12

Theories of Team Agency

NATALIE GOLD AND ROBERT SUGDEN*

In decision theory, it is almost universally presupposed that agency is invested in individuals: each person acts on her own preferences and beliefs. A person's preferences may take account of the effects of her actions on other people; she may, for example, be altruistic or have an aversion to inequality. Still, these are *her* preferences, and she chooses what *she* most prefers. Opposing this orthodoxy is a small body of literature which allows *teams* of individuals to count as agents, and which seeks to identify distinctive modes of *team reasoning* that are used by individuals as members of teams. This idea has been around for some time, having been proposed in different forms by David Hodgson (1967), Donald Regan (1980), Margaret Gilbert (1989), Susan Hurley (1989), Robert Sugden (1993, 2003), Martin Hollis (1998) and Michael Bacharach (1999, 2006). Closely related, but less directly concerned with decision theory, is the literature of collective intentions, exemplified by the work of Raimo Tuomela and Kaarlo Miller (1988), John Searle (1990) and Michael Bratman (1993). These ideas have yet to capture the attention of mainstream decision theory.

There seems to be a suspicion either that team reasoning is a particular case of individual reasoning, distinguished only by the particular assumptions it makes about preferences, or that it is not *reasoning* in the true sense of the word. The main contribution of the present paper is to represent team reasoning explicitly, as a *mode of reasoning* in which propositions are manipulated according to well-defined rules—an approach that has previously been used by Natalie Gold and Christian List (2004). Our basic building block is the concept of a *schema of practical reasoning*, in which conclusions about what actions should be taken are inferred from explicit

premises about the decision environment and about what agents are seeking to achieve. We use this theoretical framework to compare team reasoning with the individual reasoning of standard decision theory, and to compare various theories of team agency and collective intentionality.

1. Two puzzles of game theory

One motivation for theories of team reasoning is that there are games that are puzzles for orthodox decision theory, in the sense that there exists some strategy that is at least arguably rational and that a substantial number of people play in real life, but whose rationality decision theory cannot explain and whose play it cannot predict. In this paper, we focus on two such puzzles, and show how the theory of team reasoning can resolve them.

The first puzzle is the Prisoner's Dilemma, shown in figure 12.1. In specifying the payoffs of this game, we require only that they are symmetrical between the players and that they satisfy two inequalities. The inequality $a > b > c > d$ encapsulates the central features of the Prisoner's Dilemma: that, for each player, the best outcome is that in which he chooses *defect* and his opponent chooses *cooperate*; the outcome in which both choose *cooperate* is ranked second; the outcome in which both choose *defect* is ranked third; and the outcome in which he chooses *cooperate* and his opponent chooses *defect* is the worst of all. The inequality $b > (a + d)/2$ stipulates that each player prefers a situation in which both players choose *cooperate* to one in which one player chooses *cooperate* and the other chooses *defect*, each player being equally likely to be the freerider. This condition is usually treated as a defining feature of the Prisoner's Dilemma.

		Player 2	
		cooperate	*defect*
Player 1	*cooperate*	*b,b*	*d,a*
	defect	*a,d*	*c,c*

$a > b > c > d$; $b > (a + d)/2$

Figure 12.1. The Prisoner's Dilemma

For each player, *defect* strictly dominates *cooperate*. Thus, in its explanatory form, conventional game theory predicts that both players will choose

defect. In its normative form, it recommends *defect* to both players. Yet both would be better off if each chose *cooperate* instead of *defect*.

Is that a puzzle? If, in fact, almost all human players of Prisoner's Dilemma games chose *defect*, and if they construed this choice as rational, it might reasonably be argued that there was nothing to be puzzled about. It would just be an unfortunate fact about rationality that the actions of rational individuals can combine to produce outcomes that, from every individual's point of view, are sub-optimal. But the truth is that, in experiments in which people play the Prisoner's Dilemma for money, anonymously and without repetition, the proportion of participants choosing *cooperate* is typically between 40 and 50 per cent (Sally, 1995). If one describes the game to ordinary people (or, indeed, to philosophers or to social scientists who have not been trained in economics), one finds a similar division of opinion about what a rational player ought to do. While some people find it completely obvious that the rational choice is *defect*, others are equally convinced that rationality requires each player to choose *cooperate*.

The Prisoner's Dilemma poses practical problems for us collectively, as citizens. Economic and social life constantly throws up real games of the Prisoner's Dilemma type. (Think of individuals' decisions about whether to vote in elections, whether to contribute to fund-raising appeals for public goods, whether to reduce consumption of carbon fuels, and so on.) It would be better for all of us if each of us was disposed to be cooperative in such games. The evidence shows that some people *do* act on this disposition in some circumstances. If we understood better what factors induced cooperation, we might find ways of structuring the social environment so as to make cooperation more common.

The Prisoner's Dilemma also poses a problem for explanatory game theory. Conventional game theory predicts that players will always choose *defect*, while in fact many players choose *cooperate*: the theory is failing to explain observed behaviour in games. There is a parallel problem for normative game theory. The theory prescribes *defect*, but many people have the strong intuition that *cooperate* is the rational choice. Of course, it is open to the game theorist to argue that that intuition is mistaken, and to insist on the normative validity of the standard analysis. In doing so, the game theorist can point out that any individual player of the Prisoner's Dilemma does better by choosing *defect* than by choosing *cooperate*, irrespective of the behaviour of her opponent. In other words, each individual player can

reason to the conclusion: 'The action that gives the best result *for me* is *defect*'. But, against that, it can be said with equal truth that the two players of the game both do better by their both choosing *cooperate* than by their both choosing *defect*. Thus, each player can also reason to the conclusion: 'The pair of actions that gives the best result *for us* is not (*defect, defect*).'[1] It seems that normative argument between these two positions leads to a stand-off.

The second puzzle is the game of *Hi-Lo*. A Hi-Lo game is a game in which each of two players chooses one element from the same set of labels, the pair of payoffs is (a_i, a_i) if both choose the same label i (with $a_i > 0$ and (0, 0) otherwise), and there is one label j such that a_j is strictly greater than every other a_i. Figure 12.2 shows a simple version of Hi-Lo, in which there are just two labels, *high* and *low*.

		Player 2	
		high	*low*
Player 1	*high*	*a, a*	0, 0
	low	0, 0	*b, b*

$a > b > 0$

Figure 12.2. Hi-Lo

Hi-Lo combines features of pure coordination games[2] and the Prisoner's Dilemma. Like a pure coordination game, this is a *common interest game*—that is, a game in which the interests of the players are perfectly aligned, signalled by the fact that, in each cell of the payoff matrix, the two players' payoffs are equal to one another. There are two pure-strategy Nash equilibria, each associated with a different label and coming about if both players choose that label. In this sense, Hi-Lo poses a coordination problem: each player wants it to be the case that they both choose the same label. The crucial difference from a pure coordination game is that, in Hi-Lo, one of the equilibria is strictly better than the other for both players. At first sight, this makes the coordination problem in Hi-Lo trivial: it seems obvious that the players should coordinate on the equilibrium they both prefer, namely (*high, high*).

Hi-Lo shares with the Prisoner's Dilemma the feature that, of the outcomes that occur if both players choose the same label, one is better than the other for both players. In this sense, Hi-Lo poses a cooperation problem: both players benefit by their both choosing *high* rather than

low just as, in the Prisoner's Dilemma, both players benefit by their both choosing *cooperate* rather than *defect*. The difference is that in the Prisoner's Dilemma, (*cooperate, cooperate*) is not a Nash equilbrium while in Hi-Lo, (*high, high*) is. It might seem that, because of this difference, the cooperation problem in Hi-Lo is trivial too.

Certainly, Hi-Lo does not pose practical problems for ordinary people, either individually or collectively. In experiments in which participants play Hi-Lo games, and in which the *high* and *low* strategies are given neutral labels, the overwhelming majority choose *high*.[3] But Hi-Lo presents a fundamental problem for game theory. From the assumptions that the players are perfectly rational (in the normal sense of maximizing expected payoff) and that they have common knowledge of their rationality, we cannot deduce that each will choose *high*. Or, expressing the same idea in normative terms, there is no sequence of steps of valid reasoning by which perfectly rational players can arrive at the conclusion that they ought to choose *high*. Many people find this claim incredible, but it is true. It is true because, from the assumption of rationality, all we can infer is that each player chooses the strategy that maximizes her expected payoff, given her beliefs about what the other player will do. All we can say in favour of *high* is that, if either player expects the other to choose *high*, then it is rational for the first player to choose *high* too; thus, a shared expectation of *high*-choosing is self-fulfilling among rational players. But exactly the same can be said about *low*. Intuitively, it seems obvious that each player should choose *high* because both prefer the outcome of (*high, high*) to that of (*low, low*); but that 'because' has no standing in the formal theory.[4]

If we are prepared to relax the classical assumption of perfect rationality, it is not particularly difficult to construct theories which purport to explain the choice of *high*. After we have stripped out any information contained in their labels, the only difference between the *high* and *low* strategies is that *high* is associated with higher payoffs; because of this, most plausible theories of imperfect rationality predict that *high* is more likely to be chosen than *low*.[5] But it seems unsatisfactory to have to invoke assumptions about imperfections of rationality in order to explain behaviour in such a transparently simple game as Hi-Lo. If we find that standard game-theoretic reasoning cannot tell players how to solve the apparently trivial problem of coordination and cooperation posed by Hi-Lo, we may begin to suspect that

something is fundamentally wrong with the whole analysis of coordination and cooperation provided by the standard theory. Conversely, if we could find a form of reasoning which recommends *high* in Hi-Lo, that might provide the key to solving the problem posed by the Prisoner's Dilemma.

The source of both puzzles seems to be located in the mode of reasoning by which, in the standard theory, individuals move from preferences to decisions. In the syntax of game theory, each individual must ask separately 'What should *I* do?' In Hi-Lo, the game-theoretic answer to this question is indeterminate. In the Prisoner's Dilemma, the answer is that *defect* should be chosen. Intuitively, however, it seems possible for the players to ask a different question: 'What should *we* do?' In Hi-Lo, the answer to *this* question is surely: 'Choose (*high, high*)'. In the Prisoner's Dilemma, 'Choose (*cooperate, cooperate*)' seems to be at least credible as an answer. Theories of team agency try to reformulate game theory in such a way that 'What should we do?' is a meaningful question. The basic idea is that, when an individual reasons as a member of a team, she considers which *combination* of actions by members of the team would best promote the team's objective, and then performs her part of that combination. The rationality of each individual's action derives from the rationality of the joint action of the team.

2. Simple team reasoning

In propositional logic, a rule of inference—a rule that allows us to derive conclusions from premises—is valid if, whenever the premises are true, so are the conclusions that are derived from them. Here is a simple example of valid reasoning (the propositions above the line are premises, while the proposition below the line is the conclusion):

(1) There are no English mountains over 1000 metres high.
(2) Snowdon is a mountain which is 1085 metres high.

Snowdon is not in England.

One can formulate principles of *practical reasoning*—that is, reasoning that leads to conclusions about what an agent should do—which satisfy analogous criteria of validity. Bacharach (2000) defines a mode of reasoning

as valid in games if it is *success-promoting*: given any game of some very broad class, it yields only choices which tend to produce success, as measured by game payoffs. The fundamental idea is that practical reasoning infers conclusions about what an agent ought to do from premises which include propositions about what the agent is seeking to achieve. Such reasoning is *instrumental* in that it takes the standard of success as given; its conclusions are propositions about what the agent should do in order to be as successful as possible according to that standard. If the agent is an individual person, the reasoning is *individually instrumental*. Schema 12.1 shows a simple example of individually instrumental reasoning.

Schema 12.1: Individual rationality

(1) I must choose either *left* or *right*.
(2) If I choose *left*, the outcome will be O_1.
(3) If I choose *right*, the outcome will be O_2.
(4) I want to achieve O_1 more than I want to achieve O_2.

I should choose *left*.

In our analysis, we will interpret the payoffs of a game as specifying what the players want to achieve as individuals or, equivalently, what counts as success for them. Following the conventions of game theory, we will treat payoffs as utility indices in the sense of expected utility theory so that, in situations of uncertainty, a player's success is measured by the expected value of her payoff. Thus, the following individually instrumental reasoning for Player 1 in the Hi-Lo game of figure 12.2 is valid:

(1) I am Player 1 in Hi-Lo.
(2) The probability that Player 2 will choose *high* is 0.5.

I should choose *high*.

This is an example of the kind of *best-reply reasoning* that is analysed in classical game theory. Of course, if we assume only that the players of Hi-Lo have common knowledge of the payoffs of the game and of their rationality, premise (2) is not available to Player 1 and so, although the schema we have described is valid, Player 1 cannot use it to get to the conclusion 'I should choose *high*'.

But now consider schema 12.2, in which (*left, right*) denotes the pair of actions 'I choose *left*, you choose *right*':

Schema 12.2: Collective rationality

(1) We must choose one of (*left, left*), (*left, right*), (*right, left*) or (*right, right*).
(2) If we choose (*left, left*) the outcome will be O_1.
(3) If we choose (*left, right*) the outcome will be O_2.
(4) If we choose (*right, left*) the outcome will be O_3.
(5) If we choose (*right, right*) the outcome will be O_4.
(6) We want to achieve O_1 more than we to achieve O_2, O_3 or O_4.

We should choose (*left, left*).

Is this schema valid? Given the symmetries between schemata 1 and 2, it seems that, if one is valid, so too is the other. Yet the two schemata seem to be potentially contradictory. For example, consider the Prisoner's Dilemma. Using a variant of schema 12.1, each player can reason to 'I should choose *defect*'; but using a variant of schema 12.2, each can reason to 'We should not choose (*defect, defect*)'. (In the second case, the reason for not choosing (*defect, defect*) is that the outcome resulting from (*defect, defect*) is one that we want less than we want the outcome of (*cooperate, cooperate*).) We suggest that the resolution of this problem is that, properly understood, the two sets of premises are mutually inconsistent.[6] The premises of schema 12.1 presuppose that *I* am an agent, pursuing *my* objectives. Those of schema 12.2 presuppose that *we* make up a single unit of agency, pursuing *our* objectives. But instrumental practical reasoning presupposes a unit of agency. If I am to reason instrumentally, I cannot simultaneously think of myself both as a unit of agency in my own right and as part of a unit of agency which includes you.[7]

We can make this feature of practical reasoning more transparent by writing schemata in forms which include premises about agency. Consider any situation in which each of a set *S* of individuals has a set of alternative *actions*, from which he must choose one.[8] A *profile* of actions assigns to each member of *S* one element of his set of alternative actions. For each profile, there is an *outcome*, understood simply as the state of affairs that comes about (for everyone) if those actions are chosen. We define a *payoff function* as a function which assigns a numerical value to every outcome. A payoff function is to be interpreted as representing what some specific agent wants to achieve: if one outcome has a higher numerical value than another, then the relevant agent wants to achieve the first more than he (or she, or it) wants to achieve the second. Now consider any individual

i, and any set of individuals *G*, such that *i* is a member of *G* and *G* is a weak subset of *S*. We will say that *i identifies with G* if *i* conceives of *G* as a unit of agency, acting as a single entity in pursuit of some single objective.[9] Finally, we define *common knowledge* in the usual way: a proposition *x* is common knowledge in a set of individuals *G* if: (i) *x* is true; (ii) for all individuals *i* in *G*, *i* knows *x*; (iii) for all individuals *i* and *j* in *G*, *i* knows that *j* knows *x*; (iv) for all individuals *i*, *j*, and *k* in *G*, *i* knows that *j* knows that *k* knows that *x*; and so on.

Letting *A* stand for any profile and *U* for any payoff function, consider the schema 12.3.

Schema 12.3: Simple team reasoning (from a group viewpoint)
(1) We are the members of *S*.
(2) Each of us identifies with *S*.
(3) Each of us wants the value of *U* to be maximized.
(4) *A* uniquely maximizes *U*.

Each of us should choose her component of *A*.

This schema captures the most basic features of team reasoning. Notice that, because of (2), the schema does not yield any conclusions unless all the members of *S* identify with this group. Because of (4), the schema yields conclusions only when a profile that is the unique maximizer of the team payoff function exists. We will not address the question of what a team reasoner should do when this is not the case but, for our purposes, the answer is not essential. Notice also that we can apply schema 12.3 in cases in which *S* contains only one individual. In this case, *S* can be written as {myself}. (1) then becomes 'I am the only member of the set {myself}'. (2) reduces to 'I identify with {myself}', which amounts to saying that the reasoning individual views herself as an agent. And then the schema represents straightforward practical reasoning by an individual agent. Thus, schema 12.3 encompasses both individual and team reasoning.

Schema 12.3 represents a mode of reasoning that can be used by people *as a group*. What does it mean for a number of people to reason as a group? One way to make sense of this is to imagine those people in an open meeting, at which each of a set of premises is announced, and acknowledged as true by each person. Then, the inference to be drawn from those premises is announced, and acknowledged as valid by each

person. In such a setting, it is common knowledge among the members of the group that each of them accepts the relevant premises. That this is common knowledge does not need to be stated explicitly in the schema; it is not an additional premise, but a presupposition of the whole idea of reasoning as a group.

For many purposes, however, it is more convenient to represent team reasoning from the viewpoint of an individual team member. If we adopt this approach, schema 12.3 can be rewritten as in schema 12.4.

Schema 12.4: Simple team reasoning (from an individual viewpoint)

(1) I am a member of *S*.
(2) It is common knowledge in *S* that each member of *S* identifies with *S*.
(3) It is common knowledge in *S* that each member of *S* wants the value of *U* to be maximized.
(4) It is common knowledge in *S* that *A* uniquely maximises *U*.

I should choose my component of *A*.

We now consider the implications of schema 12.4 for Hi-Lo and the Prisoner's Dilemma, on the assumption that, in each game, it is common knowledge that each player identifies with the two-player group {Player 1, Player 2}. This assumption is used merely as a convenient starting point; later, we will relax it.

First, consider Hi-Lo. In order to apply schema 12.4, we need to define a payoff function *U* to represent what each individual wants to achieve, given that she identifies with {Player 1, Player 2}. We shall assume that, when a player identifies with a group, she wants to promote the combined interests of its two members, at least insofar as those interests are affected by the game that is being played. Thus, the values of *U* can be interpreted as measures of the welfare of the group {Player 1, Player 2}. Since the two players' payoffs are equal, irrespective of which actions are chosen, it is natural to make the values of *U* equal to the players' common payoffs. Then (*high, high*) is the profile that uniquely maximizes *U*, and so (provided there is common knowledge of the rules of the game), each player can use schema 12.4 to reach the conclusion that she should choose *high*.

Now, consider the Prisoner's Dilemma. Again, we need to define a payoff function U for the group {Player 1, Player 2}. If we assume that U treats the players symmetrically, we need to specify only three values of this function: the payoff when both players choose *cooperate*, which we denote u_C, the payoff when both choose *defect*, which we denote u_D, and the payoff when one chooses *cooperate* and one chooses *defect*, which we denote u_F (for 'freeriding'). It seems unexceptionable to assume that U is increasing in individual payoffs, which implies $u_C > u_D$. Given the condition $b > (a + d)/2$, it is natural also to assume $u_C > u_F$. Then the profile of actions by Player 1 and Player 2 that uniquely maximises U is (*cooperate, cooperate*). If there is common knowledge of the rules of the game, each player can use schema 12.4 to reach the conclusion that she should choose *cooperate*.

3. Is team reasoning necessary to solve the Prisoner's Dilemma?

In the analysis we have just outlined, a rational player of the one-shot Prisoner's Dilemma can choose *cooperate*. For many game theorists, this conclusion is close to heresy. For example, Ken Binmore (1994: 102–17, quotation from p. 114) argues that it can be reached only by 'a wrong analysis of the wrong game': if two players truly face the game shown in figure 12.1, then it follows from the meaning of 'payoff' and from an unexceptionable concept of rationality that a rational player must choose *defect*. His argument works as follows. Consider Player 1. She knows that her opponent must choose either *cooperate* or *defect*. The inequality $a > b$ tells us that, if Player 1 knew that Player 2 would choose *cooperate*, Player 1 would want to choose, and would choose, *defect*. The inequality $c > d$ tells us that, if Player 1 knew that Player 2 would choose *defect*, Player 1 would want to choose, and would choose, *defect*. So (Binmore concludes) we need only a principle of dominance to conclude that, whatever Player 1 believes about what Player 2 will do, Player 1 should choose *defect*. Binmore recognizes that rational individuals may sometimes choose *cooperate* in games in which *material payoffs*—that is, outcomes described in terms of units of commodities which people normally prefer to have more of rather than less, such as money, or years of not being in prison—are as in figure 12.1.

But that just shows that the payoffs that are relevant for game theory—the payoffs that govern behaviour—differ from the material ones. The first stage in a game-theoretic analysis of a real-life situation should be to find a formal game that correctly represents that situation.

Thus, in response to the problem of explaining why *cooperate* is sometimes chosen in games whose material payoffs have the Prisoner's Dilemma structure, the methodological strategy advocated by Binmore is that of *payoff transformation*: we should look for some way of transforming material payoffs into game-theoretic ones which makes observed behaviour consistent with conventional game-theoretic analysis. It has been followed by various theorists who have proposed transformations of material payoffs to take account of psychological or moral motivations that go beyond simple self-interest.

One of the earliest proposals of this kind was made by Amartya Sen (1974, 1977). Sen distinguishes between 'rationality' (as this is usually understood in economics) and 'morality'. He points out that many different codes of morality would prescribe cooperation in the Prisoner's Dilemma. More generally, many moral codes value actions which 'sacrific[e] some individual gain—given the action of others—for the sake of a rule of good behaviour by all which ultimately makes everyone better off'. Sen proposes an approach in which this core moral principle is 'expressed in the form of choice between preference patterns rather than between actions' (1974: 77–8). His idea is to represent different attitudes towards behaviour in a given game as different orderings over the outcomes of the game (or over the strategy profiles that generate those outcomes). Some of these orderings are egoistic, but others are not. Although a person's moral principles are described by a meta-ranking of the set of alternative orderings, her actual behaviour is explained by whichever of those orderings she chooses to act on. A person who chooses to act on non-egoistic preferences is said to act on *commitment*, and her choices are said to be *counterpreferential* (1977: 91–3). Nevertheless, Sen's account of what is involved in such action retains the formal structure of conventional decision and game theory: the unit of agency is the individual, and each individual's actions are governed by some ordering over outcomes. In game-theoretic terms, commitment induces a transformation from egoistic payoffs to 'counterpreferential' ones.

More recently, and in slightly different ways, Ernst Fehr and Klaus Schmidt (1999) and Gary Bolton and Axel Ockenfels (2000) have proposed

that, for any given level of material payoff for any individual, that individual dislikes being either better off or worse off than other people. Matthew Rabin (1993) proposes that each individual likes to benefit people who act with the intention of benefiting him, and likes to harm people who act with the intention of harming him.[10]

The theory of team reasoning can accept Binmore's instrumental conception of rationality, but rejects his implicit assumption that agency is necessarily vested in individuals. We can interpret the payoffs of a game, as represented in a matrix like that of figure 12.1, as showing what each player wants to achieve *if she takes herself to be an individual agent*. In this sense, the interpretation of the payoffs is similar to that used by Binmore: payoffs are defined, not in material terms, but in terms of what individuals are seeking to achieve. The theory of team reasoning can replicate Binmore's analysis when it is applied to players who take themselves to be individual agents: if Player 1 frames the game as a problem 'for me', the only rational choice is *defect*. However, the theory also allows the possibility that Player 1 frames the game as a problem 'for us'. In this case, the payoffs that are relevant in determining what it is rational for Player 1 to do are measures of what she wants to achieve *as a member of the group* {*Player1*, *Player2*}; and these need not be the same as the payoffs in the standard description of the game.

Thus, there is a sense in which team reasoning as an explanation of the choice of *cooperate* in the Prisoner's Dilemma depends on a transformation of payoffs from those shown in figure 12.1. However, the kind of transformation used by theories of team reasoning is quite different from that used by theorists such as Fehr and Schmidt. In team reasoning, the transformation is not from material payoffs to choice-governing payoffs; it is from payoffs which govern choices for one unit of agency to payoffs which govern choices for another. Thus, payoff transformation takes place as part of a more fundamental *agency transformation*.

One might wonder whether we need to transform *both* payoffs *and* agency. If payoffs have been transformed so that they represent the welfare of the two players as a group, doesn't conventional game theory provide an explanation of why each individual chooses *cooperate*? Not necessarily. Consider a Prisoner's Dilemma in which $a = 10$, $b = 8$, $c = 6$ and $d = 0$, and assume that value of the payoff function for the group {Player 1, Player 2} is the average of the payoffs for the two individuals. Then we have $u_C = 8$, $u_D = 6$ and $u_F = 5$. If we treat Player 1 and Player 2 as individual

agents, each of whom independently seeks to maximize the value of U, we have the game shown in figure 12.3. The structure of this game will be familiar: it is a variant of Hi-Lo, in which *cooperate* corresponds with *high* and *defect* with *low*. Conventional game theory does not show that rational players of this game will choose *cooperate*. To show that, we need a transformation of the unit of agency (figure 12.3).

		Player 2	
		cooperate	*defect*
Player 1	*cooperate*	8, 8	5, 5
	defect	5, 5	6, 6

Figure 12.3. A Prisoner's Dilemma with transformed payoffs

By using the concept of agency transformation, team reasoning is able to explain the choice of *high* in Hi-Lo. Existing theories of payoff transformation cannot do this. Further, it is hard to see how *any* such theory could credibly make (*high, high*) the unique solution of Hi-Lo. Let us interpret the Hi-Lo payoffs as material payoffs, and consider possible transformations. Suppose that, following Fehr and Schmidt and Bolton and Ockenfels, we introduce assumptions about players' attitudes towards the distribution of material payoffs. In every possible outcome, the two players' material payoffs are equal. Whatever the players' attitudes to inequality, it seems that their subjective ranking of the outcomes must correspond with the ranking of material payoffs. Thus, a game which is Hi-Lo in material payoffs will remain Hi-Lo after payoff transformation. Alternatively, suppose we follow Rabin and assume that each player wants to reciprocate other players' 'kindness' or 'unkindness' towards him. In a situation in which both players choose *low*, Player 1 is benefiting Player 2 to the maximum degree possible, given Player 2's action; and vice versa. So each is reciprocating the other's 'kindness'. Reciprocity in Rabin's sense does not affect the equilbrium status of (*low, low*).

One might reasonably expect a theory of rational choice to account for the intuition that *high* is the rational choice in a Hi-Lo game, and to explain why this strategy is in fact chosen by apparently rational players. The theory of team reasoning meets this requirement. Once the components of this theory are in place, very little more is needed to explain the choice of *cooperate* in the Prisoner's Dilemma. All that is needed in addition is the assumption that (*cooperate, cooperate*) is the best profile of actions for the two

players together. That assumption is hardly controversial: it is presupposed in most accounts of the significance of the Prisoner's Dilemma—whether that is understood as a puzzle for game theory or as a model of real-world problems of cooperation.

4. Comparing theories of team agency

The reasoning represented by schema 12.4 is the common core of theories of team agency. However, there are various theories of team agency, which differ in important ways. They differ in their hypotheses about how teams are formed, or how individual agents come to identify with groups. Group formation has been claimed to be: a requirement of rationality/ morality, a response to the psychological impetus of framing, the result of explicit mutual commitment, or a consequence of non-rational assurance. Further, in schema 12.4, there is common knowledge within the group *S* that each member of *S* identifies with *S*. In some situations, this is not a realistic assumption. It may be the case that some members of *S* (that is, the group with which team reasoners identify) do not identify with *S*. In this case, we can define the *team T* as those members of *S* who do identify with *S*. The various theories differ in how they recommend that members of *T* should reason in such cases.

4.1. Team agency required by rationality/morality

The first theorists to discuss team reasoning did so in the context of moral and rational requirements on action. Hodgson (1967) was the first person to use the Hi-Lo game, as part of an argument that rule utilitarianism does not reduce to act utilitarianism. Regan (1980) proposed a form of team reasoning in his theory of *cooperative utilitarianism*. Regan's theory is normative; it is commended to all of us in our capacities as rational and moral agents. The fundamental principle of this theory is that 'what each agent ought to do is to co-operate, with whoever else is co-operating, in the production of the best consequences possible given the behaviour of non-co-operators' (1980: 124). For the moment, consider a world where everyone is a rule utilitarian. In that case, Regan's rational and moral agents reason according to schema 12.4. In a similar vein, Hurley (1989: 136–59) proposes that we (as rational and moral agents) should specify agent-neutral

goals—that is, goals of which it can simply be said that they ought to be pursued, rather than they ought to be pursued by some particular agent. Then we should 'survey the units of agency that are possible in the circumstances at hand and ask *what the unit of agency, among those possible, should be*'; and we should 'ask ourselves *how we can contribute to the realization of the best unit possible in the circumstances*'.

Regan's theory gives recommendations for cases in which not everyone is a cooperative utilitarian. The logic of these recommendations can be represented by a variant of team reasoning called *restricted team reasoning* by Bacharach (2006). This applies to cases in which it is known that certain specific members of S do not identify with S. It is formalized in schema 12.5. Let A_T be a profile of actions for the members of T. Then:

Schema 12.5: Restricted team reasoning

(1) I am a member of T.
(2) It is common knowledge in T that each member of T identifies with S.
(3) It is common knowledge in T that each member of T wants the value of U to be maximized.
(4) It is common knowledge in T that A_T uniquely maximises U, given the actions of non-members of T.

I should choose my component of A_T.

In cooperative utilitarianism, each of us is told to join with as many others as are willing to do the same, and to cooperate with them in trying to achieve the overall good of all people (or perhaps the good of all sentient beings—but *not* just the common good of the members of T). In terms of schema 12.5, S is the set of all people, T is the set of cooperative utilitarians, and the value of U is a utilitarian measure of overall goodness. Regan claims that a cooperative utilitarian *ought* to identify with S, and that she *ought* to want to maximize U. Whilst Regan's theory tells us which group we should identify with, and what we should want to maximize, Hurley does not nominate any particular group or any particular goal as being the rational one to pursue. Nevertheless, for Hurley, the idea seems to be that rationality requires each person to choose the unit of agency in which she participates, and that this choice should be governed by goals which are independent of the unit of agency.

4.2. Team agency as the result of framing

In contrast, Bacharach's (2006) theory does not allow the unit of agency to be chosen, and does not admit the concept of a goal that is not the goal of some agent. For Bacharach, whether a particular player identifies with a particular group is a matter of 'framing'. A *frame* is the set of concepts a player uses when thinking about her situation. In order to team reason, a player must have the concept 'we' in her frame. Bacharach proposes that the 'we' frame is normally induced or *primed* by Hi-Lo games, but is primed less reliably by the Prisoner's Dilemma. Both games have a property that Bacharach calls *strong interdependence*. Roughly, a game has this property if it has a Nash equilibrium which is Pareto-dominated by the outcome of some feasible strategy profile. (For a more formal definition, see Bacharach, 2006.) Although Bacharach proposes that the perception of this property increases the probability of group identification, he does not claim that games with this property *invariably* prime the 'we' frame. More specifically:

> In a Prisoner's Dilemma, players might see only, or most powerfully, the feature of common interest and reciprocal dependence which lie in the payoffs on the main diagonal. But they might see the problem in other ways. For example, someone might be struck by the thought that her coplayer is in a position to double-cross her by playing [*defect*] in the expectation that she will play [*cooperate*]. This perceived feature might inhibit group identification. (2006, chapter 2, section 4.2)

The implication is that the 'we' frame *might* be primed; but, alternatively, a player may see the game as one to be played by two separate individual agents. That either framing is psychologically possible reflects the sense in which the Prisoner's Dilemma itself is puzzling. One the one hand, the positions of the two players are completely symmetrical, which prompts one to focus on strategy profiles in which the two players' actions are symmetrical. Then, comparing the outcomes of (*cooperate, cooperate*) and (*defect, defect*), one sees that the two players have a common interest in their both choosing *cooperate*. This line of thought leads naturally to a conception of the game as a problem 'for us'. On the other hand, if one looks at the outcomes of (*cooperate, defect*) and (*defect, cooperate*), one sees a conflict of interest between the two players: by choosing *defect* when one's opponent chooses *cooperate*, one can gain at her expense. This line of thought leads to a conception of the game as one in which the two players

are in opposition, each facing her own decision problem. As a metaphor or model, Bacharach often refers to the famous drawing (used in *Gestalt* psychology) which can be seen either as a duck or a rabbit. In the same way, the Prisoner's Dilemma can be seen by a player either as a problem 'for me' or as a problem 'for us'. Thus, we should not assume it to be common knowledge that the players of the Prisoner's Dilemma identify with {Player 1, Player 2}.

In Bacharach's theoretical framework, this dualism is best represented in terms of *circumspect team reasoning*. We now present this mode of reasoning in the form of a reasoning schema. As before, let S be the set of individuals with which team-reasoners identify, and let T be any subset of S, interpreted as the set of individuals who in fact identify with S.[11] Suppose there is a random process which, independently for each member of S, determines whether or not that individual is a member of T; for each individual, the probability that he is a member of T is ω, where $\omega > 0$. We define a proposition p to be *T-conditional common knowledge* if: (i) p is true; (ii) for all individuals i in S, if i is a member of T, then i knows p; (ii) for all individuals i and j in S, if i is a member of T, then i knows that if j is a member of T, then j knows p; and so on. (As an illustration: imagine an underground political organization which uses a cell structure, so that each member knows the identifies of only a few of her fellow-members. New members are inducted by taking an oath, which they are told is common to the whole organization. Then, if T is the set of members, the content of the oath is T-conditional common knowledge.) We define a *protocol* as a profile of actions, one for each member of S, with the interpretation that the protocol is to be followed by those individuals who turn out to be members of T. Let P be any protocol. The schema is:

Schema 12.6: Circumspect team reasoning

(1) I am a member of T.
(2) It is T-conditional common knowledge that each member of T identifies with S.
(3) It is T-conditional common knowledge that each member of T wants the value of U to be maximized.
(4) It is T-conditional common knowledge that P uniquely maximizes U, given the actions of non-members of T.

I should choose my component of P.

We can apply this schema to the Prisoner's Dilemma by setting $S =$ {Player 1, Player 2} and by defining U as before. Let ω (where $0 < \omega \leqslant 1$) be the probability that, for any individual player of the Prisoner's Dilemma, the 'we' frame comes to mind; if it does, the player identifies with {Player 1, Player 2}. Assume that, if this frame does *not* come to mind, the player conceives of himself as a unit of agency and thus, using best-reply reasoning, chooses the dominant strategy *defect*. We can now ask which protocol maximizes U, given the value of ω. Viewed from within the 'we' frame, the protocol (*defect, defect*) gives a payoff of u_D with certainty. Each of the protocols (*cooperate, defect*) and (*defect, cooperate*) gives an expected payoff of $\omega u_F + (1 - \omega)u_D$. The protocol (*cooperate, cooperate*) gives an expected payoff of $\omega^2\ u_C + 2\omega(1 - \omega)u_F + (1 - \omega)^2 u_D$. There are two possible cases to consider. If $u_F \geq u_D$, then (*cooperate, cooperate*) is the U-maximizing protocol for all possible values of ω. Alternatively, if $u_D > u_F$, which protocol maximizes U depends on the value of ω. At high values of ω, (*cooperate, cooperate*) is uniquely optimal; at low values, the uniquely optimal protocol is (*defect, defect*).[12]

If we assume *either* that $u_F \geq u_D$ *or* that the value of ω is high enough to make (*cooperate, cooperate*) the uniquely optimal protocol, we have a model in which players of the Prisoner's Dilemma choose *cooperate* if the 'we' frame comes to mind, and *defect* otherwise. Bacharach offers this result as an explanation of the observation that, in one-shot Prisoner's Dilemmas played under experimental conditions, each of *cooperate* and *defect* is usually chosen by a substantial proportion of players. He also sees it as consistent with the fact that there are many people who think it completely obvious that *cooperate* is the only rational choice, while there are also many who feel the same about *defect*. Bacharach can say that both sets of people are right—in the same way that two people can both be right when one says that the drawing they have been shown is a picture of a duck and the other says it is a picture of a rabbit.

Bacharach claims that schema 12.6 is valid, with the implication that, for any given individual, *if she identifies with S and wants U to be maximized*, it is instrumentally rational for her to act as a member of T, the team of like-minded individuals. He does not claim that she *ought* to identify with any particular S, or that she *ought* to want any particular U to be maximized. In the theory of circumspect team reasoning, the parameter ω is interpreted as a property of a psychological mechanism—the probability

that a person who confronts the relevant stimulus will respond by framing the situation as a problem 'for us'. The idea is that, in coming to frame the situation as a problem 'for us', an individual also gains some sense of how likely it is that another individual would frame it in the same way; in this way, the value of ω becomes common knowledge among those who use this frame.

4.3. Team agency produced by commitment

Another variety of team agency has it that a group is constituted by public acts of promising, or by public expressions of commitment by its members. This latter idea is central to Gilbert's (1989) analysis of 'plural subjects'. Although Gilbert is more concerned with collective attitudes than with collective action, her analysis of how a plural subject is formed might be applied to the formation of teams. There are also hints of this approach in the work of Hollis (1998). Hollis suggests that Rousseau's (1762/ 1988) account of the social contract, with its 'most remarkable change in man', can be understood as a transition from individual to group agency that takes place through a collective act of commitment.

However, the commitment and the framing understandings of group identity may be more similar than Bacharach's and Gilbert's formal analyses suggest. In notes for a chapter of his book that remained unwritten at the time of his death, Bacharach (2000) records the following train of thought about the formation of teams:

> My current … idea is something like this: Something in the situation prompts the parties to see that they have action possibilities which provide joint agency possibilities which have possible outcomes of common interest. Each finds herself in a frame which features concepts which describe the conceived possible actions, describe the conceived possible outcomes, and present some of these outcomes positively. Each of us sees that we could write a paper together, or have a pleasant walk round the garden together, or bring down the appalling government together.
>
> The holism of frames comes in here. One concept belonging to the frame may bring others with it, or only be activated if others are. Some actions only get conceived if one gets the idea of certain possible outcomes, and conversely. One such situation is that created by one of us being so prompted, then making a verbal suggestion to the other(s), as in 'Would you like to dance?'

This account has many similarities with Gilbert's, except that what Gilbert describes in the language of agreement and tacit understanding,

Bacharach describes in terms of framing. Bacharach's concept of framing allows one person to choose an action with the intention of affecting *someone else's* frame. While Gilbert would treat the saying of the words 'Would you like to dance?' as the first stage in a process which may lead to a common understanding that the two people are a plural subject in relation to a dance, Bacharach treats it as part of a process by which two people influence one another's frames. Bacharach cannot say (as Gilbert might) that each individual *chooses* (or *agrees*, or *commits himself*) to view the situation in the 'we' frame. But he can, and does, say that group identification tends to be primed by an individual's recognition that the members of the putative group have common interests that can be furthered by joint action. So, although Bacharach's individuals cannot choose to create teams with the rational intention of solving problems of coordination or cooperation, such problems tend to induce the kind of mutual adjustment of frames that he describes.

Although Gilbert does not offer an explicit model of collective choice, we suggest that schema 12.4 is compatible with her general approach, provided that 'identifying with' the group *S* is understood as some kind of conscious and public act of commitment. On this interpretation, however, the schema is not one of instrumental reasoning. Rather, the rationality of acting as a member of a team derives from the rationality of fulfilling one's commitments or intentions. Focusing on collective attitudes rather than collective actions, Gilbert claims that membership of a plural subject imposes obligations to uphold 'our' attitudes. This claim is conceptual rather than moral: roughly, the idea is that a plural subject is formed by an exchange of commitments, and that to make a commitment is to impose on oneself an obligation to act on it. For Gilbert, there is no problem that *S* (the group with which individuals identify) may be different from *T* (the set of individuals who in fact identify with *S*). In commitment-based theories, it is natural to assume that the set of individuals who act as a team is the same as the group with which they identify, provided we can assume that individuals keep their commitments.

4.4. Team agency and assurance

Schemata 4, 5 and 6 share an important common feature. In each case, the conclusion is an *unconditional* proposition of the form 'I should choose my component of the best profile'. The unconditional form of this conclusion

is crucial in the resolution of the problem posed by Hi-Lo. In contrast, the best-reply reasoning of classical game theory leads to conclusions about what one agent should do *conditional on what other agents can be expected to do*. Thus, in Hi-Lo, best-reply reasoning leads only to the conclusion 'If I expect my opponent to choose her component of the best profile, I should choose mine', and so to an infinite regress.

Schema 12.4 has an additional feature, not shared by the other two team reasoning schemata. Even though it yields an unconditional conclusion, it tells an individual member of S to choose his component of the best profile only in situations in which it also tells the other members to choose theirs. Further, these are always situations in which the other players identify with S: they are framing the decision problem as one 'for us'. Thus, if each of them is rational, each will act on the conclusions of schema 12.4, as applied to her case. And, since it is common knowledge in S that everyone identifies with S, each player can work all this out. So, whenever schema 12.4 tells an individual to choose his component in the best profile, that individual has the *assurance* that the others (if rational) will choose theirs too. To put this another way: when the individual chooses his component in the best profile, he can construe this as his part of a collective action that is in fact taking place.

However, this property of assurance does not carry over to the theories of restricted and circumspect team reasoning. In these theories, each member of the team T identifies with S. Thus, each member of T wants the value of U to be maximized, where U represents what people want *as members of* S. Each member of T is told to do his part in a joint action by T to maximize U, given the behaviour of non-members of T; he can be assured only that *the other members of* T will do their parts. For example, consider the Prisoner's Dilemma under the assumption that $u_F > u_D$ (that is, it is better for the two players as a group that one of them chooses *cooperate* and the other chooses *defect* than that both choose *defect*). Suppose that Player 1 identifies with the group $S = \{\text{Player 1, Player 2}\}$; but suppose also (the case of restricted team reasoning) that Player 1 knows that Player 2 does not identify with S, or (the case of circumspect team reasoning) that Player1 knows that the probability that Player 2 identifies with S is close to zero. Restricted and circumspect team reasoning both lead to the conclusion that Player 1 should choose *cooperate*, even though he knows (or is almost certain) that Player 2 is taking a free ride.

On some interpretations of the concept of team reasoning, it involves an idea of *reciprocity*: each member of a team does her part on the understanding that others will do theirs. If team reasoning is viewed in this way, assurance may seem an essential concept. This provides the starting point for yet another interpretation of the status of team reasoning.

Sugden (2003) presents a 'logic of team reasoning' without making any claims for its validity. In his analysis, a 'logic' is merely an internally consistent system of axioms and inference rules. An individual actor may *endorse* a particular logic, thereby accepting as true any conclusions that can be derived within it, but the theorist need not take any position about whether the axioms of that logic are 'really' true or whether its inference rules are 'really' valid. Team reasoning is then represented as a particular inference rule which, as a matter of empirical fact, many people endorse. Thus, following this approach, one might re-interpret schema 12.4 as specifying the inference rule 'From (1), (2), (3) and (4), infer "I should choose my component of A" '.

On this interpretation, however, schema 12.4 does not guarantee assurance. Recall that this schema recommends an individual to choose his component of the U-maximizing profile of actions by the members of the group S, only in situations in which it also recommends the other members to choose their components. So, *if* it can be assumed that schema 12.4 has a validity that is transparent to all rational people, and *if* it can be assumed that each member of S is confident that the others are rational, *then* each member of S has the assurance that, when he chooses his component, the others will choose theirs. But Sugden's approach does not acknowledge agent-neutral concepts of 'validity' and 'rationality'. It maintains assurance in a different way, which we now formulate as a reasoning schema.

Following David Lewis (1969) and Robin Cubitt and Sugden (2003), Sugden uses a theoretical framework in which the central concept is *reason to believe*. To say that a person has reason to believe a proposition p is to say that p can be inferred from propositions that she accepts as true, using rules of inference that she accepts as valid. On the analogue of the definition of common knowledge, there is *common reason to believe* a proposition p in a set of individuals T if: (i) for all individuals i in T, i has reason to believe p; (ii) for all individuals i and j in T, i has reason to believe that j has reason to believe p; (iii) for all individuals i, j, and k in T, i has reason to believe that j has reason to believe that k has reason to believe p; and so on.

The following definition is also useful. Within a set of individuals T, there is *reciprocal reason to believe* that some property q holds for members of T if (i) for all individuals i and j in T, where $i \neq j$, i has reason to believe that q holds for j; (ii) for all individuals i, j, and k in T, where $i \neq j$ and $j \neq k$, i has reason to believe that j has reason to believe that q holds for k; and so on. To see the point of this latter definition, consider the Prisoner's Dilemma and let q be the property 'chooses *cooperate*'. In a schema of practical reasoning which is intended to be used by Player 1 in deciding how to play the Prisoner's Dilemma, we cannot allow the premise that, in the group {Player 1, Player 2}, there is common reason to believe that Player 1 chooses *cooperate*. That would make it a premise that Player 1 has reason to believe that he himself will choose *cooperate*, when the whole point of using the schema is to determine which action he should choose. However, we *can* allow the premise that there is reciprocal reason to believe that members of the group {Player 1, Player 2} choose *cooperate*, and there may be circumstances in which such a premise would be natural. For example, suppose that Player 1 and Player 2 have played the Prisoner's Dilemma many times before, and on every such occasion, both have chosen *cooperate*. They are about to play again, and there is no obvious difference between this interaction and all its predecessors. Then, by induction, Player 2 might have reason to believe that Player 1 will choose *cooperate*. Attributing this reasoning to his opponent, Player 1 might have reason to believe that Player 2 has reason to believe that Player 1 will choose *cooperate*; and so on.

Sugden's formulation of team reasoning can be represented as shown in schema 12.7.

Schema 12.7: Mutually assured team reasoning

(1) I am a member of S.
(2) I identify with S and acknowledge U as its objective.
(3) In S, there is reciprocal reason to believe that each member of S identifies with S and acknowledges U as the objective of S.
(4) In S, there is reciprocal reason to believe that each member of S endorses and acts on mutually assured team reasoning.
(5) In S, there is common reason to believe that A uniquely maximizes U.

I should choose my component of A.

This schema is not presented as a mode of *valid* reasoning. It is merely a mode of reasoning that any person might (or might not) endorse; a person commits herself to team reasoning by endorsing the schema. Notice that premises (2) and (3) refer to 'acknowledging *U* as the objective of *S*' rather than 'wanting *U* to be maximized'. On Sugden's account, a team reasoner who identifies with a group stands ready to do her part in joint actions in pursuit of the group's objective; but she does not necessarily take this objective as *hers* in the stronger sense of wanting to pursue it even if other members of the group do not reciprocate.

Schema 12.7 is recursive: premise (4) refers to the endorsement of the schema itself. That this is not circular can be seen from an analogy. Consider an international treaty which includes among its conditions that it will come into effect only if and when it has been ratified by a certain number of nations; once this condition is met, it is binding on every nation that has ratified it. To ratify such a treaty is to make a commitment which is binding from that moment, but which is activated only if enough others make the same commitment. Analogously, to endorse mutually assured team reasoning is to make a unilateral commitment to a certain form of practical reasoning, but this reasoning does not generate any implications for action unless one has assurance that others have made the same commitment. Such assurance could be created by public acts of commitment of the kind considered by Gilbert. But it could also be induced by repeated experience of regularities of behaviour in a population. For example, suppose that in some population, some practice of mutual assistance (say, giving directions to strangers when asked) is generally followed in anonymous encounters. Each individual might interpret the existence of the practice as evidence that premises (3), (4) and (5) are true. If so, each individual would be assured that others would choose their components of the U-maximizing profile. But he would still have to decide whether team reasoning was a mode of reasoning that he wanted to endorse.

5. Collective intentions

Whilst the problems that motivated the literature on team agency are about why agents would take certain *actions*, the literature on collective intentions analyses agents' *mental states*. When an agent deliberates about what she

ought to do, the result of her reasoning is an intention. An intention is interposed between reasoning and an action, so it is natural to treat the intentions that result from team reasoning as collective intentions.

An early analysis of collective intentionality is the work of Tuomela and Miller (1988). The essential features of this analysis can be presented as follows for the case of a two member group, whose members are Player 1 and Player 2. Consider some 'joint social action' A which comprises actions A_1 and A_2 for the respective individuals. According to Tuomela and Miller, Player 1 has a *we-intention* with respect to A if: (i) Player 1 intends to do A_1, (ii) Player 1 believes that Player 2 will do A_2, (iii) Player 1 believes that Player 2 believes that Player 1 will do A_1, and so on (p. 375). This analysis reduces we-intentions to individual intentions and a network of mutual beliefs.

An apparently unsatisfactory feature of this analysis is that it seems to treat every Nash equilibrium as a case of collective intentionality. For example, consider the version of the Hawk–Dove game shown in figure 12.4.

		Player 2	
		dove	*hawk*
Player 1	*dove*	2, 2	0, 3
	hawk	3, 0	–5, –5

Figure 12.4. Hawk–Dove

As an example of such a game, think of two individuals in a state of nature who come into conflict over some valuable resource. To play *dove* is to offer to share the resource but to back down if the other attempts to take it all; to play *hawk* is to demand the whole resource, backed by a readiness to fight for it. We assume that fighting is costly for both parties, and that the utility value of a half share of the resource is greater than half of the utility value of the whole.

This game has two pure strategy Nash equilibria: (*hawk, dove*) and (*dove, hawk*). Consider the first of these. Suppose it is common knowledge between Player 1 and Player 2 that, in interactions like this, the player in the position of Player 1 almost always chooses *hawk* and the one in the position of Player 2 almost always chooses *dove*. Expecting Player 2 to play *dove*, Player 1 forms the intention to play *hawk*. Expecting Player 1 to play *hawk*, Player 2 forms the intention to play *dove*. Given all this, does each

player have a we-intention with respect to the pair of strategies (*hawk, dove*)? On Tuomela and Miller's analysis, it seems that they do.

We say 'seems that' because Tuomela and Miller's core analysis comes with various qualifications. In particular, it applies only to 'joint social actions', defined as 'situations in which some agents act together, usually or often with the purpose of achieving some joint goal' (p. 367); this goal is 'normally (but not necessarily) the goal to perform the total action [in our notation, A]' (p. 370). Tuomela and Miller also add a condition to the effect that when Player 1 performs A_1, 'he does it in order for the participating agents to succeed in doing [A]' (p. 376). Possibly, these conditions are intended to exclude cases like the Hawk–Dove example; but if so, how these cases are excluded remains obscure.[13]

Intuitively, the Hawk–Dove case does not seem to be an instance of we-intentions. At any rate, it seems unlike the examples that are treated as paradigm cases in the literature of collective intentions: two people singing a duet, two people pushing a car, two players on the same football team executing a pass. But what makes the Hawk–Dove case different?

Searle (1990) tries to answer this question. He undertakes to show that we-intentions cannot be reduced to combinations of I-intentions—that we-intentions are 'primitive' (p. 404). He presents a critique of Tuomela and Miller's analysis and then proposes his own. The critique is persuasive at the intuitive level but, on closer inspection, turns out to be question-begging. Searle asserts that 'the notion of a we-intention … implies the notion of *cooperation*' (p. 406), and construes cooperation in terms of 'collective goals' (pp. 405, 411). He says that, in cases of collective intentionality, individual I-intentions are 'derivative from' we-intentions 'in a way we will need to explain' (p. 403). But he offers no analysis of the concepts of 'cooperation' or 'collective goal', and the explanation of the sense in which I-intentions derive from we-intentions never materializes.

Searle analyses collective intentions with reference to a case in which Jones and Smith are preparing a hollandaise sauce together, Jones stirring while Smith pours. On Searle's analysis, Jones's description of what is going on is 'We make the sauce by means of Me stirring and You pouring'. The intention in Jones's mind is: 'We intend to make the sauce by means of Me stirring' (p. 412). Searle suggests that the we-intention to make the sauce by means of Jones's stirring is like an intention to fire a gun by means of pulling the trigger. The idea seems to be that the I-intention to stir is *part*

of the we-intention to make the sauce. This is not quite the *derivation* of I-intentions from we-intentions that Searle said we needed.

Whatever one makes of this analysis,[14] it does not resolve the problem with which Searle began. We can still ask why, in the Hawk–Dove example, Player 1 and Player 2 don't have a collective intention with respect to (*hawk, dove*). What is wrong with saying that, in Player 1's mind, there is a we-intention to play the combination (*hawk, dove*) by means of Player 1 playing *hawk* and Player 2 playing *dove*? The only answer Searle's analysis can give is that, in playing that combination of strategies, Players 1 and 2 are not 'cooperating' in pursuit of a 'collective goal'. But those concepts are left unanalysed.

What is missing, we suggest, is an analysis of practical reasoning in which cooperative reasoning—or reasoning about how to achieve a collective goal—can be distinguished from straightforward individual reasoning. This is what theories of team reasoning provide.

Bratman (1993) offers a rather different account of collection intentionality in his analysis of 'shared cooperative activity'. The activities that Bratman has in mind are ones in which individuals coordinate their actions over a period of time: each has to adjust his behaviour continuously so as to keep it aligned with the behaviour of the other, as in the case of two people singing a duet. Bratman explicitly rules out the kind of 'pre-packaged cooperation' that can be represented by the choice of strategies in a normal-form game, such as the Prisoner's Dilemma or Hi-Lo (p. 339). He uses what he calls a 'planning conception of intention' (p. 330), whereby an intention is an action-guiding mental state that is maintained over an interval of time. Thus, for example, someone might have an individual intention to paint her house; this intention would then guide the formation of 'sub-plans' for buying paint, cleaning walls and so on. Bratman argues that shared cooperative activities are governed in a corresponding way by continuing intentions, but in this case the intentions are *collective*.

Collective intentions, as analysed by Bratman, reflect what he sees as the characteristic features of shared collective activity. In such an activity, a set of agents (for ease of exposition, say a pair) coordinate their actions in some joint enterprise. The successful pursuit of this enterprise requires the continuous 'meshing' of the separate sub-plans of the two agents. In Bratman's analysis, each agent has the intention that 'we' perform the

joint activity through the meshing of 'my' sub-plans with 'yours'. This intention is not linked to any *particular* combination of sub-plans; rather, it expresses a commitment to engage with the other in a process of 'mutual responsiveness' and 'mutual support' which is directed towards the meshing of sub-plans *in general*. Thus, Bratman's conception of collective intentionality can be thought of as the counterpart in the domain of intentions of group identification in the domain of practical reasons: it expresses a disposition to reason and to act as a member of a group. It leaves open one of the main questions that the theory of team reasoning tries to answer: *how* the members of a group coordinate their actions.

6. Conclusion

We have presented a number of alternative theories of team reasoning, which differ on several dimensions—in particular: how to deal with cases in which not every member of the relevant group can be relied on to identify with the group; whether group identification is a product of psychological framing or conscious commitment; whether each individual's engaging in team reasoning is conditional on assurance that others engage in it too; and, if so, whether assurance is generated by common knowledge of the psychology of framing, by joint commitment, or by experience. But despite these unresolved issues, we believe that our analysis of has shown that team reasoning is just as coherent and valid as the best-reply reasoning of conventional game theory.

Notes

⋆ Previous versions of this paper were presented at a workshop on Rationality and Commitment at the University of St Gallen, a conference on Logic, Games and Philosophy: Foundational Perspectives in Prague, the Collective Intentionality IV conference in Siena, and the Kline Workshop on Collective Rationality at the University of Missouri at Columbia. We thank the participants at these meetings, and Nicholas Bardsley, for comments. The paper uses material from our contributions as editors to an unfinished book by the late Michael Bacharach, now published as Bacharach (2006).

1. In order to conclude that (*cooperate, cooperate*) is the best pair of strategies for them, the players have to judge the payoff combinations (a, d) and (d, a) to be worse 'for them' than (b, b).
2. A pure coordination game is identical with Hi-Lo except that, for all labels i, a_i takes the same value $a > 0$.
3. In an as yet unpublished experiment, Nicholas Bardsley presented fifty-six Dutch students with two Hi-Lo games. In one game, the ratio of the money payoffs to *high* and *low* was 10:1; in the other it was 10:9. In each case, fifty-four subjects (96 per cent) chose *high*.
4. For fuller statements of this argument, see Hodgson (1967), Sugden (1993) or Bacharach (2006).
5. For example, suppose that each player believes that his opponent is just as likely to choose one strategy as the other. Then both will choose *high*. Or suppose that each player believes that his opponent believes that he is just as likely to choose one strategy as the other. Then each player will expect his opponent to choose *high*, and so choose *high* as a best reply. Or …
6. Gold (2005) shows a technical sense in which this is so, within a model of reasoning involving the manipulation of propositions.
7. This is not to deny the psychological possibility that a person might simultaneously experience motivational or affective pulls towards both individual and group identity. Our claim is merely that such conflicting pulls *cannot be resolved by instrumental reasoning*. Consider an analogous case in conventional choice theory. What if an individual faces a choice between two options, feels motivational pulls towards each of them, but cannot settle on any firm preference (or on a firm attitude of indifference)? Clearly, this case is psychologically possible; but if a person is unsure of her own objectives, instrumental rationality cannot tell her what they should be.
8. In game-theoretic language, this is a *game form*. A game form consists of a set of players, a set of alternative strategies for each player, and, for each profile of strategies that the players might choose, an outcome. In contrast, a *game* is normally defined so that, for each profile of strategies, there is a vector of numerical payoffs, one payoff for each player.
9. We will say more about how agents might come to conceive of *S* as a unit of agency in later sections of the paper.

10. Rabin's formulation of this hypothesis is not fully compatible with conventional game theory, since it allows each player's utility to depend on his beliefs about other players' beliefs about the first player's choices. (These second-order beliefs are used in defining the first player's beliefs about the second player's intentions.) Rabin's theory uses a non-standard form of game theory, *psychological game theory* (Geanakoplos, Pearce and Stachetti, 1989). However, as David Levine (1998) shows, the main features of Rabin's theory can be reconstructed within conventional game theory.
11. One might wonder why we can't simply substitute 'identifies with T' for 'identifies with S' in premise (2) of schema 12.6, and interpret U as a measure of what is good for T. But Bacharach's theory of framing commits him to these premises as we have written them. His hypothesis is that group identification is an *individual's* psychological response to the stimulus of a particular decision situation. It is not itself a group action. (To treat it as a group action would, in Bacharach's framework, lead to an infinite regress.) Thus, group identification is conceptually prior to the formation of the 'team' of people who identify with the group.
12. We can normalize the payoff function by setting $u_C = 1$ and $u_D = 0$. Then, given that $u_F < 0$, the critical value of ω is $\omega^* = 2u_F/(2u_F - 1)$. The protocol (*cooperate, cooperate*) is optimal if and only if $\omega \geq \omega^*$, (*defect, defect*) is optimal if and only if $\omega \leq \omega^*$. There is no non-zero value of ω at which (*cooperate, defect*) or (*defect, cooperate*) is optimal.
13. These conditions may rule out some dominant-strategy Nash equilibria as cases of collective intention. For example, in the case of the Prisoner's Dilemma, one might deny that Player 1 plays *defect* in order for Players 1 and 2 to succeed in playing (*defect, defect*); rather, Player 1 plays *defect* because that is best for him, irrespective of what Player 2 does.
14. Nicholas Bardsley (2005) criticizes Searle's analysis and offers an alternative, intended to be compatible with team reasoning. Bardsley's alternative to 'We intend to make the sauce by means of me stirring' as Jones's intention would be: 'I intend my part of the combination (Jones stirs, Smith pours) in circumstances that you and I have this very intention, all of which is to make the sauce'. In Bardsley's analysis, Smith's intention has exactly the same sense as Jones's (although 'my', 'you' and 'I' have different references in the two cases). The self-reference in

these intentions is analogous with the recursiveness of mutually assured team reasoning.

References

Bacharach, Michael. 1999. 'Interactive Team Reasoning: A Contribution to the Theory of Cooperation'. *Research in Economics* 53: 117–47.

Bacharach, Michael. 2000. 'Scientific Synopsis'. Unpublished manuscript (describing initial plans for *Beyond Individual Choice*).

Bacharach, Michael. 2006. *Beyond Individual Choice: Teams and Frames in Game Theory*. Natalie Gold and Robert Sugden (eds),. Princeton: Princeton University Press.

Bardsley, Nicholas. 2005. 'On Collective Intentions: Collective Action in Economics and Philosophy'. Unpublished manuscript. Paper previously presented at 'Collective Intentions IV' conference, University of Siena, October 2004.

Binmore, Ken. 1994. *Playing Fair*. Cambridge, Mass.: MIT Press.

Bolton, Gary and Axel Ockenfels. 2000. 'ERC—a Theory of Equity, Reciprocity and Competition'. *American Economic Review* 90: 166–93.

Bratman, Michael. 1993. 'Shared Intention'. *Ethics* 104: 97–113.

Cubitt, Robin and Robert Sugden. 2003. 'Common Knowledge, Salience and Convention'. *Economics and Philosophy* 19: 175–210.

Fehr, Ernst and Klaus Schmidt. 1999. 'A Theory of Fairness, Competition and Cooperation'. *Quarterly Journal of Economics* 114: 817–68.

Geanakoplos, John, David Pearce and Ennio Stacchetti. 1989. 'Psychological Games and Sequential Rationality'. *Games and Economic Behavior* 1: 60–79.

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

Gold, Natalie. 2005. 'Framing and Decision Making: A Reason-Based Approach'. Unpublished D.Phil thesis, University of Oxford.

Gold, Natalie and Christian List. 2004. 'Framing as Path-Dependence'. *Economics and Philosophy* 20: 253–77.

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

Hollis, Martin. 1998. *Trust within Reason*. Cambridge: Cambridge University Press.

Hurley, Susan. 1989. *Natural Reasons*. Oxford: Oxford University Press.

Levine, David. 1998. 'Modelling Altruism and Spitefulness in Experiments'. *Review of Economic Dynamics* 1: 593–622.

Lewis, David. 1969. *Convention: A Philosophical Study*. Cambridge, Mass.: Harvard University Press.

Rabin, Matthew. 1993. 'Incorporating Fairness into Game Theory and Economics'. *American Economic Review* 83: 1281–1302.

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

Rousseau, Jean-Jacques. 1988 [1762]. 'On Social Contract'. In Alan Ritter and Julia Conaway Bondanella (eds), *Rousseau's Political Writings*. New York: Norton.

Sally, David. 1995. 'Conversation and Cooperation in Social Dilemmas: A Meta-Analysis of Experiments from 1958 to 1992'. *Rationality and Society* 7: 58–92.

Searle, John. 1990. 'Collective Intentions and Actions'. In P. Cohen, J. Morgan, M. E. Pollack (eds), *Intentions in Communication*. Cambridge, Mass.: MIT Press, pp. 401–15.

Sen, Amartya. 1974. 'Choice, Orderings and Morality'. In R. Körner (ed.), *Practical Reason*. Oxford: Blackwell [page references are to the paper as reprinted in Amartya Sen (1982), *Choice, Welfare and Measurement* (Oxford: Blackwell)].

Sen, Amartya. 1977. 'Rational Fools: a Critique of the Behavioral Foundations of Economic Theory'. *Philosophy and Public Affairs* 6: 317–44 [page references are to the paper as reprinted in Amartya Sen (1982), *Choice, Welfare and Measurement* (Oxford: Blackwell)].

Sugden, Robert. 1993. 'Thinking as a Team: Toward an Explanation of Nonselfish Behavior'. *Social Philosophy and Policy* 10: 69–89.

Sugden, Robert. 2003. 'The Logic of Team Reasoning'. *Philosophical Explorations* 6: 165–81.

Tuomela, Raimo and Kaarlo Miller. 1988. 'We-Intentions'. *Philosophical Studies* 53: 367–89.