THE JOURNAL OF PHILOSOPHY

VOLUME CIV, NO. 3, MARCH 2007

COLLECTIVE INTENTIONS AND TEAM AGENCY*

Collective intentions are those intentions associated with joint actions. The paradigm cases include two people singing a duet, painting a house, pushing a car, and taking a walk together. In these situations, there is a sense in which "we" intend the joint action, as well as a sense in which "I" intend my part in it. For instance, if we go to Edinburgh together and, as a part of this, I check the train timetable and you buy the tickets, then there is a sense in which our actions result from "our" intention to go to Edinburgh. This collective intention is not reducible to the summation of our individual intentions and, as such, is unlike the case where two agents happen to do the same thing independently. For instance, if I plan to go to Edinburgh and coincidentally meet you on the train, there is a use of the word 'we' in which it may be that "we went to Edinburgh" because "I went to Edinburgh" and "you went to Edinburgh," but this is not a joint action and the intentions behind it are not collective intentions. Analyses of collective intentions seek to elucidate the features that are peculiar to the joint action case and to explain how collective intentions are connected to individual intentions.

The literature on collective intentions is exemplified by the work of Raimo Tuomela and Kaarlo Miller,[1] John Searle,[2] and Michael Bratman.[3] A general problem for these accounts is how to differentiate collective intentions from the mutually-consistent individual intentions that lie behind Nash equilibrium behavior in games. In game theory, the

*We would like to thank Nick Bardsley, Michael Bratman, Margaret Gilbert, Philip Pettit, and Wlodek Rabinowicz for helpful discussions.

[1] Tuomela and Miller, "We-intentions," *Philosophical Studies*, LIII (1988): 367–89.

[2] Searle, "Collective Intentions and Actions," in Philip R. Cohen, Jerry Morgan, Martha E. Pollack, eds., *Intentions in Communication* (Cambridge: MIT, 1990), pp. 401–15.

[3] Bratman, "Shared Cooperative Activity," *Philosophical Review*, CI (1992): 327–41, and "Shared Intention," *Ethics*, CIV (1993): 97–113.

0022-362X/07/0403/109–37

behavior of the players of a noncooperative game is defined to be in Nash equilibrium if each player's chosen strategy is optimal for her, given the strategies chosen by the others. Nash equilibrium is usually interpreted as a situation in which each player acts as a rational individual agent, holding true beliefs about the actions of the others. It is clear that not all Nash equilibria are joint actions. However, the core analyses provided by Tuomela and Miller, Searle, and Bratman seem to imply that all Nash equilibrium situations are instances of collective intentions. Cases in which Nash equilibria are not joint actions are excluded only by stipulation or by the addition of further conditions which are just as problematic as the original concept of collective intention. Tuomela and Miller stipulate that the definition of collective intention includes the condition that the action is joint, Searle that collective intentions involve cooperation in pursuit of collective goals. This amounts to saying that the special feature of collective intentions that distinguishes them from the intentions behind Nash equilibrium behavior is that they are associated with cooperative activity, but this is something that we already knew prior to the analysis. Bratman adds conditions which require each agent to be responsive to the behavior of the other as the joint action proceeds and if unexpected problems occur, but these conditions are stated only informally, and rely on a pre-analytic understanding of the nature of cooperative activity.[4]

We shall argue that these problems arise because, in the literature of collective intentions, the focus is on the properties of these intentions as mental states, rather than on the mental processes by which they were formed. Thus, anything that is distinctive to cooperative activity has to be represented as a distinctive feature of the corresponding intentions. An alternative approach is to analyze the practical reasoning by which individual agents choose to engage in cooperative activity and then to consider whether this reasoning has special features which lead to collective intentions.

A starting point for such an analysis can be found in a body of decision-theoretic literature on *team agency*. This seeks to extend standard game theory, where each individual asks separately "What should *I* do?" to allow teams of individuals to count as agents and for players to ask the question "What should *we* do?" This leads to *team reasoning*, a distinctive mode of reasoning that is used by members of teams, and which may result in cooperative actions. When an agent deliberates about what she ought to do, the result of her reasoning

[4] Bratman, "Shared Cooperative Activity."

is an intention. An intention is interposed between reasoning and action, so it is natural to treat the intentions that result from team reasoning as collective intentions. The concept of intention seems to imply not reasoning again, or at least that the result of reasoning is a default plan which is revised only under special circumstances. So, if the distinctive feature of collective intentions is to be found in the reasoning by which they were formed, then an analysis that focuses on the intentions themselves will miss the feature that makes collective intentions collective. In this paper, we will show how team reasoning fills the gaps in the collective intention literature and provide a framework in which various theories of group agency can be compared.

I. COLLECTIVE INTENTIONS AND NASH EQUILIBRIUM

An early analysis of collective intentionality is the work of Tuomela and Miller (*op. cit.*). The essential features of this analysis can be presented as follows for the case of a two member group, whose members are P1 and P2. Consider some "joint social action" A which comprises actions A_1 and A_2 for the respective individuals. According to Tuomela and Miller, P1 has a *we-intention* with respect to A if: (i) P1 intends to do A_1, (ii) P1 believes that P2 will do A_2, (iii) P1 believes that P2 believes that P1 will do A_1, and so on (*op. cit.*, p. 375). This analysis reduces we-intentions to individual intentions and a network of mutual beliefs.

An unsatisfactory feature of this analysis is that it seems to treat every Nash equilibrium as a case of collective intentionality. In a Nash equilibrium, each individual's action is a best response to her true beliefs about the other's action. Since these are intentional actions, this is equivalent to saying that each individual's intention is adapted to her true beliefs about the actions of the other. For example, consider the version of the Hawk-Dove game shown in Figure 1 (following).

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 P1 and P2 that, in interactions like this, the player in the position of P1 almost always chooses *hawk* and the one in the position of P2 almost always chooses *dove*. Expecting P2 to play *dove*, P1 forms the intention to play *hawk*. Expecting P1 to play *hawk*, P2

		P2	
		dove	*hawk*
P1	*dove*	2, 2	0, 3
	hawk	3, 0	−5, −5

Figure 1: Hawk-Dove

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.

The problem becomes even more acute when we consider the Prisoner's Dilemma, a two-person version of which is shown in Figure 2.

As an example of such a game, think of two individuals who must each decide how much of a valuable resource to take. This resource renews itself each season but the amount available next season will depend on how much is left at the end of this one (think of fish in the sea). Payoffs reflect the amount of the resource that each player gets over two seasons. To play *cooperate* is to take a moderate amount of the resource this season; to play *defect* is to take a large amount. The rate of depletion is such that, if only one player takes a large amount then the increased amount she gets this season outweighs the decrease next season, but if one player takes a large amount this season then the other player does better by also taking a large amount now before the resource becomes depleted. For each player, whatever the other player does, her best move is to play *defect*. However, the outcome in which both players *defect* and the resource is severely depleted is worse for each player than that in which they both *cooperate* and conserve enough of the resource to renew itself for next season.

The unique Nash equilibrium of this game is (*defect*, *defect*). So, in the Nash equilibrium, P1 will form the intention to play *defect*. Because playing *defect* is a dominant strategy, if there is common knowledge of rationality, P1 will expect P2 to do likewise (though it is in her best interests to form the intention to *defect* regardless of her expectation about P2's action). So, again, the Nash equilibrium behavior seems to fall under Tuomela and Miller's analysis of collective intentions. But, intuitively, the exploitation of the resource is not a joint action. It is not true to say that each player intends that "we" exploit the common resource. Conversely, the non-Nash equilibrium cooperative practice of conserving the resource does involve a collective intention.

Although Tuomela and Miller's core analysis seems to include the intentions behind the Nash equilibria in these games, it comes with various qualifications. In particular, it applies only to "joint social

		P2	
		cooperate	*defect*
P1	*cooperate*	**3, 3**	**1, 4**
	defect	**4, 1**	**2, 2**

Figure 2: The Prisoner's Dilemma

actions," defined as "situations in which some agents act together, usually or often with the purpose of achieving some joint goal" (*op. cit.*, p. 367); this goal is "normally (but not necessarily) the goal to perform the total action [in our notation, A]" (*op. cit.*, p. 370). Tuomela and Miller also add a condition to the effect that when P1 performs A_1, "he does it in order for the participating agents to succeed in doing [A]" (*op. cit.*, p. 376). 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 P1 plays *defect* in order for P1 and P2 to succeed in playing (*defect, defect*); rather, P1 plays *defect* because that is best for him, irrespective of what P2 does. Possibly, these conditions are also intended to exclude cases like the Hawk-Dove example; but if so, how these cases are excluded remains obscure.

Searle also criticizes Tuomela and Miller's analysis for including cases where there is no collective intentionality (*op. cit.*). He proposes that we-intentions cannot be reduced to such combinations of I-intentions and beliefs—that we-intentions are "primitive" because of their distinct phenomenology (*op. cit.*, p. 404)—and undertakes an analysis of his own. The critique is persuasive at the intuitive level but, on closer inspection, turns out to be question-begging. Searle says that, in cases of collective intentionality, individual I-intentions are "derivative from" we-intentions "in a way we will need to explain" (*op. cit.*, p. 403). He then analyzes collective intentions with reference to a case in which Jones and Smith are preparing a hollandaise sauce together, Jones stirring while Smith pours. 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 distinctive problem of collective intentions is how to characterize each individual's intention without having *either* Jones's individual intention causing the whole joint action *or* having Jones intend Smith's action, and vice versa. 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," and the intention in Jones's mind is: "We intend to make the sauce by means of Me stirring" (*op. cit.*, p. 412). 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, it does not resolve the problem with which Searle began. We can still ask why, in the Hawk-Dove example, P1 and P2 do not have a collective intention with respect to (*hawk*, *dove*). What is wrong with saying that, in P1's mind, there is a we-intention to play the combination (*hawk*, *dove*) by means of P1 playing *hawk* and P2 playing *dove*? Searle asserts that "the notion of a we-intention...implies the notion of *cooperation*" (*op. cit.*, p. 406) and says that cooperation is construed in terms of "collective goals" (*op. cit.*, pp. 405, 411), so he might answer that, in playing the strategy combination (*hawk*, *dove*), P1 and P2 are not "cooperating" in pursuit of a "collective goal." But those concepts are left unanalyzed.

So Searle's analysis, like Tuomela and Miller's, distinguishes between genuinely collective intentions and the individual intentions of Nash equilibrium only by appealing to an intuitive understanding of the concepts of "cooperation" (or "joint action") and "collective goal" or ("joint goal"). From the perspective of the literature of collective intentionality, the problem is to find a way of making the cooperative pursuit of a collective goal a property of the corresponding intention. The difficulties that these writers have found in grappling with this problem suggest that a different line of approach might be more useful. Our approach will be to analyze the nature of practical reasoning associated with the pursuit of a collective goal, and then to investigate the intentions that it produces.

Bratman offers a rather different account of collective intentionality in his analysis of "shared cooperative activity." For Bratman, the key feature of shared cooperative activity is that, for each agent, I must intend that we J "in part *because of* your intention that we J."[5] This must be known by both agents. (In "Shared Intention," there is an explicit common knowledge condition. This much gets us the core analysis of View 3 in that paper.) It is not clear that Bratman's basic analysis, any more than Tuomela and Miller's or Searle's, excludes Nash equilibrium play in Hawk-Dove. In Nash equilibrium, each individual's action is adapted to her beliefs about the actions of the other and, since these are intentional actions, that is equivalent to saying that each individual's intention is adapted to her beliefs about the intentions of the other. If P1 intends the Nash equilibrium outcome

[5] "Shared Cooperative Activity," p. 333 (with italics); "Shared Intention," p. 104 (without italics).

(*hawk, dove*), and believes that this outcome will come about in part because P2 will play dove, and that P2 will play dove because he intends to, and that P2 intends to play dove because he intends the outcome (*hawk, dove*) and because he believes that (*hawk, dove*) will come about in part because P1 will play *hawk*, and so on, then it seems that playing (*hawk, dove*) will be a shared cooperative activity in Bratman's sense. Bratman might object that it is too glib to interpret P1's intending that *J* come about "because of" P2's intention as the idea that P1 believes that P2 has the corresponding intention and acts on the basis of this belief. But his expansion of it—"I intend that our performance of the joint activity be in part explained by your intention that we perform the joint activity; I intend that you participate as an intentional agent in a joint activity that, as I know, you too intend"[6]—is opaque. The subsequent discussion reveals that P1 must intend the "efficacy" of P2's intention, or that the outcome is achieved through P2's intention to achieve it. It is not clear how this would exclude there being a collective intention with respect to the pair of strategies (*hawk, dove*).

An alternative response is that the Hawk-Dove case is excluded from the scope of Bratman's analysis because the two players move once only, and simultaneously. Bratman uses what he calls a "planning conception of intention,"[7] whereby an intention is an action-guiding mental state that is maintained over an interval of time; the cooperative activities that he has in mind are ones in which individuals coordinate their actions over 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. Analogously, Bratman argues that "we" might have a collective intention to paint the house which guides the formation of "my" and "your" sub-plans.[8] In game-theoretic terms, this transforms the problem into one of dynamic choice. We might represent the shared cooperative activity as a possible outcome of an extended game comprising a sequence of "stage games" (for instance, a series of Hawk-Dove games), with a stage-game strategy counting as a "subplan." Bratman says that the subplans of players A and B for a joint action *J mesh* if there is some way that "we could *J* that would...involve the successful execution of those subplans."[9] So the

[6] "Shared Intention," p. 104.
[7] "Shared Cooperative Activity," pp. 330–31.
[8] Adding these subplans gets the schema on pp. 333–34 of "Shared Cooperative Activity," and View 4 in "Shared Intention."
[9] "Shared Cooperative Activity," p. 332; "Shared Intention," p. 106.

intentions behind a sequence of Nash equilibria in the stage games (for instance, "we intend to perform the sequence in which, in every stage game, A plays *hawk*, B plays *dove*") would be a collective intention.

However, it is clear that Bratman would not be happy with this characterization of his views. He says that his view implies more than that each player "sees the other's intentions as data for [her] deliberations, albeit as data that are potentially affected by [her] own decision"—in other words, it requires more than Nash equilibrium. It requires that each agent "aims at the efficacy of the intention of the other" and "embrace[s] as her own end the efficacy of the other's relevant intention."[10] How he would discriminate between shared intentions and Nash equilibrium behavior is clearer in his "Shared Cooperative Activity" paper, which includes further conditions, particularly that each agent's intention is "minimally cooperatively stable" (*op. cit.*, p. 338). This requires that there are at least some circumstances in which she would help the other agent if a new problem arose. Bratman also specifies that the agents' attitudes have to be appropriately connected to the joint action which, he says, explicitly distinguishes shared cooperative activity from the kind of "pre-packaged cooperation" that can be represented by the choice of strategies in a game. If our roles in an activity are fully determined in advance and we can each carry out our part with no further interaction then, whilst the planning stage may have been a shared cooperative activity, the noninteractive performance of the plan itself is pre-packaged cooperation. Although both shared cooperative activity and pre-packaged cooperation involve our bringing about that we *J* through "mutual responsiveness of intention," shared cooperative activity involves the additional feature of "mutual responsiveness in action." To illustrate mutual responsiveness in action, Bratman gives the example of singing a duet, each singer accommodating herself to the actions of the other. The intuition is clear enough, but there is no analysis of how mutual responsiveness works. So the status of mutual responsiveness is similar to that of "collective intention" and "joint action" prior to Bratman's work—something we have an intuitive understanding of, but want an analysis of.

We are also left wondering how Bratman's project relates to other analyses of collective intentions. He differentiates it from the analyses provided by Tuomela and Miller and Searle by saying that his shared intentions are states of affairs made up of the interrelated intentions and beliefs of the people who share them; they are not intentions of a

[10] "Shared Intention," p. 107.

special kind, held by individual agents.[11] However, we see Bratman's use of the planning conception of intention as a much more fundamental distinction, in that it makes collective intentions prior to practical reasoning about cooperative activity. We will argue that his theory may be best understood as one which analyzes group agency in terms of the mental states of the individual agents that compose the group, and which provides the background circumstances in which individual agents use team reasoning.

II. TEAM REASONING

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 groups or "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,[12] Donald Regan,[13] Margaret Gilbert,[14] Susan Hurley,[15] Robert Sugden,[16] Martin Hollis,[17] Michael Bacharach,[18] and Elizabeth Anderson.[19] 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. These are puzzles of *cooperation*, in that the strategies offer what Bacharach calls *scope for common gain*, or the possibility of making a Pareto improvement (that is, at least one player is made

[11] "Shared Intention," pp. 99, 103.

[12] Hodgson, *Consequences of Utilitarianism* (New York: Oxford, 1967).

[13] Regan, *Utilitarianism and Cooperation* (New York: Oxford, 1980).

[14] Gilbert, *On Social Facts* (New York: Routledge, 1989).

[15] Hurley, *Natural Reasons* (New York: Oxford, 1989).

[16] Sugden, "Thinking as a Team: Toward an Explanation of Nonselfish Behavior," *Social Philosophy and Policy*, X (1993): 69–89; "The Logic of Team Reasoning," *Philosophical Explorations*, VI (2003): 165–81.

[17] Hollis, *Trust within Reason* (New York: Cambridge, 1998).

[18] Bacharach, "Interactive Team Reasoning: A Contribution to the Theory of Cooperation," *Research in Economics*, LIII (1999): 117–47; *Beyond Individual Choice*, Natalie Gold and Robert Sugden, eds., (Princeton: University Press, 2006).

[19] Anderson, "Unstrapping the Straitjacket of 'Preference': A Comment on Amartya Sen's Contributions to Philosophy and Economics," *Economics and Philosophy*, XVII (2001): 21–38.

better off and no player is made worse off) on an outcome predicted by a standard solution concept (*op. cit.*).

One such puzzle is the Prisoner's Dilemma. Conventional game theory predicts that players will always choose *defect*, while in fact many players choose *cooperate*. There is scope for common gain because if both players choose *cooperate* then they are both better off than if they both choose *defect*. 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 percent.[20] The theory is failing to explain observed behavior 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 behavior 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*)."[21] It seems that normative argument between these two positions leads to a stand-off.

A second puzzle is the game of Hi-Lo. This is a game in which each player chooses one element from the same set of *labels*. As in a pure coordination game, the two players get the same strictly positive payoff if both choose the same label and (0, 0) otherwise. However (and in contrast to a pure coordination game, in which all labels give the same payoff), there is one label that gives a strictly higher payoff from coordination for each player than the others. Figure 3 shows a simple version of Hi-Lo, in which there are just two labels, *high* and *low*.

Hi-Lo combines features of pure coordination games 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

[20] David Sally, "Conversation and Cooperation in Social Dilemmas: A Meta-Analysis of Experiments from 1958 to 1992," *Rationality and Society*, VII (1995): 58–92.

[21] In order to conclude that (*cooperate*, *cooperate*) is the best pair of strategies for them, the players have to judge the payoff combinations (4, 1) and (1, 4) to be worse "for them" than (3, 3).

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

Figure 3: Hi-Lo

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: there is scope for common gain because 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*.[22] 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

[22] Nicholas Bardsley, Judith Mehta, Chris Starmer, and Sugden, "The Nature of Salience Revisited: Cognitive Hierarchy Theory *versus* Team Reasoning" (unpublished manuscript, 2006), report an experiment where 56 student subjects were presented 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, 54 subjects (96 percent) chose *high*.

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 favor 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.[23]

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*.[24] But it seems unsatisfactory to have to invoke assumptions about imperfections of rationality in order to explain behavior 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-

[23] For fuller statements of this argument, see Hodgson, *Consequences of Utilitarianism*; Sugden, "Thinking as a Team: Toward an Explanation of Nonselfish Behavior"; or Bacharach, *Beyond Individual Choice*.

[24] 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....

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. In the same way that individual reasoning leads to individual intentions, it is natural to suppose that team reasoning leads to collective intentions. In this section of the paper we show how the we-intentions of individual agents would arise as the result of team reasoning.

We 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.[25] 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. 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. One can formulate principles of practical reasoning which satisfy analogous criteria of validity. Following Bacharach, we define a mode of practical reasoning for a given class of games as *valid* if it is *success-promoting*: given any game in that class, it yields only choices which tend to produce success, as measured by game payoffs.[26] 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

[25] Gold and List, "Framing as Path-Dependence," *Economics and Philosophy*, xx (2004): 253–77.

[26] Bacharach, *Beyond Individual Choice*, pp. 7–10.

individual person, the reasoning is *individually instrumental*. Here is a simple example of individually instrumental reasoning, in which (1) to (4) are premises and the proposition below the line is the conclusion:

Schema 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 "official" payoffs of a game (as represented in the matrix which defines the game, Figures 1, 2 and 3 being examples) 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.

Now consider the following schema, in which (*left*, *right*) denotes the pair of actions "I choose *left*, you choose *right*":

Schema 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. Both present instrumental practical reasoning, where an agent deliberates about what to do in order to achieve its goals. The only difference is the level of agency. If Schema 1 represents valid reasoning for an individual agent, Schema 2 is its parallel for a group agent.

As far as we can see, the only grounds for objecting to Schema 2 while accepting Schema 1 is to claim that the concept of group agency itself is incoherent. This claim can construed in two different ways.

On the first construal, the claim is that expressions such as 'we want to achieve' and 'we should choose' are mysterious, because wanting and choosing are mental states and mental states must be located in individual minds. In the rest of this section of the paper, we try to answer this objection by showing how the "we" concepts of Schema 2 correspond with "I" concepts. We assert that Schemata 1 and 2 are both forms of valid instrumental reasoning and that they are valid *for agents*, defined as those entities that use these modes of reasoning (individual agents in one case, group agents in the other).

On the second construal, it is a claim about social ontology. The counter-claim that groups *can* be agents in this metaphysical sense is defended by Carol Rovane[27] and Philip Pettit.[28] For both Rovane and Pettit, an agent is characterized by a commitment to "rational unity"; it can be any entity that has the ability to form states that play the role of intentional attitudes, such as judgments and beliefs, and that can take steps to ensure that these states are consistent with each other. If the defining feature of agency is the ability to engage in rational deliberation, then the bounds of agency need not coincide with the individual. Rovane and Pettit's is a normative concept of agency; it presupposes principles of rationality which are constitutive of being an agent, and derives the metaphysical possibility of group agency from these principles.

In this paper, we do not make any ontological or metaphysical claims; rather, we describe a type of reasoning which the members of group agents use. This means that we are using a more inclusive definition of group agent than Rovane and Pettit. On their definition some of the examples of ephemeral groups in the collective intentions literature, such as two strangers pushing a car, will not exhibit agency. But these groups may use team reasoning. In section III, we will present a range of alternative accounts of what group agency is and how it comes about; some of these accounts are compatible with Rovane and Pettit's definition, while others are not. In this section, we confine ourselves to showing that there are practical reasoning schemata that use "we" concepts, which individual people can (and, we claim, sometimes do) use, which are similar in structure to modes of individual practical reasoning.

We begin by noting what might seem to be a potential contradiction between Schemata 1 and 2. For example, consider the

[27] Rovane, *The Bounds of Agency* (Princeton: University Press, 1988).
[28] Pettit, *Weakness of the Will and Practical Irrationality* (New York: Oxford, 2003).

Prisoner's Dilemma. Using a variant of Schema 1, each player can apparently reason to "I should choose *defect*"; but using a variant of Schema 2, each can apparently reason to "We should not choose (*defect*, *defect*)." We suggest that the resolution of this problem is that, properly understood, the two sets of premises are mutually inconsistent.[29] Thus, while both schemata are valid, they cannot be used simultaneously to infer contradictory conclusions. The premises of Schema 1 presuppose that *I* am an agent, pursuing *my* objectives. Those of Schema 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.[30]

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. 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.[31] 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.[32] Now consider any individual *i*, and any

[29] Gold shows a technical sense in which this is so, within a model of reasoning involving the manipulation of propositions in *Framing and Decision Making: A Reason-Based Approach* (Unpublished D.Phil thesis, University of Oxford, 2005).

[30] 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.

[31] 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.

[32] Thus, the "official" payoffs of a game are the values of payoff functions which represent what each player would want to achieve as an individual agent.

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. 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 following schema:

Schema 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 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 3 encompasses both individual and team reasoning.

Schema 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 3 can be rewritten as follows:

Schema 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 maximizes U.

I should choose my component of A.

Schema 4 represents team reasoning as a mode of reasoning that can be used by an individual group member. If it is common knowledge that every member of the group identifies with the group and wants the group's payoff function to be maximized, and if it is also common knowledge that a particular profile of actions A uniquely maximizes that function, then premises (1) to (4) are accessible to each member. Each can then reason independently to the conclusion that she should choose her component of A.

Team reasoning was originally introduced to explain how, when individuals are pursuing collective goals, it can be rational to choose strategies that realize scope for common gain. But it also provides an account of the formation of collective intentions. Team reasoning, as represented by Schema 4, results in the formation of intentions. An individual who accepts premises (1) to (4) and infers "I should choose my component of A" has reason to form the intention to choose that component. That intention, if formed, is a mental state of an individual person. Nevertheless, references to the group are noneliminable parts of the reasoning process that led to the formation of the intention. Thus, it is natural to regard the intentions that result from team reasoning as collective intentions.

Interpreted in this way, our analysis does everything Searle asked for. The "we-ness" of we-intentions is primitive, in the sense that group agency is a noneliminable part of the reasoning process by which these intentions are formed. The presence of group agency also explains the distinctive phenomenology of we-intentions. Collective goals (in the form of the group payoff function U) play a fundamental role in the analysis. I-intentions are derived from we-intentions in the move from Schema 2, through Schema 3, to Schema 4. The resulting in-

dividual intention dovetails with an alternative to Searle's analysis proposed by Nicholas Bardsley,[33] which is intended to be compatible with team reasoning.[34]

Our analysis allows us to distinguish collective intentions from the individual intentions that exist in Nash equilibrium, and thus to solve the general problem we identified in the literature of collective intentions. This distinction can be made by referring to the unit of agency in the reasoning process that led to the formation of the relevant intentions. Nash equilibrium, as normally understood, is a relationship between the strategy choices of players who reason as individuals (or, equivalently, each of whom identifies with the one-person group which contains only himself). Correspondingly, the intentions which underlie Nash equilibrium are the result of individual reasoning. In contrast, collective intentions are the product of team reasoning.

As an illustration, we return to the Hawk-Dove example discussed in section I. In this example, it is common knowledge between P1 and P2 that, in Hawk-Dove interactions, the player in the position of P1 almost always chooses *hawk* and the one in the position of P2 almost always chooses *dove*. Expecting P2 to play *dove*, P1 forms the intention to play *hawk*. Expecting P1 to play *hawk*, P2 forms the intention to play *dove*. We posed the question: Does each player have a we-intention with respect to (*hawk*, *dove*)? To answer this question, we investigate the reasoning process by which the intentions were formed. Given his belief that P2 will almost certainly choose *dove*, P1 reasons as follows:

(1) I must choose either *hawk* or *dove*.
(2) If I choose *hawk*, the outcome will (almost certainly) be a payoff of 3.
(3) If I choose *dove*, the outcome will (almost certainly) be a payoff of 2.
(4) I want to achieve a payoff of 3 more than I want to achieve a payoff of 2.

I should choose *hawk*.

Reasoning similarly, P2 reaches the conclusion "I should choose *dove*." Both chains of reasoning follow the "individual rationality"

[33] Bardsley, "On Collective Intentions: Collective Action in Economics and Philosophy," *Synthese* (forthcoming, 2007).

[34] 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 'you' and 'I' refer to a different people in the two cases).

model of Schema 1. We conclude that P1's and P2's intentions with respect to (*hawk*, *dove*) are not collective.

It is an implication of our analysis that a given pattern of behavior can be intended either individually or collectively, depending on the reasoning which led to it. For example, compare the following two stories about Hi-Lo. In the first story, P1 and P2 reason as individuals. Each forms the belief that the other is just as likely to play one strategy as the other. Since this implies that the expected individual payoff from choosing *high* is greater than that from choosing *low*, each forms the intention to choose *high*. In this case, (*high*, *high*) is a Nash equilibrium in the standard sense, and the players' intentions are not collective. In the second story, each of P1 and P2 identifies with the group {P1, P2} and takes the group payoff function to be one which is uniquely maximized by the profile (*high*, *high*); all of this is common knowledge between them. Each forms the intention to play *high* as his component in this profile. These intentions are collective.[35]

III. THEORIES OF GROUP AGENCY

The pure theory of team reasoning, as presented in the previous section of this paper, presupposes that there can be group agency; but it is not reliant on any particular theory of how group agency comes about or of what the group agent should take as its goals. The literature of team reasoning offers a range of possible answers to these questions. In this section, we briefly review various theories of group agency. Our aim is not to adjudicate between them, but merely to indicate the variety of philosophically significant ways in which group agency can be interpreted. In the process, we locate Bratman's analysis within our framework.

III.1. Team Agency Required by Rationality or Morality. The first theorists to discuss team reasoning did so in the context of moral and rational requirements on action. In these accounts, what we have called "identifying with" a group is construed as an individual's response to such requirements.

Hodgson was the first person to use the Hi-Lo game, as part of an argument that rule utilitarianism does not reduce to act utilitarianism (*op. cit.*). Expanding on this argument, Regan 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

[35] Because conventional game theory does not recognize group agency, we cannot appeal to convention to resolve the question of whether (*high*, *high*) is properly called a "Nash equilibrium" in this case. It is a Nash equilibrium in relation to the official payoffs of the game, but the players are not motivated by these payoffs in the standard way.

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" (*op. cit.*, p. 124). In a world in which everyone is a cooperative utilitarian, Regan's rational and moral agents reason according to Schema 4; S is the set of all agents, and the value of U is a measure of the overall good of the world.

Regan's theory also 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.[36] As in Schema 4, S denotes the group with which team reasoners identify. However, it is allowed that not all members of S in fact identify with S. The set of those members of S (the "team" for the purposes of team reasoning) who do so identify is denoted by T. Let A_T be a profile of actions for the members of T. Then restricted team reasoning is represented by the following schema:

Schema 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 maximizes U, given the actions of nonmembers 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 the world. In terms of Schema 5, S is the set of all people, T is the set of cooperative utilitarians, and the value of U is a measure of overall goodness.

The idea that team reasoning can be required by morality is not limited to utilitarians. Anderson presents a Kantian argument for the rationality of team reasoning and for its morality when applied to the universal community of humanity (*op. cit.*). Anderson presents this idea while explaining how, in games which offer scope for common gain, individuals can rationally treat their joint strategy profile as the unit of selection. Her analysis of the "rational basis for committed ac-

[36] Bacharach, *Beyond Individual Choice.*

tion" in such games follows the general logic of Schema 4. The analysis applies to cases in which a group of individuals identify themselves as members of that group, and see their actions as jointly advancing a common goal. Under these conditions, it is rational for each member of the group to choose her component of the strategy profile that best advances the goal. This conclusion is reached by "collective deliberation," structured by the principle that "whatever can count as a reason for action for one member of the collective must count as a reason for all"; it is rational by virtue of the individual's having reason "to do my part in what we are willing together" (*op. cit.*, pp. 28–30).

Anderson does not claim that practical reason requires individuals to identify with any particular groups. Rather, identity is prior to rational choice, in the sense that "what principle of choice it is rational to act on depends on a prior determination of personal identity, of who one is" (*op. cit.*, p. 30). However, she holds out the hope that there may be "further principles of rational self-identification" which will tell us which groups we should identify with (*op. cit.*, p. 32). She maintains that morality requires us to transcend our various identities and harmonize their demands, by identifying with a community that comprehends them all—the Kantian Kingdom of Ends—but she admits to having no argument that this is rationally required (*op. cit.*, p. 37). Similarly Kantian aspirations can be found in Hollis's sketch of how team reasoning might be used by "citizens of the world" (*op. cit.*, pp. 150–63).

In a similar vein, Hurley 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 (*op. cit.*, pp. 136–59). 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.*" Like Anderson, Hurley does not nominate any particular group as the rational one to identify with, or any particular goal as the rational one to pursue. Nevertheless, 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.

III.2. Team Agency as the Result of Framing. In contrast, Bacharach's 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.[37] For

[37] Bacharach, *Beyond Individual Choice.*

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 induced or *primed* by games which 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.[38] 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. In particular, although the Prisoner's Dilemma exhibits strong interdependence, Bacharach allows that a player of this game may frame it either as a single-agent problem for "us" or as a game to be played by two separate individual agents.

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. 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 identities 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 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.

[38] For a more formal definition, see Bacharach, *Beyond Individual Choice.*

(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 nonmembers of T.

I should choose my component of P.

Bacharach applies this theory to the Prisoner's Dilemma, setting S = {P1, P2} and interpreting U as a measure of the value of the outcome of the game to P1 and P2 together; he assumes that U is maximized when both players choose *cooperate*. If the value of U is higher when just one player chooses *cooperate* than when neither does, or if the value of ω is sufficiently high, the uniquely optimal protocol is (*cooperate, cooperate*). This gives 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.

Bacharach claims that Schema 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.

III.3. Team Agency and Assurance. Up to now, we have treated schemata of practical reasoning as modes of *valid* reasoning, 'validity' being interpreted instrumentally. Thus, team reasoning has been presented as an element of the theory of *rational* choice. Sugden offers an alternative approach, presenting a "logic of team reasoning" without making any claims for its validity.[39] In this analysis, a "logic" is merely an internally consistent system of axioms and inference rules.

[39] Sugden, "The Logic of Team Reasoning."

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.

Sugden is particularly concerned with the following *assurance* question: When a team reasoner chooses her component of the profile that maximizes her team's payoff, does she have the assurance that other members of the team are choosing their components too? If one asserts (as Sugden does not) that the relevant schema of team reasoning is valid, one can give the following straightforward answer: if the team members know one another to be rational, this knowledge provides the necessary assurance. To see why, consider the simplest schema of team reasoning, Schema 4. Notice that this schema tells an individual member of S to choose his component of the U-maximizing profile only in situations in which it also tells the other members to choose theirs. Thus, if each member is rational in the sense of being capable of valid practical reasoning and having the motivation to act on the conclusions it generates, each will act on the conclusions of Schema 4, as applied to his case. And, since it is common knowledge in S that everyone identifies with S, each player has the resources to work all this out. So, whenever Schema 4 tells an individual to choose his component of the U-maximizing profile, that individual has the assurance that the others, *if rational*, will choose theirs too.[40] But since Sugden's approach does not acknowledge agent-neutral concepts of "validity" and "rationality," assurance has to be generated in a different way.

Following David Lewis[41] and Robin Cubitt and Sugden,[42] 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 than she accepts as true, using rules of inference that she accepts as valid. On the analogue of

[40] This argument extends to the cases of restricted and circumspect team reasoning, in the sense that each team-reasoning individual has the assurance that other members of T, the "team" of like-minded individuals, will choose their components of the profile or protocol which maximizes the value of the team's payoff function. However, a member of T may know with certainty (in the case of restricted team reasoning) or with high probability (in the case of circumspect team reasoning) that other members of S will act as individual agents—for example, by choosing *defect* in the Prisoner's Dilemma.

[41] Lewis, *Convention: A Philosophical Study* (Cambridge: Harvard, 1969).

[42] Cubitt and Sugden, "Common Knowledge, Salience, and Convention," *Economics and Philosophy*, XIX (2003): 175–210.

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.[43] 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 (say) P1 in deciding how to play the Prisoner's Dilemma, we cannot allow the premise that, in the group {P1, P2}, there is common reason to believe that P1 chooses *cooperate*. That would make it a premise that P1 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 {P1, P2} choose *cooperate*, and there may be circumstances in which such a premise would be natural. For example, suppose that P1 and P2 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, P1 might have reason to believe that P2 will choose *cooperate*. Attributing similar reasoning to his opponent, P1 might have reason to believe that P2 has reason to believe that P1 will choose *cooperate*; and so on.

Sugden's formulation of team reasoning can be represented as the following schema:

Schema 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 every member of S identifies with S and acknowledges U as the objective of S.

[43] The concept we define here is similar to that of "reciprocal" belief, used by Frederic Schick, "Surprise, Self-Knowledge, and Commonality," this JOURNAL, XCVII, 8 (August 2000): 440–53.

(4) In S, there is reciprocal reason to believe that every 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 presented merely as 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.

This schema 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 a kind we will discuss in the following subsection. But it could also be induced by repeated experience of regularities of behavior 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.

III.4. Team Agency Produced by Commitment. A final 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 analysis of "plural subjects" (*op. cit.*). 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 (*op. cit.*). Hollis suggests that Rousseau's account of the social contract,[44] 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.

Although Gilbert does not offer an explicit model of collective choice, we suggest that Schema 4 is compatible with her general approach, provided that "identifying with" the group S is understood as some kind of conscious act of commitment. On this interpretation, the schema is asserted to be rational, but not in an instrumental sense. 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.

We suggest that Bratman's analysis of shared intention can be understood as an account of group agency produced by commitment, even though he does not elaborate on how that commitment comes about. Given Bratman's "planning concept of intention," we can identify two levels of intentions: agents form high-level, *strategic* intentions, which guide the practical reasoning that leads to low-level, *tactical* intentions. Strategic intentions set the framework within which subsequent tactical reasoning takes place. Bratman's analysis is at the strategic level. Recall that, in shared cooperative activity, each agent has the intention that "we" perform some joint activity through the meshing of "my" subplans with "yours." This intention is not linked to any *particular* combination of subplans; 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*. In this sense, the shared intentions are "interlocking" and "end-providing."[45]

[44] Rousseau, "The Social Contract" (1762) in Alan Ritter and Julia Conaway Bondanella, eds., *Rousseau's Political Writings* (New York: Norton, 1988).

[45] "Shared Cooperative Activity" p. 335.

In later work, Bratman associates shared intentions with "shared valuing," a shared policy that says what considerations are to be given justifying significance and which provides a background framework for shared deliberation. This would be the analogue of the team payoff function.[46] It seems that, for Bratman, collective intentionality expresses a disposition to reason and act as a member of a group in relation to the objective of executing some broadly-defined joint activity. His analysis leaves open the question of *how*, at the tactical stage, the members of a group coordinate their actions so that together they achieve their joint objective.

Since this is the central question addressed by the theory of team reasoning, the two approaches can be seen as complementary. Bratman's planning conception of shared intention can be thought of as the counterpart in the domain of intentions of group identification in the domain of practical reasons. Seen in this light, Bratman is providing a reductionist account of group agency itself, in terms of the mental states of the individual members of the group.

IV. CONCLUSION

We have identified a general problem in the literature of collective intentions, that of how to distinguish between the "we-intentions" that lie behind cooperative actions and the mutually consistent "I-intentions" that lie behind Nash equilibrium behavior. We have argued that the problem arises because, for this purpose at least, the analyses in the collective intention literature have the wrong starting point. The key difference between the two kinds of intention is not a property of the intentions themselves, but of the modes of reasoning by which they are formed. Thus, an analysis which starts with the intention has already missed what is distinctively collective about it. In our analysis, collective intentions are the product of a distinctive mode of practical reasoning, team reasoning, in which agency is attributed to groups. We have presented the core features of team reasoning in the form of explicit schemata, discussed a range of possible accounts of group agency, and shown how existing theories of collective intentions fit into this framework.

NATALIE GOLD

University of Edinburgh

ROBERT SUGDEN

University of East Anglia

[46] Bratman, "Shared Valuing and Frameworks for Practical Reasoning," in R. Jay Wallace, Philip Pettit, Samuel Scheffler, and Michael Smith eds., *Reason and Value: Themes from the Moral Philosophy of Joseph Raz* (New York: Oxford, 2004), pp. 1–27.