose that Row and Column have played the Dilemma Game against one another many times, and both players have always chosen *cooperate*. They are about to play the game again. One might argue that, by the canons of inductive reasoning, there is (in the set of players {Row, Column}) reciprocal reason to believe that each player will choose *cooperate*. But each player can still ask whether he or she has reason to make this choice.

[8] I use 'reason to believe' in the sense of Lewis (1969) and Cubitt and Sugden (2003).

I now present a schema of team reasoning that can be used by each player in any game that has two or more players. The set of players is $N = \{1, \ldots, n\}$; $\mathbf{s}^* = (s_1^*, \ldots, s_n^*)$ is any strategy profile in that game. The propositions P1 to P3 are premises that 'I' (one of the players) accept; the proposition C is a conclusion that 'I' infer from those premises. I as author am not asserting that this schema 'really' is valid. Rather, it is a schema that any player *might* endorse. Were she to do so, she would *take it to be* valid.

Schema of Cooperative Team Reasoning

(P1) In N, there is common reason to believe that $\mathbf{s}^*$ is a mutually beneficial practice.

(P2) In N, there is reciprocal reason to believe that each player will choose her component of $\mathbf{s}^*$.

(P3) In N, there is reciprocal reason to believe that each player endorses and acts on the Schema of Cooperative Team Reasoning with respect to N.

(C) I should choose my component of $\mathbf{s}^*$ (or some other strategy that is unconditionally at least as beneficial for every player).[9]

The concept of 'endorsing and acting on the Schema of Cooperative Team Reasoning' is the analogue of group identification in Bacharach's theory. To endorse the schema is to dispose oneself to treat N as a unit of agency and to play one's part in its joint actions. The schema itself prescribes what that part is. For each player i (and leaving aside the complication of the 'or some other strategy …' clause in C), that part is i's component of a strategy profile $\mathbf{s}^*$ for which there is common reason to believe in its being mutually beneficial (P1) and for which there is reciprocal reason to believe in its being chosen (P2). However, the schema has implications for each player's choices only if there is assurance that all players endorse it (P3).

The status of P3 in the schema is analogous with that of a clause in a contract between two parties stating that the contract is to be activated if and when both parties have

[9] A strategy s_i' for some player i is *unconditionally at least as beneficial as* s_i^* *for* some player j if and only if, irrespective of the strategy choices of players other than i, i's choice of s_i' guarantees that j's payoff will be at least as great as $u_j(\mathbf{s}^*)$. The clause in parentheses allows a team-reasoning player to deviate from $\mathbf{s}^*$ if she can be certain that no one would be harmed by her doing so.

signed it. The first party to sign such a contract makes a unilateral commitment to abide by the terms of the contract, but those terms do not require anything of her unless and until the second party signs. Similarly, if a player commits herself to act on the Schema of Cooperative Team Reasoning, that commitment makes no demands on her unless there is reciprocal reason to believe that every player has made the same commitment.

One might ask why P3 is needed in addition to P2. It would certainly be possible to postulate a reasoning schema (call it the Simple Schema) in which C can be inferred merely from P1 and P2. Roughly speaking, a player who endorses the Simple Schema commits herself to the *individual* action of choosing her component of a mutually beneficial practice when other players can be expected to choose theirs. This is an intelligible moral principle, but it does not involve the idea of *joint* intention or *joint* action. For example, consider the Trust Game, with **s*** defined as the mutually beneficial practice (*send*, *return*). Consider how B might reason about the game, given that she has reason to believe that A will choose *send* (or indeed, given that she knows that A has already chosen *send*). If she endorses the Simple Schema, she does not need to enquire into A's intentions in order to conclude that she should choose *return*. But this makes it difficult to represent the idea that she intends her action as a repayment of A's trust.

In contrast, suppose that in the Trust Game, A and B each endorse the Schema of Cooperative Team Reasoning, and that there is reciprocal reason for them to believe that this is the case. Further, suppose that there is reciprocal reason for them to believe that A will choose *send* and B will choose *return*. The latter beliefs might be supported by inductive inferences from previous observations of *send* and *return* in Trust Games – perhaps previous games played between A and B, or perhaps games played by other pairs of players drawn from some population of which they are both members. Then A and B can each infer they should choose their respective components of the mutually beneficial practice (*send*, *return*), *with the joint intention of participating in that practice*. In choosing *send*, A acts on his part of this intention, trusting B to act on her part of it. B repays A's trust by doing so.

Now consider how this argument extends to the Market Game. In the Market Game, (*send*, *return*) is the strategy profile that is uniquely recommended to individually rational and self-interested players who have reciprocal reason to believe one another to be individually rational and self-interested. Thus, A might choose *send* and B might choose *return*, each acting on an individual intention to pursue his or her self-interest, as suggested by Adam Smith's account of how we get our dinners. But there is another possibility: A and

B might both endorse the Schema of Cooperative Team Reasoning. If there is reciprocal reason for them to believe that this is the case, and if there is reciprocal reason for them to believe that A will choose *send* and that B will choose *return*, they can choose *send* and *return* with the joint intention of participating in a mutually beneficial practice.

5. Conclusion

I have described a form of team reasoning which, if followed by each member of a group of interacting individuals, can support mutually beneficial cooperation. This reasoning is carried out separately by each individual, but each individual reasons *as a member of the group*, with the intention of playing her part in practices that are mutually beneficial for group members.

In some cases, such as that of a second mover in the Trust Game who has the opportunity to take advantage of the first mover's trust, this reasoning can lead her to perform actions that are contrary to her self-interest, given the actual or expected behaviour of other group members. But in such cases, the team-reasoner does not construe her action as a sacrifice of her individual interests to achieve some 'social' end, such as rewarding a co-player's kindness or punishing his unkindness. Nor does she think of herself as adopting a collective goal that transcends her private interests. Rather, she views her action as her part of a practice that, if followed by all members of the group, will benefit all of them; and since she has reason to believe that the others will participate (or have already done so), she expects to share in the benefits of the practice.

However, and perhaps just as significantly, there are cases such as the Market Game in which team reasoning leads individuals to perform actions that *are* in their self-interest, given the actual or expected behaviour of other group members. Nevertheless, the team-reasoner's intention in so acting is not self-interest, but mutual benefit. Thus, contrary to the implication of Smith's remarks about butchers, brewers and bakers, ordinary market transactions do not have to be understood as expressing self-interest on each side. To say this is not to make the claim that Smith rightly rejected, namely that market behaviour is motivated by benevolence. In a well-ordered society, market transactions can express intentions for mutually beneficial cooperation.

References

Akerlof, George (1982): “Labor Contracts as Partial Gift Exchange”. In: *Quarterly Journal of Economics* 97, p. 543-569.

Anderson, Elizabeth (1993): *Value in Ethics and Economics*. Cambridge, MA: Harvard University Press.

Bacharach, Michael (1993): “Variable Universe Games”. In: *Frontiers of Game Theory*. Ken Binmore, Alan Kirman and Piero Tani (Eds.). Cambridge, MA: MIT Press, p. 255-275.

Bacharach, Michael (1999): “Interactive Team Reasoning: A Contribution to the Theory of Cooperation”. In: *Research in Economics* 53, p. 117-147.

Bacharach, Michael (2006): *Beyond Individual Choice.* Natalie Gold and Robert Sugden (eds). Princeton, NJ: Princeton University Press.

Bacharach, Michael, Gerardo Guerra and Daniel Zizzo (2007): “The Self-fulfilling Property of Trust: An Experimental Study”. In: *Theory and Decision* 63, p. 349-388.

Bardsley, Nicholas (2007): “On Collective Intentions: Collective Action in Economics and Philosophy”. In: *Synthese* 157, p. 141-159.

Bardsley, Nicholas, Judith Mehta, Chris Starmer, and Robert Sugden (2010): “Explaining Focal Points: Cognitive Hierarchy Theory *versus* Team Reasoning.” In: *Economic Journal* 120, p. 40-79.

Battigalli, Pierpaolo and Martin Dufwenberg (2007): “ Guilt in Games”. In: *American Economic Review: Papers and Proceedings* 97, p. 171-176.

Berg, Joyce, John Dickhaut and Kevin McCabe (1995): “Trust, Reciprocity, and Social History”. In: *Games and Economic Behavior* 10, p. 122-142.

Bolton, Gary and Axel Ockenfels (2000): “ERC: A Theory of Equity, Reciprocity and Competition”. In: *American Economic Review* 90, p. 166-193.

Bratman, Michael (1993): “Shared Iintention”. In: *Ethics* 104, p. 97-113.

Bruni, Luigino and Robert Sugden (2008): “Fraternity: Why the Market Need Not Be a Morally Free Zone”. In: *Economics and Philosophy* 24, p. 35-64.

Bruni, Luigino and Robert Sugden (2013): “Reclaiming Virtue Ethics for Economics”. In: *Journal of Economic Perspectives* 27(4), p. 141-164.

Casajus, André (2001): *Focal Points in Framed Games: Breaking the Symmetry*. Berlin: Springer-Verlag.

Crawford, Vincent, Uri Gneezy, and Yuval Rottenstreich (2008): “The Power of Focal Points Is Limited: Even Minute Payoff Asymmetry May Yield Large Coordination Failures”. In: *American Economic Review* 98, p. 1443-1458.

Cubitt, Robin and Robert Sugden (2003): “Common knowledge, salience and convention: a reconstruction of David Lewis’s game theory”. In: *Economics and Philosophy* 19, p. 175- 210.

Falk, Armin, Ernst Fehr and Urs Fischbacher (2003): “On the Nature of Fair Behavior. In: *Economic Inquiry* 41, p. 20-26.

Fehr, Ernst and Klaus Schmidt (1999): “A Theory of Fairness, Competition and Cooperation”. In: *Quarterly Journal of Economics* 114, p. 817-868.

Gilbert, Margaret (1989): *On Social Facts*. London: Routledge.

Hobbes, Thomas (1651/ 1962): *Leviathan*. London: Macmillan.

Hodgson, David (1967): *Consequences of Utilitarianism.* Oxford: Clarendon Press.

Hume, David (1740/ 1978): *A Treatise of Human Nature*. Oxford: Clarendon Press.

Janssen, Maarten (2001): “Rationalising Focal Points”. In: *Theory and Decision* 50, p. 119-148.

Lewis, David (1969): *Convention: A Philosophical Study*. Cambridge, MA: Harvard University Press.

List, Christian and Philip Pettit (2011): *Group Agency: The Possibility, Design, and Status of Corporate Agents*. Oxford: Oxford University Press.

Neumann, John von and Oskar Morgenstern (1947): *Theory of Games and Economic Behavior*, 2nd edition. Princeton, NJ: Princeton University Press.

Pelligra, Vittorio (2005): “Under Trusting Eyes: The Responsive Nature of Trust”. In: *Economics and Social Interaction*. Benedetto Gui and Robert Sugden (Eds). Cambridge: Cambridge University Press, p. 195-124.

Rabin, Matthew (1993). “Incorporating Fairness into Game Theory and Economics”. In: *American Economic Review* 83, p. 1281-1302.

Regan, Donald (1980): *Utilitarianism and Cooperation.* Oxford: Clarendon Press.

Sandel, Michael J. (2012): *What Money Can’t Buy: The Moral Limits of Markets*. New York: Farrar, Straus and Giroux.

Schelling, Thomas (1960): *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.

Searle, John (1990): “Collective Intentions and Actions”. In: *Intentions in Communication*. Philip Cohen, Jerry Morgan and Martha Pollack (Eds). Cambridge, MA: MIT Press, p. 401-415.

Smith, Adam (1776/ 1976): *An Inquiry into the Nature and Causes of the Wealth of Nations*. Oxford: Clarendon Press.

Sugden, Robert (1984): “Reciprocity: The Supply of Public Goods through Voluntary Contributions”. In: *Economic Journal* 94, p. 772-787.

Sugden, Robert (1985): “Liberty, Preference and Choice”. In: *Economics and Philosophy* 1, p. 213-229.

Sugden, Robert (1991): “Rational Choice: A Survey of Contributions from Economics and Philosophy”. In: *Economic Journal* 101, p. 751-785.

Sugden, Robert (1993): “Thinking as a Team: Toward an Explanation of Nonselfish Behavior”. In: *Social Philosophy and Policy* 10, p. 69-89.

Sugden, Robert (1995): “A theory of Focal Points”. In: *Economic Journal* 105, p. 533-550.

Sugden, Robert (2011): “Mutual Advantage, Conventions and Team Reasoning”. In: *International Review of Economics* 58, p. 9-20.

Tuomela, Raimo and Kaarlo Miller (1988): “We-Intentions”. In: *Philosophical Studies* 53, p. 367-389.

Figure 1: The Trust Game

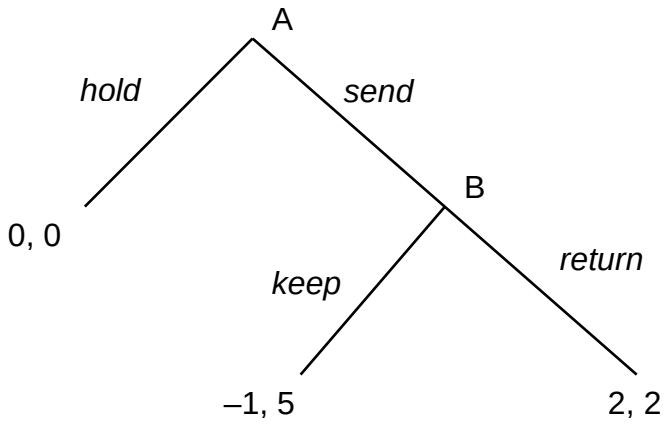

Figure 2: The Market Game

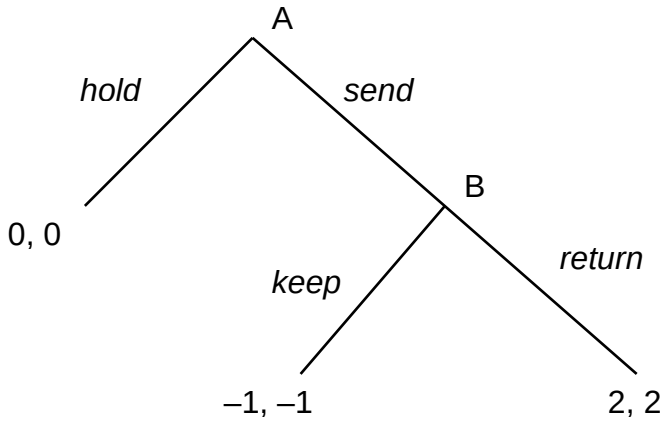

Figure 3: The Confidence Game