Economics and Philosophy, **26** (2010) 291–320 Copyright © Cambridge University Press
doi:10.1017/S0266267110000386

TWO KINDS OF WE-REASONING

Raul Hakli, Kaarlo Miller and Raimo Tuomela
University of Helsinki

People sometimes think in terms of 'we' referring to a group they belong to. When making decisions, they frame the decision problem as: 'What should we do?' instead of 'What should I do?'. We study one particular approach to such 'we-reasoning', economist Michael Bacharach's theory of 'team reasoning', and relate it to philosopher Raimo Tuomela's distinction between 'I-mode' reasoning and 'we-mode' reasoning. We argue that these theories complement each other: Tuomela's philosophical theory provides a conceptual framework augmenting Bacharach's theory, and Bacharach's mathematical results support Tuomela's view on the irreducibility of the we-mode to the I-mode. We-mode reasoning can explain some kinds of human cooperative behaviour left unexplained by standard game theory. Standard game theory is not well-equipped to deal with we-mode reasoning but it can be extended by the methods developed by Bacharach. However, we argue that both standard game theory and Bacharach's theory require more attention to the information-sharing stages that precede actual decision making, and we describe a stage-based model of we-reasoning.

1. INTRODUCTION

People frequently find themselves in situations in which they have to make joint decisions or decisions that involve several people. Such situations have been studied in game theory and economics on the one hand, and in philosophical theories of collective action on the other. In this paper we aim to discuss some connections between these two research fields.

Several people have read and commented on earlier versions of the manuscript, and we thank them for their comments: Sabine Englschalk, Lorena Jaume-Palasí, Aki Lehtinen, Pekka Mäkelä, Olli Niinivaara, Martin Rechenauer and two anonymous referees. This research was supported by the Academy of Finland.

https://doi.org/10.1017/S0266267110000386 Published online by Cambridge University Press

Non-cooperative game theory is the most developed theory of rational decision making in situations involving several agents. However, there are seemingly simple situations, describable as two-person normal form games, which are problematic for the theory. Two well-known games of that sort, Prisoner's Dilemma (PD) and Hi-Lo, have recently received much attention, because the predictions that game theory makes seem to clash with either empirical results or intuitive rationality judgements (see e.g. Sally 1995; Colman *et al.* 2008a). Two attempts to overcome these difficulties have been suggested, one based on what has been called *preference transformation* and another based on what has been called *agency transformation*.

The first approach starts from the observation that the experimental settings that attempt to represent a certain game, e.g. a PD, in fact do not necessarily have PD structure. In experimental settings, the payoffs typically represent some objective features of the situation like monetary rewards. These *objective payoffs* are not necessarily equal to the *subjective utilities* that correspond to the agents' preferences which are assumed to represent all the factors that the agents take to be relevant in the decision-making situation. In particular, preferences can be social or other-regarding: They may be affected by such factors as other agents' preferences, or social norms, e.g. norms concerning fairness (see e.g. Bowles and Gintis 2006). In order to take such factors into account, various transformation functions have been put forward for transforming objective payoffs to proper utility functions that represent the subjects' real preferences (Griesinger and Livingston 1973; Tuomela 2000: 279–289; Guala 2006). Such transformations may result in a game that is different from the original description of the interaction situation. That explains why people exhibit behaviour that differs from the recommendations made for the original game.

The second possibility is not to rely on preference transformation but on what Bacharach (2006: 90) calls agency transformation. This means that the agent conceives the situation – not as a decision-making problem for individual agents – but as a decision-making problem for the group conceived as an agent. The idea is to conceptualize the situation, or 'frame' it, from the group's point of view: What would the group (e.g. a team) choose if it were the agent making choices? To answer this question, we need concepts that allow us to speak about what a group prefers, especially we need a notion of *group* preference.

Both agency transformation and preference transformation, when the latter is applied to account for the influence of group to an agent's preferences, can be seen as attempts to model what has been called *we-thinking*, thinking in terms of a 'we' instead of an 'I'. We concentrate on practical *we-reasoning*, which we take to be a subclass of we-thinking in general: Practical we-reasoning is we-thinking that aims at a decision

https://doi.org/10.1017/S0266267110000386 Published online by Cambridge University Press

about how to act. For instance, instead of asking 'What should *I* do in this situation?', the agents may ask themselves 'What should *we* do in this situation?'. The latter expression is ambiguous, however, since it can be understood in at least two different ways, as a combination of two questions: 'What should you do and what should I do?', or it can be understood as 'What should our group, having you and me as its members, do?'. An important aim of the present paper is to clarify these vague and broad characterizations of we-reasoning by distinguishing between two different types of we-reasoning: we-reasoning in the pro-group I-mode and we-reasoning in the we-mode. Pro-group I-mode reasoning answers 'What should I do as a private person acting in part for the group?', and we-mode reasoning answers two types of questions: 'What should our group do?' and 'What should I do as a group member as my part of our group's action?'.

The study of preference and agency transformations lies outside game theory proper. A central assumption in game theory is that it takes for granted a type of explanatory *methodological individualism*: explanations of action and of social phenomena more generally are to be couched in terms of intentional attitudes of individual agents. Another assumption, which concerns static games of complete and perfect information such as PD and Hi-Lo that can be represented as normal form matrices, is what we call the *givenness of preferences*: the preferences of the agents are exogenously given and are common knowledge among the agents involved in the situation.

There is a large body of literature concerned with criticism and defence of these and other central assumptions of standard game theory. Our aim here is not to criticize game theory in general. However, we want to point out that the assumptions do not always hold. In particular, standard game theory does not seem to apply to situations in which social groups are involved as agents and in which individuals act as group members exhibiting strong forms of we-reasoning. In situations in which the action alternatives and the preferences of the agents are not common knowledge the standard method of studying normal form games is not applicable either. One possibility would be to study dynamic games of incomplete information (see e.g. Aumann and Heifetz 2002) in which the agents may not know each others' preferences. However, since the interesting cases that have been studied in experiments are describable as normal form games (instead of more complex dynamic games), we will here consider ways in which agents could arrive at situations in which all the information contained in the matrices is common knowledge. This is an essential prerequisite for achieving reliable results on people's behaviour in situations like the PD and the Hi-Lo.

Our project is complementary to game theory since we will study also cases in which the above two assumptions do not hold, and we employ

https://doi.org/10.1017/S0266267110000386 Published online by Cambridge University Press

a richer conceptual framework found in philosophical action theory that employs concepts like reasons, goals and intentions. We do this (1) by relating the philosophical studies of collective intentionality to game-theoretic studies, and (2) by giving illustrative examples of processes by which preferences are communicated, modified and brought into common awareness.

As to (1), philosophical theories of collective intentionality study such phenomena as collective intentional behaviour (e.g. joint action), collective attitudes (e.g. group beliefs) and collective reasoning (e.g. collective decision making). A major question in the literature concerns whether the analysis of these phenomena requires conceptually or ontologically irreducible collectively intentional mental states or actions – for instance, collective attitudes that are not mere aggregates of individual attitudes – or whether the phenomena can be reduced to individual intentional states and actions. According to standard rational choice theory, cooperation and collective intentionality phenomena can be adequately explained without reference to irreducible group attitudes. A contrary approach is taken by philosophers such as Margaret Gilbert (1989), Raimo Tuomela (1995, 2000, 2007) and John Searle (1990, 1995) who argue that irreducible group attitudes are required for understanding cooperative behaviour. In economics, a related 'team thinking' approach has been advocated by Robert Sugden (1993, 2003) and Michael Bacharach (1999, 2006).

Here we adopt the *I-mode/we-mode* distinction made by Raimo Tuomela (2007), apply it to collective reasoning and decision making, and show its connection to Bacharach's (1999) theory of 'unreliable team interaction'. Intuitively, thinking and acting in the we-mode is thinking and acting as a proper group member, group understood as 'we' from the members' point of view. It contrasts with the I-mode that concerns thinking and acting as a private person. As standard game theory is limited to I-mode type of decision making, we argue that there is also we-mode type of decision making that people exhibit. Moreover, this kind of decision making can lead to better results (in a sense to be explained) than I-mode type of decision making.

As to (2), we relax the assumption of givenness of preferences and divide the decision-making process into consecutive stages. Theories of rational choice generally make the assumption that a complete description of the situation (consisting of the available action alternatives and the preferences or utilities over them) is available and is common knowledge among the agents. No attention is paid to how it is possible that the agents have become aware of these specifics of the situation. We give examples of processes that temporally precede and eventually result in situations that can be described as game-theoretic decision problems, that is, situations in which all information concerning the

https://doi.org/10.1017/S0266267110000386 Published online by Cambridge University Press

action alternatives and preferences of the agents have become common knowledge.

Finally, we relax both assumptions simultaneously and study situations in which agents form group preferences and select actions accordingly. Group preferences can be expressed by the group members as 'our group prefers that . . .'. The question where such group preferences come from is a problem for a theory of collective reasoning, and it is an important lacuna that accounts of team reasoning have not yet properly addressed (see e.g. Sugden 2000). We will here side with Gilbert (1989, 2001) and Tuomela (1995, 2007), who take group preferences to be based on the collectively accepted attitudes of the group in question. These collective attitudes are themselves typically *results* of deliberation and decision making and thus cannot be assumed to exist prior to deliberation. Modelling this kind of deliberation is another reason for distinguishing between different stages of we-reasoning.

Rational multi-agent deliberation involving we-thinking and we-reasoning covers the four areas depicted below:

	Reasoning prior to a decision problem	Reasoning in a decision problem
I-mode reasoning	A	B
We-mode reasoning	C	D

Game theory is basically a theory of reasoning and decision making focusing on B. Here we aim to discuss also reasoning of kinds A, C and D.

In this article we compare the we-reasoning account with Michael Bacharach's theory of interactive team reasoning. The primary objective of this comparison is to find out and create connections between game-theoretic accounts of collective reasoning and philosophical accounts of social action, in particular between Bacharach's theory of interactive team reasoning and the we-mode theory developed by Raimo Tuomela. First, it is shown that the we-mode account and the team reasoning account predict the same choices, e.g. in the Hi-Lo game. Second, the we-reasoning approach can be applied to account for the stages that precede the stage that team reasoning account starts from. Third, it is an important benefit of we-reasoning that it can provide the parameters that team reasoning presupposes, in particular, group preferences.

The outline of the article is as follows. In section 2, we introduce the central elements of our conceptual framework, especially the I-mode/we-mode distinction. Furthermore, we relate the distinction to Michael Bacharach's (1999) theory of 'unreliable team interaction' that attempts to extend game theory to situations in which groups can be viewed as agents and individual agents can reason as group members. We apply his results to show that I-mode decision making can produce different

https://doi.org/10.1017/S0266267110000386 Published online by Cambridge University Press

results from we-mode decision making: the equilibria in these cases may differ. More informatively, in these cases equilibria that are suboptimal for the group are eliminated in favour of optimal equilibria. This is one way in which we-mode reasoning can produce better results than I-mode reasoning.

In Section 3, we discuss the different stages of we-reasoning. We give examples of processes that lead agents to share information about their preferences and of processes that lead them to form group preferences. We argue that these preceding processes, in particular when they are of the we-mode type, can prevent agents from ending up in coordination problems with multiple equilibria and in collective action dilemmas with suboptimal equilibria. This is another way in which we-mode reasoning can produce better results than I-mode reasoning.

In Section 4, we recapitulate the main results of the paper and discuss their philosophical consequences.

2. WE-REASONING

Here we study we-reasoning and its relation to game theory. First, we present Tuomela's (2007) general distinction between the I-mode and the we-mode, and characterize the differences between we-reasoning in these two modes. Then we present Bacharach's (1999) methods of extending game theory to account for decision making based on we-reasoning. Finally, we relate Tuomela's philosophical framework to Bacharach's game-theoretic one, and we argue that Bacharach's results support the irreducibility of the we-mode to the I-mode.

2.1 The We-mode and the I-mode

Raimo Tuomela (2007) has presented a philosophical theory of sociality. According to his theory, a proper understanding of social phenomena requires irreducible group concepts, such as group goals, group values and we-thinking. His central conceptual tool is the distinction between the I-mode and the we-mode. The distinction is general and can be applied to various social phenomena. Here we apply it to decision making and we-reasoning.

According to Tuomela's theory, to think and act in the *we-mode* is to think and act fully as a group member. In contrast, to think and act in the *I-mode* is to think and act as a private person. The I-mode further divides into *pro-group I-mode* and *plain I-mode*. The pro-group I-mode is concerned with promoting the group's interests, whereas the plain I-mode is only concerned with promoting the interests of individuals even in group contexts. The we-mode approach is conceptually based on the intuitive idea that a group, not its individual members, is the primary agent. However, as a group's action must be based on its members' action,

https://doi.org/10.1017/S0266267110000386 Published online by Cambridge University Press

collective agency requires that the individual agents act as group members by doing their parts in satisfying the group-preferred alternative.

The we-mode framework is based on group concepts and especially a group's performative (and in this sense constructive) acceptance of attitudes and other group properties for the group. Formulating group acceptance in terms of its members' collective acceptance, we can use phrases like 'We hereby accept goal G as our group's goal' to express member-level collective acceptance. We assume throughout the paper that for everything that has been collectively accepted, it is common knowledge among the group members that it has been collectively accepted. The idea of a group agent can be based on an intuitive analogy: Analogously to intentional action of an individual agent, intentional action of a group agent must be based on reasons for actions. Analogously to an individual having to coordinate the movements of her body parts, a group's members must coordinate their action in order to achieve group goals. Analogously to an individual agent who is committed to her intended actions, the group members must be committed as a group, i.e. collectively committed, to the group's actions. Thus, we get three distinguishing criteria of the we-mode: authoritative group reason, (the satisfaction of a) collectivity condition and collective commitment, which are elaborated below.

1. The group members take the accepted attitudes or contents (e.g. that G is our group's goal) to be *group reasons* for their proper functioning as members. The members are assumed to recognize a group reason as authoritative, hence generally preemptive and also presumptive (viz. sufficient if conflicting group reasons are not present) reason for their action as group members. An authoritative group reason both requires the group members to act and directs their actions.
2. The contents of the attitudes accepted by and for the group can be assumed to satisfy a *collectivity condition*. Formulated for the special case of goal satisfaction this condition connects the members tightly as follows: necessarily (because of being based on the group members' reflexive construction of a goal as the group's goal), the goal is satisfied for a member if and only if it is satisfied for all (other) members.
3. The members are *collectively committed* to what they have accepted for the group. In the case of a group accepting a goal, the members collectively bind themselves to the goal and can be released only by the group. A central difference between the I-mode and the we-mode concerns commitments: In the plain I-mode I am committed to myself to furthering my own interests. In the pro group I-mode I am committed to myself to furthering the group's interests. In the we-mode I am committed to the group to furthering the group's interests.

https://doi.org/10.1017/S0266267110000386 Published online by Cambridge University Press

The notion of commitment here need not be binding in the game-theoretic sense.

To elucidate, thinking and acting as a group member is based on what the group decides, orders or requires (etc.). Here the group members voluntarily give some of their authority to the group. The group members form attitudes for their group by constructive collective acceptance. Examples are given by 'We as a group intend to build a bridge together' (group intention), 'We as a group believe that the euro is our currency' (group beliefs). Such attitudes can serve as group reasons, thus as premisses in the group members' practical reasoning.

Traditional game theory is, due to its commitment to methodological individualism, confined to the I-mode: the individuals are the only agents and they select actions aiming at satisfying their preferences in light of their expectations of the actions of the other agents who also are aiming at the satisfaction of their preferences. Game theory is neutral with respect to what the agents take into account in their preferences, and this makes it possible to accommodate some elements of we-reasoning into the theory. The preferences may be self-regarding or other-regarding, or they may aim at the benefit of the group's goals and interests. The last-mentioned case falls in the category of pro-group I-mode in the classification above.

A strategic aspect characteristic of the I-mode is that a rational agent tries to satisfy her preferences by selecting strategies that are best replies to others' expected strategies. In contrast, in the we-mode approach the group is viewed as an agent that functions strategically in a decision situation. For illustration, compare individual-agent games from the point of view of an individual qua a private person and from the point of view of the group consisting of all the agents involved: an individual agent can only select her strategy, but a group agent can in a sense select outcomes, because it can specify the strategies of all the agents in the group. In order to secure a specific outcome, it is necessary for an individual agent to act jointly with others.

The selection of an outcome is based on what the group prefers, and this in turn may, but need not, be functionally (in the mathematical sense) based on the preferences of the individual agents. The individual agents who reason in the we-mode and act as group members then realize the outcome by selecting those strategies that lead to it, assuming that the other group members do their parts. We-mode reasoning does not inherently involve (individual) strategic thinking like the I-mode does, although strategic aspects may sometimes be relevant, in particular, when some individuals in the group are reasoning in the I-mode or when there is more than one group involved.

Where game-theorists talk about selection of a strategy, philosophers interested in practical reasoning talk about the formation of an intention.

https://doi.org/10.1017/S0266267110000386 Published online by Cambridge University Press

Similarly, a group selecting an outcome corresponds to the group forming a collective intention, here indeed an intention attributable to the group. The individual agents realize the intention by performing their part-actions. Schematically, a paradigm case of we-reasoning that we study can be presented as follows:

(a) Group g intends to paint the house. (*group intention*)
(b) We (the members of g) jointly intend to paint the house together. (*joint intention*)
(c) I, qua a member of g and a participant in joint intention, we-intend to paint the house together with the others. (*we-intention*)
(d) I, qua a member of g and a participant in our joint intention to paint the house together with the others, intend to perform my part of our painting the house. (*part-performance intention*)

Here (a) is a group reason for (b), (b) is a reason for (c), while clause (d) follows from (c) on analytical grounds. Next we consider ways of modelling such we-reasoning in a game-theoretic framework.

2.2 We-reasoning in a game-theoretic framework

Here we apply the I-mode/we-mode distinction (sketched above and treated in detail in Tuomela 2007: ch. 2) to decision making in situations that can be described by normal form game matrices. We study processes of we-reasoning concentrating on the difference between we-mode reasoning and pro-group I-mode reasoning. The difference is not in the aims of the agents because in both cases the agents aim at the benefit of the group. Rather, the difference is in the reasoning process: It is individualistic in the I-mode case (both plain I-mode and pro-group I-mode). In game-theoretic terms, agents reasoning in the plain I-mode select actions that, given their expectations of the other agents' actions, best satisfy their preferences, which may be other-regarding in the sense of being affected by other agents' preferences. Agents reasoning in the pro-group I-mode select actions that, given their expectations of the other agents' actions, best satisfy their preferences, which are group-regarding, in the sense that they reflect the group's interests. We concentrate on the case in which the agent adopts the group's preferences as her own. An agent reasoning in the we-mode selects an action that is her component in the outcome that best satisfies the group's preferences.

All these distinctions can also be extracted from Michael Bacharach's (1999) game-theoretic framework. We-mode reasoning corresponds to his notion of team reasoning, pro-group I-mode reasoning corresponds to his notion of functioning as a benefactor to the group, whereas the plain I-mode reasoning corresponds to what he, somewhat misleadingly perhaps,

https://doi.org/10.1017/S0266267110000386 Published online by Cambridge University Press

calls individualistic reasoning with 'egoistic' utilities. Using Bacharach's framework, we argue that pro-group I-mode and we-mode decision making may produce different results, and that, in general, the we-mode helps to overcome social dilemma situations.

The we-mode, in contrast to the I-mode, also guarantees the achievement of Pareto-optimal equilibria in coordination games with one Pareto-optimal outcome, for instance the Hi-Lo game (in which both HH and LL are equilibrium outcomes):

	H	L
H	3, 3	0, 0
L	0, 0	1, 1

Intuitively, a we-mode reasoner chooses her part-action on the basis of the outcome that benefits the group most.

In general such reasoning presupposes the existence of a group utility function. We will later discuss the nature and formation of group preferences and group utility functions, but here we just assume that they exist. To simplify the exposition we will talk about intending to maximize the group's utility although what the agents intend is rather to do what most benefits the group. The following practical inference schemas illustrate the difference between the two cases. In the we-mode an agent would reason as follows:

1. We intend to maximize group utility.
2. Outcome HH uniquely maximizes group utility.
3. Therefore, I will perform my component in HH, viz. H.

In the pro-group I-mode she would reason as follows:

1. You and I intend to maximize group utility.
2. If you choose H, my choosing H maximizes group utility.
3. If you choose L, my choosing L maximizes group utility.
4. Therefore, I will perform ?

First, only the we-mode reasoning strictly entails some kind of conclusion, and, second, given that you reason analogously, only the we-mode strictly guarantees that we succeed in maximizing group utility. Although in both cases the matrix is commonly known, the pro-group I-mode leads to the same results only if the agent resorts to some controversial additional assumptions, such as the so-called payoff dominance principle by Harsanyi and Selten (1988: 355–357) that can conflict in risk-involving contexts with the often more appropriate risk dominance principle; see also Bacharach (1999: 135–137) and Colman *et al.* (2008a, 2008b).

https://doi.org/10.1017/S0266267110000386 Published online by Cambridge University Press

To be sure, our we-mode approach has not been formalized, partly because of its conceptual richness that goes beyond the narrow framework of game theory. Even our central criteria of the we-mode, viz. group reason, collectivity condition and collective commitment, do not seem to be game-theoretically formalizable. However, Bacharach's (1999) distinction between team reasoners and benefactors of a team is closely related to the respective notions of the we-mode and the pro-group I-mode in the context of non-cooperative games. The relationships between the two accounts can be stated concisely in the following theses:

(1) We-mode reasoning and Bacharach's team reasoning yield the same action recommendations in game-theoretic choice situations.
(2) Pro-group I-mode reasoning, in cases in which agents adopt the group preferences, and Bacharach's reasoning as a team benefactor yield the same action recommendations.

Claims (1) and (2) are assumptions that partially translate our framework into the game-theoretic one. With respect to (1), both reasoning methods have the same pattern: Given a group's preference ordering, the agents choose their part actions on the basis of the group-preferred outcome. With respect to (2), the pro-group I-moder and the team benefactor rely on the same pattern of reasoning: they adopt the group's preferences and select strategies that are best responses to other agents' predicted actions. In addition to similar reasoning patterns in Bacharach's team reasoning account and the we-mode theory, further parallels can be drawn between the two theories: Group utility in Bacharach's account plays the role of group reason, the team-reasoning pattern corresponds to the collectivity condition, and his 'group identification' to the notion of collective commitment.

Given the connections stated in our theses (1) and (2), it follows from Theorem 2 of Bacharach (1999), to be stated later in this paper, that the pro-group I-mode and the we-mode do not in all cases entail the same equilibrium behaviours. This is shown by comparing Bayesian games in which the players share the group's utility function (that each is trying to maximize individually) with games in which the players acting as a group try to maximize (expected) group utility in the we-mode sense (that is, by each selecting the best outcome in terms of group utility and then inferring their own part-actions). This is a strong argument for making and applying the distinction between the we-mode and the pro-group I-mode in one's theory. A simple example that illustrates the result is the Hi-Lo situation presented earlier: If the members are acting for the group, the we-mode approach only accepts the HH equilibrium whereas the pro-group I-mode account also allows for the LL equilibrium to be realized.

https://doi.org/10.1017/S0266267110000386 Published online by Cambridge University Press

To discuss this further, we will introduce some of the definitions of Bacharach's (1999) paper, partially replacing his terminology with ours. Thus, when Bacharach refers to team reasoning, we will refer to we-mode reasoning and we will speak of groups instead of teams.[1] When Bacharach refers to individual reasoning (as a special case of team reasoning with a singleton team consisting of only one agent), we will refer to I-mode reasoning. Instead of Bacharach's benefactor reasoning (individual reasoning pursuing group utility) we speak of reasoning in the pro-group I-mode, and instead of individual reasoning with egoistic payoffs we talk about reasoning in the plain I-mode.

In Bacharach's theory, an *unreliable team interaction* (UTI) is a structure $\langle M, A, U, S, T, \Omega \rangle$, where

- $M = \{M^1, \ldots, M^K\}$, where for each $k = 1, \ldots, K$, M^k is a subset of $\{1, \ldots, N\}$, is the set of group agents, viz. teams, which are collections of individual agents at least probabilistically disposed to we-reason together,
- $A = A_1 \times \cdots \times A_N$, where for each $i = 1, \ldots, N$, A_i is a set of acts of i,
- $U = \{U^1, \ldots, U^K\}$, where $U^k : A \to \mathbb{R}$ is the utility function of M^k,
- $S = S_1 \times \cdots \times S_N$, where S_i is the set of outside signals of i,
- $T = T_1 \times \cdots \times T_N$, where for each $i = 1, \ldots, N$, T_i is a subset of $\{1, \ldots, K\}$, is the set of participation states of i (that is, the value k of a participation state or mode t_i of individual i indicates the group agent M^k that i reasons for: in I-mode case $M^k = \{i\}$ and in the we-mode it is a larger group, typically $\{1, \ldots, N\}$),
- Ω is a probability measure on $Z = S \times T$.

Note that although groups have preferences over the possible outcomes, acts are defined only for individuals, so groups act only through their individual members in this model. Bacharach defines an *agent protocol* as a function from an agent's outside signals to actions and a *protocol for group agent, viz. the team*, M^k as an ordered tuple of agent protocols of the agents in M^k specifying the acts of each agent in the group. An agent, i, follows the protocol of the group she is active in, that is, the group as denoted by her participation state t_i. Thus, supposing that the protocol for M^k is $\alpha^k = (\alpha_i^k)$ for each agent i in M^k, then for each t_i in T_i and act a_i in A, if $t_i = k$, then $a_i = \alpha_i^k(s_i)$.

An *equilibrium of a UTI* is now a profile of protocols for each team such that each protocol is optimal for its team given the protocols of the

[1] We adopt a simplified account of group agents and teams for our present technical purposes and say that group agents are collections of individuals intentionally connected by their (probabilistic) disposition to use we-reasoning in appropriate decision-making contexts (e.g. in collective action dilemma situations).

https://doi.org/10.1017/S0266267110000386 Published online by Cambridge University Press

other teams. Assuming that the signal space S is finite, expected utility is defined as follows:

$$EU^k(\alpha) = \sum_{(s,t)\in Z} \Omega(s,t)U^k\left(\alpha_1^{t_1}(s_1), \ldots, \alpha_N^{t_N}(s_N)\right).$$

Then α^* is an equilibrium of the UTI $\langle M, A, U, S, T, \Omega\rangle$ iff α^* maximizes $EU^k(\alpha^k, \alpha^{*-k})$ where α^{*-k} denotes the acts of every other agent except k in protocol α^*.

As an example, consider the game matrix below containing both individual and group preferences. The first number in each cell represents the preference of the row player (the singleton group agent $M^1 = \{1\}$), the second the preference of the column player ($M^2 = \{2\}$), and the third the group preference (here the utility for the group $M^3 = \{1,2\}$ is simply an average of individual utilities):

	C	D
C	3, 3, 3	0, 4, 2
D	4, 0, 2	2, 2, 2

Suppose that each player reasons in the we-mode with probability ω. We will ignore the outside signal and only consider the participation state of each single agent. Thus, the possible signals are

- (1,2) (both are in the I-mode),
- (3,2) (1 is in the we-mode and 2 is in the I-mode),
- (1,3) (1 is in the I-mode and 2 is in the we-mode), and
- (3,3) (both are in the we-mode).

If both players operate in the I-mode (that is, if $\omega = 0$), this is an ordinary Prisoner's Dilemma and D is the rational choice for both players. However, with larger values of ω, (D,D,CC) becomes an UTI equilibrium. This is equivalent with CC being a best response for the group {1, 2} given (D,D) as the individual protocols, and D being a best response for an individual given (D,CC) as the protocols for the other individual and the group. We can calculate the expected utilities for the group {1, 2} as follows, given (D,D):

$$\begin{aligned} EU^{\{1,2\}}(CC) &= \omega^2 U^{\{1,2\}}(C,C) + \omega(1-\omega)U^{\{1,2\}}(C,D) \\ &\quad + \omega(1-\omega)U^{\{1,2\}}(D,C) + (1-\omega)^2 U^{\{1,2\}}(D,D), \\ EU^{\{1,2\}}(DD) &= \omega^2 U^{\{1,2\}}(D,D) + \omega(1-\omega)U^{\{1,2\}}(D,D) \\ &\quad + \omega(1-\omega)U^{\{1,2\}}(D,D) + (1-\omega)^2 U^{\{1,2\}}(D,D) \\ &= U^{\{1,2\}}(D,D), \text{ irrespective of the value of } \omega. \end{aligned}$$

https://doi.org/10.1017/S0266267110000386 Published online by Cambridge University Press

Similarly, we get $EU(CD)$ and $EU(DC)$, and the expected utilities for the individuals (taking here 1 as the reference individual), given (D,CC):

$$EU^{\{1\}}(C) = \omega^2 U^{\{1\}}(C, C) + \omega(1-\omega)U^{\{1\}}(C, D) + \omega(1-\omega)U^{\{1\}}(C, C) + (1-\omega)^2 U^{\{1\}}(C, D),$$

$$EU^{\{1\}}(D) = \omega^2 U^{\{1\}}(C, C) + \omega(1-\omega)U^{\{1\}}(C, D) + \omega(1-\omega)U^{\{1\}}(D, C) + (1-\omega)^2 U^{\{1\}}(D, D).$$

Supposing that the probability of agents reasoning in the we-mode is 0.9, we get the four matrices below. The label in the top left corner represents the group protocol that specifies actions for both group members when they are acting in the we-mode. In the first matrix, the group protocol is that both select C, in the second, player 1 selects C and 2 selects D, in the third, 1 selects D and 2 selects C, and in the fourth one, both select D.

CC	C	D
C	3, 3, 3	2.7, 3.1, 2.9
D	3.1, 2.7, 2.9	2.81, 2.81, 2.81

CD	C	D
C	0.3, 3.9, 2.1	0, 4, 2
D	0.49, 3.69, 2.09	0.2, 3.8, 2

DC	C	D
C	3.9, 0.3, 2.1	3.69, 0.49, 2.09
D	4, 0, 2	3.8, 0.2, 2

DD	C	D
C	2.01, 2.01, 2.01	1.8, 2.2, 2
D	2.2, 1.8, 2	2, 2, 2

Consider the first outcome in the first matrix which gives the expected utilities for players 1 and 2 and the group {1, 2} resulting from the protocol (C,C,CC) which assigns strategy C for player 1 playing in the I-mode, C for player 2 in the I-mode and strategy CC for the group (meaning that 1 selects C if in the we-mode and 2 also selects C if in the we-mode). It can be seen that player 1 would benefit from changing her strategy to D so (C,C,CC) cannot be a UTI equilibrium. Consider protocol (D,C,CC) in the lower left corner of the first matrix then. Player 2 would now benefit from switching to strategy D. Now we are at the lower right corner. Neither 1 nor 2 would benefit from changing strategy and neither does the group because the group utility in the lower right corner of the other matrices is smaller than in the first matrix. Thus (D,D,CC), meaning that I-moders

https://doi.org/10.1017/S0266267110000386 Published online by Cambridge University Press

should select D and we-moders C, is a UTI equilibrium of the game. In fact, it is the only equilibrium, and this holds not only for $\omega = 0.9$ but for all values $0 < \omega < 1$.

This kind of we-mode reasoning in which the agents select actions directly on the basis of the group's utility function may give results different from those obtainable from pro-group I-mode reasoning in which the agents transform their preferences by adopting the group's utility function as their own. To see this, we consider Bacharach's results further. Bacharach (1999) proves a lemma saying that if α^* is an equilibrium of an UTI $I = \langle M, A, U, S, T, \Omega \rangle$ then for each k and each $i \in M^k$, α_i^{*k} maximizes EU^k subject to the conditions that $t_i = k$ and that each $j \neq i$ follows α^*. This is needed to prove Bacharach's main result showing that UTIs differ from Bayesian games with the same utility function. More specifically, not all equilibria of Bayesian games generated by an UTI I are equilibria of I. This result can be regarded as showing that rational we-mode action is not functionally equivalent to pro-group I-mode action where the agents adopt the group utility but reason rationally in individualistic fashion.

First, Bacharach specifies what it is for a Bayesian game to be generated by an UTI as follows: The Harsanyi game generated by I is the Bayesian game $H_I = \langle M, A, U, S, T, \Omega \rangle$, where all other components are as in I but the utility functions depend on the signal so that if $t_i = k$ then $U_i = U_{M^k}$, that is, the signal (here taken to represent the *type* of the agent) specifies the group whose utility function the agent will adopt. With probability ω an agent will be a *team benefactor* (pro-group I-moder) and adopt the group utility function as her own whereas with probability $(1 - \omega)$ she will be what Bacharach calls an *egoist* (plain I-moder) and maintain her own individual utility function.

As an example, consider the Bayesian game generated by the UTI given above. The following matrices will represent the structure of the game with probabilities ω^2, $\omega(1 - \omega)$, $(1 - \omega)\omega$ and $(1 - \omega)^2$, respectively:

I	C	D
C	3, 3	2, 2
D	2, 2	2, 2

$p_{\mathrm{I}} = \omega^2$

II	C	D
C	3, 3	2, 4
D	2, 0	2, 2

$p_{\mathrm{II}} = \omega(1 - \omega)$

III	C	D
C	3, 3	0, 2
D	4, 2	2, 2

$p_{\mathrm{III}} = (1 - \omega)\omega$

IV	C	D
C	3, 3	0, 4
D	4, 0	2, 2

$p_{\mathrm{IV}} = (1 - \omega)^2$

We can now calculate the expected utilities to find out the equilibria of this game, and with $\omega = 0.9$ it turns out that the equilibria are ($C_B D_E$,$C_B D_E$) and

https://doi.org/10.1017/S0266267110000386 Published online by Cambridge University Press

(D_BD_E,D_BD_E), where, for instance, a strategy C_BD_E means that given the signal that one is a benefactor, one should choose C, and given the signal that one is an egoist one should choose D. In this example, the result is that an egoist should choose D but a benefactor may rationally choose either C or D. Thus, this kind of pro-group I-mode reasoning is incapable of eliminating the Pareto-inefficient equilibrium. The situation resembles Hi-Lo-type of games but is perhaps even more perplexing. Even in the following game (which is the same as the above matrix in which both agents share the group utility function), individualistic reasoning cannot uniquely recommend the strategy C to the players because in addition to CC also DD is an equilibrium:

I	C	D
C	3, 3	2, 2
D	2, 2	2, 2

Intuitively, C is the only rational strategy, because here, unlike in Hi-Lo, there is not even a conditional reason to select D instead of C because C always gives at least as high a utility as D.

Given our assumptions (1) and (2) above, Bacharach's theorem[2] illustrated by the example above entails our thesis (3):

(3) I-mode and we-mode reasoning, even when aiming at maximizing the same utility function, may produce different equilibria. Furthermore, the we-mode tends to create more collective order than the pro-group I-mode: It can decrease the amount of equilibria but it cannot increase them.

Since similar examples can easily be multiplied, the theorem gives a central rationality-based argument for the we-mode superseding over the (pro-group) I-mode, that is, for the following claim (4):

(4) If the group utility function is Paretian[3], then in several cases we-mode reasoning is rationally more rewarding for everyone in the utility maximizing sense than is I-mode reasoning.

This claim is partly an empirical one; involving that in some cases the we-mode yields a better result in terms of maximization of expected

[2] **Theorem 2** (Bacharach, 1999) *(i) for all UTIs I, for each equilibrium α of I there is a Bayesian equilibrium of H1 with which α agrees; (ii) for some UTIs I there is a Bayesian equilibrium of H1 with which no equilibrium of I agrees*.

[3] For all outcomes p, p', if p is at least as rewarding as p' to every agent in group G, then group utility $U^G(p) \geq U^G(p')$ (Bacharach 2006: 88).

https://doi.org/10.1017/S0266267110000386 Published online by Cambridge University Press

individual utility for group members rationally acting as group members than does the pro-group I-mode (see also Tuomela 2007: ch. 7). Together with the assumption that rational agents aim at maximizing utility, it in some cases entails a recommendation according to which agents should reason in the we-mode rather than in the I-mode.

The group members' switch to we-mode thinking, the 'we-mode switch', that is, reframing of a situation in we-mode terms, involves viewing the situation from the point of view of a group agent and its preferences and utilities. This together with the Paretian assumption can be seen to lead to group-rational mutual cooperation in such situations which without reframing would be identical to games with mutual interdependence of interests such as PD and Chicken, to the Hi joint outcome in the Hi-Lo situation, and so on. In other words, it leads to optimal choice for the group and tends to enhance its members' well-being, *ceteris paribus*. More precisely, this switch needs to occur at least in a certain proportion of the group members, not necessarily in all of them. Bacharach has investigated several situations in which the switch is expected to occur, given the relative frequency probability of switching (that he, contrary to us, at least in his 1999 paper takes to occur due to external circumstances).[4]

Finally, (1), (2) and (3) entail thesis (5):

(5) We-mode reasoning is not reducible to pro-group I-mode reasoning, i.e. it is not definable by or functionally constructable from I-mode reasoning. This is because it employs a different reasoning mechanism that relies on groups as the basic agents of reasoning.

To show the irreducibility of the we-mode to the pro-group I-mode, it suffices to show that there are cases where we-mode reasoning necessarily leads to different equilibria than does pro-group I-mode reasoning. Bacharach's mentioned theorem shows precisely this, given our bridge assumptions (1) and (2).

We also subscribe to the empirical claim that in many cases our account of we-mode reasoning gives a more correct description of how people actually reason than the I-mode description given by traditional game theory, but we will not argue for it here. For evidence, we note that there are empirical studies the results of which are better explained by

[4] To study the effects of different types of reasoning as in the previous example, we have formulated a spreadsheet file by means of which the changes in expected utilities and their maximization in arbitrary two-person games can be calculated as a function of the probability of being a we-moder versus being an I-moder given the same shared group utility in both cases. The file is available at http://www.cs.helsinki.fi/~hakli/uti-equilibria.html.

https://doi.org/10.1017/S0266267110000386 Published online by Cambridge University Press

the hypothesis that people use we-mode reasoning instead of competing I-mode hypotheses that are based on, for example, utility transformations. For recent results and discussion, see Colman *et al.* (2008a, 2008b), Krueger (2008), Sugden (2008) and Van Lange (2008), and also Guala *et al.* (2009).

Bacharach's theory thus allows game-theoretic modelling of both I-mode and we-mode decision making unlike traditional game theory that is only concerned with I-mode decision making. Furthermore, Bacharach's extension of game theory is a generalization of standard game theory in the sense that in cases in which the only agents are individuals and groups play no role, the individuals can be modelled as singleton groups with a group utility function that is identical to the agent's utility function, and the original Nash equilibria are preserved. The result can be taken to indicate that I-mode thinking is a special case of we-mode thinking in which the 'we' is limited to the agent herself.

It should be noted, however, that Bacharach's theory has some technical problems. One, which Bacharach mentions himself, is that the probability of agents acting in the we-mode is an exogenous variable; it is fixed beforehand and does not vary with agents nor depend on the interaction situation. As he notes (1999, 2006: ch. 2), the probability of an agent's identification with the group that stimulates team reasoning is affected by the perceived features of the interaction situation. We address this problem in the next section where we discuss the formation of group attitudes.

Another problem (that he does not mention) is that defining equilibria in terms of expectations makes his theory sensitive to changes in parameter values in the sense that changing the utility values can sometimes change the equilibria even though the preference orderings are preserved. Consider an instance of PD discussed by Bacharach (1999) in which the group utility is simply the average of individual utilities:

	C	D
C	3, 3, 3	−3, 4, 0.5
D	4, −3, 0.5	1, 1, 1

Let the probability of we-mode be 0.8. The strategy combination (D,D,CC) then is the only UTI equilibrium. If we change the individual utilities (without changing the individual preference orders) in the diagonal outcomes from 4 to 10, we get the matrix

	C	D
C	3, 3, 3	−3, 10, 3.5
D	10, −3, 3.5	1, 1, 1

https://doi.org/10.1017/S0266267110000386 Published online by Cambridge University Press

The equilibrium is not changed although the group preference ordering changed from $CC \succ DD \succ CD = DC$ to $CD = DC \succ CC \succ DD$. However, if we make another change from 10 to 11, we get

	C	D
C	3, 3, 3	−3, 11, 4
D	11, −3, 4	1, 1, 1

All the preference orderings remain unchanged but the UTI equilibria of this matrix are (D,D,CD) and (D,D,DC) instead of (D,D,CC), which is no longer an equilibrium. It might be worth investigating in detail precisely when phenomena like the above kind of artificial change of equilibria occur. We wish to emphasize, however, that cases like the plain Hi-Lo clearly show the existence of different equilibria resulting from we-mode and pro-group I-mode reasoning.

3. THREE STAGES OF WE-REASONING

As we have seen, both standard game theory and its extensions to we-reasoning make several tacit assumptions. Normal form game matrices presuppose that the agents have all-things-considered preferences expressed as numerical utilities and that there is common knowledge concerning them. Preferences can be understood in several different ways (see e.g. Hampton 1994; Rawling 2003; Binmore 2009), for instance, as representations of agents' propositional attitudes (beliefs and desires), modeller's tool of abstraction, or something revealed in the choices agents make. If it is assumed that the preferences are common knowledge among the agents, they cannot be only abstractions from, or something revealed in, agents' choices but have to represent something that the agents can be aware of prior to choice. Yet, how these preferences can become common knowledge is not explained but it is simply assumed that they are common knowledge. If they represent real attitudes of the agents, it is evident that an interaction situation may change the original preferences once the other agents' preferences become known. Such preference changes have been considered in the study of preference transformations, but it is not enough to take into account transformations that the agents perform or undergo privately, the *transformed preferences* should still become common knowledge. Once common knowledge is acquired, it may 'trigger' transformations, which have to become new common knowledge, which in turn may trigger new transformations, etc.

In this section we study the *processes* which may lead to situations describable as game-theoretic decision problems, that is, situations that can be modelled in terms of decision matrices the contents of which are

https://doi.org/10.1017/S0266267110000386 Published online by Cambridge University Press

common knowledge among the agents involved. These processes can be described as processes of we-reasoning. We-reasoning leading to the creation of a matrix can be performed in the I-mode or in the we-mode. Here we concentrate on the we-mode case and assume that there already is a group which has collectively accepted goals, in particular, the goal to find out what each group member wants, viz. the goal to create a game-theoretic matrix. In the I-mode case there need not be a group goal, but only individuals with a (shared) goal to create a matrix. We consider a collectively accepted goal X, thus a *we-mode goal* X that is based on collective acceptance and that satisfies the three criteria discussed above: authoritative group reason, the collectivity condition and collective commitment. Accordingly, first, goal X gives a member a 'favouring' reason to perform her part to satisfy X. Second, she cannot satisfy X for herself only. Third, she is collectively committed to satisfying X with her fellow members.

Here we apply the we-mode goal approach to the collective reasoning process, viz. to we-reasoning from the beginning till the very end. The participants' *first* goal is to acquire reliable information concerning the different action alternatives and their evaluation. The *second* goal is to form a joint intention on the basis of the information acquired in the first stage. The *third* goal is to satisfy the joint intention acquired in the second stage. Only the third goal directly involves interfering with the external course of events. The first two can be regarded as cases of collective problem solving.

This three-stage division can be made very clearly because each of the stages ends with the satisfaction of the corresponding goal. The first one ends when the parties have acquired the new information and when this information is common knowledge. The second one ends once they have reached the joint intention to act, and the third one ends once this intention is carried out. In a game-theoretic metaphor, the first goal is to create a decision matrix, the second goal is to choose an outcome from the matrix, and the third goal is to satisfy the chosen outcome. The we-mode goal account not only provides clear criteria for distinguishing different stages in 'the stream of we-reasoning' but it also serves as a unifying framework for accounting for the entire process.

Accordingly, the collective reasoning process that leads to the satisfaction of the joint intention is divided into three different stages of we-reasoning, WR1, WR2 and WR3 for short. For logical reasons, each preceding reasoning stage must have been carried through before any subsequent process can begin. The accomplishment of WR1 is a presupposition for WR2 and the accomplishment of WR2 is a presupposition of WR3. Each of the three stages of we-reasoning aims at giving a solution to a different problem for the group.

https://doi.org/10.1017/S0266267110000386 Published online by Cambridge University Press

The previous section was concerned with the decision-making stages, that is, WR2 and WR3, which are the stages that game theory and Bacharach's theory of unreliable team interactions deal with. Individual preference matrices are taken as given in game theory and their creation is discussed in section 3.1. Bacharach's UTIs presuppose, in addition, group preferences and the parameter Ω, and we will discuss them in section 3.2.

3.1 Creation of an individual preference matrix

Group members often end up in situations where joint action could be desirable but they are not aware of how other members would rank the available outcomes. When the agents try to deliberate privately, by themselves, they may end up in incomplete matrices, or, even if they manage to create a complete normal form matrix by themselves, e.g. one that mirrors the objective payoffs, they do not know if it would serve as a trustworthy basis for further deliberation and action, or they do not know if the others are aware of it. In these cases they can join forces to find a solution in the we-mode by adopting a we-mode goal to create a matrix. For simplicity, we talk about the creation of a matrix, but it of course need not be produced in matrix form, let alone in written form. Game-theoretic matrices are tools for modelling; what the agents are sharing and producing is the information represented in the matrix.

Here they have two possibilities for a we-mode goal: the members might be satisfied in creating a standard game-theoretic matrix, viz. all-things-considered individual preference matrix, or they might attempt to create a group preference matrix. The difference between the two is that in the latter case there will be a single common preference ranking that is collectively accepted for the group. In the former case, the members have their separate preference rankings that are acknowledged by the others. These separate rankings may, but need not, converge, and a collective choice can be attempted on the basis of them without trying to create a common group preference ranking. We will here discuss the formation of such individual preference matrices, and the formation of group preferences is discussed in section 3.2.

Consider a two-person two-choice case where the players are you and I. First we have to find out what each of us wants to be the case, thus what our individual all-things-considered preferences are concerning the outcomes CC, CD, DC, DD. Note that WR1 could have been preceded by another type of we-reasoning concerning what in fact it is possible for the agents to do on a given occasion, i.e. what the external opportunities for action are. However, which of these possibilities are relevant is up to the agents. Accordingly, they can 'we-reason' to produce a list of the

https://doi.org/10.1017/S0266267110000386 Published online by Cambridge University Press

relevant options that are available. This we-reasoning could be termed WR0. However, even though the relevant available options would be set and known at the outset, what need not be known is how the agents value these options, and WR1 deals with this. The we-reasoning involved in WR1 can be represented by a syllogism that concerns means and ends. Here we consider the case where the parties are collectively committed to constructing a matrix mapping their individual all-things-considered preferences of which there will be common knowledge. The following simple practical inference describes each participating person's we-reasoning in the first person form:

1. We intend to achieve a genuine common matrix of our preferences and to acquire mutual awareness of it.
2. Unless (a) I rank the available outcomes into a preference order, and unless (b) I sincerely express my preference order to you, we will not achieve this goal (the genuine matrix of which there is common knowledge).
3. Therefore, I intend to see to it that (a) and (b).

This pattern of reasoning can be considered as 'conceptually binding' (if one accepts the premises, one cannot rationally deny the conclusion). Suppose that the parties have succeeded in carrying out the conclusions of their corresponding practical reasoning. As a result, given that the flow of information is not blocked or distorted, they have obtained a matrix M1 the contents of which are common knowledge. The acquired matrix, M1, may represent everyone's preferences in which case WR1 is over and the members continue to WR2 on the basis of M1. But it may also contain information that leads to a preference change in which case M1 is transformed into another matrix, M2, and they continue on the basis of M2, or M2 is transformed to M3, and so on.

For example, consider a group, a family consisting of you and me, hiking in the Alps. To reach our destination, we have to decide which route to take and whether to go together or to take separate ways. Assume that there are two choices: One route is steep and short, the other is not so steep but longer. It is common knowledge that we both have two possible actions, S and L, denoting the steep and long routes, respectively. Assume that we both enjoy the company of one another and consequently want to take the same route. In addition, you think that we should take the longer route together since you believe that I am an inexperienced hiker. In the absence of other considerations I would like us to take the steep and short route but, recalling that you are an expert, I find it more important that we do as you prefer since I think that you are an expert. But I also mistakenly think that you prefer the steep route so that, all things considered, I prefer SS to LL. We then tell each other our preferences and what we get is the

https://doi.org/10.1017/S0266267110000386 Published online by Cambridge University Press

preference structure of the battle of sexes game that can be expressed in normal form as follows:

	L	S
L	1, 3	0, 0
S	0, 0	3, 1

When I learn what your preferences are, I get new relevant information that affects my preferences: You do not prefer the steep route like I thought but prefer the long route instead. As I attach a lot of weight to your opinion, my all-things-considered preferences are changed accordingly, and I communicate my preference change to you. Your preferences are not changed, and we agree that our situation has the preference structure of a Hi-Lo game:

	L	S
L	3, 3	0, 0
S	0, 0	1, 1

The resulting matrix is a *genuine preference matrix* if and only if the order of the individual utility figures matches the order of the preferences of the individuals. If a group member lapses from the we-mode and strategically announces a false preference order, the goal cannot be satisfied and the resulting matrix will not be genuine, even if the other group member would falsely believe so. One can always try to arrive at a matrix by I-mode reasoning, but in that case the resulting matrix need not be genuine, as the parties can keep their true preference rankings to themselves.

As pointed out by Rubinstein (1991), a good game-theoretic model should describe the features of the interaction situation, not as they are realized in the world, but as they are perceived by the decision makers. To reach such a description in experimental situations, something like the we-reasoning process of stage WR1 could be simulated by allowing the subjects to exchange information about their preferences over the possible outcomes until these preferences stabilize. Accomplishing this in practice involves several difficult questions such as how to ensure that the subjects communicate their actual preferences, and, if this is solved, how to design experiments in which the subjects end up with final matrices representing situations that the experimenters are interested in, for instance, PD type of situations. However, these problems should be addressed if one wants to make any kind of comparison between game-theoretic predictions and people's actual behaviour. In order to be able to draw accurate conclusions about people's behaviour in various interaction situations it is necessary to make sure that the experimenters know what the game is that the subjects take themselves to be playing.

https://doi.org/10.1017/S0266267110000386 Published online by Cambridge University Press

Theoretical accounts of collective action dilemmas typically start with some matrices, e.g. PD, Chicken or Hi-Lo. Given these matrices, they may discuss various reasoning patterns or matrix transformations that lead to different collective or individual choices. 'Where the matrices came from' is not usually considered a problem that would require explanation. To account for the WR1 process is to account for the emergence of the matrices. Sometimes the situation is so obvious (the matrix is given and the parties are mutually aware of it) that the parties can directly begin with WR2. Sometimes it is enough to find only one mutually satisfactory or 'satisficing' action alternative (as in the bulletin board view by Tuomela, 2000: 59–61 and 2007: 89, and the 'offer-acceptance' model by Bach, 1995). Sometimes there is no agreement over a description of the situation or the agents realize that none of the alternatives will be mutually satisfactory and they may return to stage WR0 to expand the space of alternatives. We have emphasized the WR1 stage because in standard accounts it is typically overlooked or fused together with the subsequent stages, and this is apt to produce confusion.

In general, the following argument shows that we-mode reasoning can prevent the agents from framing an interaction situation as a PD even though the objective payoffs form a PD structure as follows:

1. In a genuine single-shot PD, individually rational agents defect.
2. However, it is not necessarily the case that rational agents frame the objective situation individualistically as a PD.
3. If one reasons in the we-mode and adopts a group perspective, the situation is not framed as a PD.
4. Therefore, the we-mode can prevent the emergence of a PD.

So far we have shown that we-mode reasoning at stage WR1 can be more beneficial than the I-mode on three accounts: (i) It can account for the emergence of the normal form matrix in cases where the I-mode fails. For example, privately they only succeed in creating an incomplete matrix. (ii) It guarantees that the resulting matrix is genuine in the sense of expressing the participants' true preferences. (iii) It can prevent the emergence of collective action dilemmas.

The prevention of collective action dilemmas can take place in various ways. For example, some agents 'by nature' persist in framing interactive situations as cooperative: they are 'disposed' to identify with their group and to team reason accordingly (cf. Bacharach 2006). For the present paper, the most interesting possibility of preventing collective action dilemmas is the we-mode creation of group preferences. This will be discussed in the following section.

https://doi.org/10.1017/S0266267110000386 Published online by Cambridge University Press

3.2 Creation of a group preference matrix

Our discussion concerning the formation of group preferences is related to the theory of social choice, which studies the aggregation of individual preference orderings to a collective preference ordering. Social choice theory has resulted in several interesting mathematical results concerning the relationship between individual and collective preferences, but it does not focus on the social processes leading to the formation of group preferences. (See Gaertner (2006) for preference aggregation and List and Pettit (forthcoming) for a good nontechnical but state-of-the-art presentation of aggregation of judgements.) Here we consider how actual group preferences can be created, and we do not assume that the group preferences over certain outcomes must be a function of individual preferences over the same outcomes.

There are also dynamic approaches that are concerned with consensus formation, see e.g. Lehrer and Wagner (1981). Such approaches could be applied to forming a rational collective preference ranking, see Tuomela (1995: 433). Also these approaches typically assume that individual preferences are primary and the collective ones are arrived at by some aggregative means.

If group preferences are not a function of the individual group members' preferences, something more needs to be said about them. As discussed, preferences of individuals can be understood to be based on the beliefs and desires of the individuals. However, the case with groups is more problematic, because it cannot be simply assumed that groups have such attitudes. According to Bacharach (1999), the group's utility function need not be reducible to the utility functions of the individuals, but in his examples group utilities are typically simple averages of the utilities of the individuals. Sugden (2000) discusses the nature of group preferences. According to him, at the level of the group they are rankings of outcomes that are revealed in the group's decisions, whereas at the level of the group members they are preferences that govern their group-related decision making and are shared by the group members. He does not say much about their formation, though.

We take group preferences to be based on the collectively accepted goals and other intentional attitudes of the group. Of any two outcomes the one that better promotes these goals can rationally be taken to be preferred by the group, and assuming that the group's preferences satisfy the standard axioms of utility theory, a group utility function can be derived. As a simple example, consider a business firm, the only goal of which is to maximize profit. This entails that the outcome that maximizes profit is the preferred outcome for the group.

As discussed earlier, the group's preferences are not necessarily functionally dependent on the group members' preferences, especially

https://doi.org/10.1017/S0266267110000386 Published online by Cambridge University Press

not if the we-mode view of a group agent accepting attitudes for itself is adopted. In the case of a business firm, its personnel may be required to act as group members in order to achieve the firm's objectives. If the firm's only goal is to maximize profit, the preferences of the personnel need not affect the group's preferences at all. However, the personnel can be seen as just a subgroup of the whole group. The latter includes the owners who have the power to decide the goals of the firm. The preferences of the group cannot in all cases be reduced to the preferences of the owners either, since the preferences of the owners may change without the change being immediately reflected in the firm's objectives.

Group preferences can be based on goals and other collectively accepted attitudes but sometimes the group preference ordering itself can be collectively accepted. Above, we discussed the WR1 process where the agents created an ordinary game-theoretic preference matrix. In a similar vein, the agents reasoning in WR1 may take reaching a group preference matrix as their goal. Here each group member gives her proposal for the group's preferences: 'According to me, the group should put the joint outcomes in the following order . . .'. As intermediate results, we may get different suggestions for a group preference ordering, and they may be transformed on the basis of other agents' suggestions and group discussion. In terms of matrices, the ultimate aim is to reach a matrix with only one number for each outcome, representing the group utility function.

As an example, consider again the case of our family hiking in the Alps. Previously we were concerned with finding out our individual preferences concerning the selection of the route, but now we evaluate the options from the group's perspective. This may affect the ordering, because we are not necessarily considering our own preferences concerning the four outcomes but our preferences concerning the group's preferences concerning these outcomes. When we evaluate the outcomes we are trying to think of our group as an agent who is doing the evaluation. As previously, all relevant things are considered but from a different viewpoint that may change the relative weights of the matters to be considered. Also what is relevant is different in the two cases. For instance, we may add more weight to the consideration that a longer route gives us more time to spend together and it thus strengthens our family – the collective agent consisting of the two of us. The final matrix agreed upon may become something like this:

	L	S
L	3	1
S	1	2

The formation of a group preference matrix may be more demanding than the formation of a game-theoretic matrix since an agreement must

https://doi.org/10.1017/S0266267110000386 Published online by Cambridge University Press

be reached. The agents may try to reach a consensus, they may need to make compromises, negotiate, present arguments, or they may end up voting on the preference ordering. Whether and how a group preference matrix is reached is a contingent matter. What is important is whether it is collectively accepted by the group members that it adequately represents the group's preferences. Collective acceptance may come about by implicit agreement and can be weaker or stronger, depending on their shared background knowledge and on the nature of their relationship. A group preference matrix is genuine if and only if the preference ordering it specifies is collectively accepted or is inferrable by any competent group member from what the group has collectively accepted.

As seen, the group preference ordering concerning certain outcomes is not necessarily reducible to individual preference orderings concerning those outcomes, for several reasons: First, the formation of group preference requires collective acceptance. Second, once collectively accepted, a preference ordering remains the group's preference ordering until it is collectively changed, even though individual preference orderings may change. Third, the collective acceptance of a preference ordering with respect to certain outcomes $O_1, \ldots, O_n$ need not be based on individuals' preference orderings concerning $O_1, \ldots, O_n$, but on their opinions on what preference ordering the group should adopt, and these two things may differ.

In addition to individual and group preferences, Bacharach's UTI matrices require the probability function Ω. In our view, Ω can be interpreted as expressing the group's view concerning the probability of group members selecting actions on the basis of the group preferences. The value of this parameter can be estimated in the WR1 process similarly to the formation of the preferences. After the group preferences have been agreed upon, the group members can give their estimates of Ω and update their estimates based on the estimates of others (perhaps by methods described by Lehrer and Wagner 1981). If the estimates finally converge, all the information required for describing the situation as a UTI becomes available and common knowledge. Moreover, the estimate of Ω now depends on the interaction situation described in the matrix, unlike in Bacharach's original theory in which Ω is an exogenous variable.

In everyday life these kinds of estimations occur frequently. A group first evaluates the outcomes and then it estimates the number of group members willing to participate in reaching the preferred outcomes. These parameters, the group utility function and the probability function, then define the equilibria that specify whether it is rational for the group to pursue the most preferred outcome or some other outcome. A group, for example a sports club, may judge that organizing a sports event would be better for the group than organizing a jumble sale because it would be more effective in terms of fund raising and it would give more publicity

https://doi.org/10.1017/S0266267110000386 Published online by Cambridge University Press

for the group. (This stage corresponds to the formation of the group preferences in WR1.) However, it would also require much more effort from the group members, so there might not be enough volunteers willing to participate in organizing the event. (This stage concerns the estimation of Ω in WR1.) Consequently, the group chooses to organize a jumble sale instead. (This stage belongs to WR2.)

If the group members can agree on a group preference ordering, they can collectively accept it so that it satisfies the three conditions of the we-mode. If they then reason in the we-mode in accordance with these group preferences, their intragroup situation cannot be described as a PD or any other collective action dilemma.

The jointly produced matrix M1, i.e. the result or the output of WR1, of which there is common knowledge, serves as input to WR2. The second we-goal would be to choose an outcome from the matrix, viz. the second goal is to reach a joint intention to act. This goal is satisfied when the agents succeed in forming we-intentions with the content of satisfying a particular outcome. Once the outcome is collectively fixed, the parties are committed to achieving it, and they can proceed to WR3 in which the agents select their part-actions that together result in achieving the outcome. Note that the reasoning processes in stages WR2 and WR3 can be performed by the individual agents autonomously as generally supposed in non-cooperative game theory. In particular, in cases in which there is a unique outcome that maximizes group utility, or in cases in which there is a unique UTI equilibrium, agents reasoning in the we-mode will be guaranteed to reach the optimal outcome. If there are several equally good alternatives, the agents will still face a coordination problem, and communication will be required in order to guarantee an optimal result. As long as such communication does not lead to all-things-considered obligations to obey the results of the communicative process, our treatment still falls within the boundaries of non-cooperative game theory, as usually defined.

4. CONCLUDING REMARKS ON THE WE-MODE VERSUS THE I-MODE

In this paper we have compared two kinds of we-reasoning in a collective decision-making situation: pro-group I-mode reasoning and we-mode reasoning. We have discussed these reasoning methods both in the game-theoretic framework of Bacharach (1999) and in the framework of philosophical action theory, in particular Tuomela's theory of joint action. We have highlighted the connections and differences between these two frameworks. Tuomela's theory adds conceptual machinery lacking in Bacharach's game-theoretic framework, whereas the results obtained by Bacharach support Tuomela's idea of the irreducibility of the we-mode to the I-mode. In addition, we have shown that our stage model of

https://doi.org/10.1017/S0266267110000386 Published online by Cambridge University Press

we-reasoning can provide parameters that Bacharach's theory takes for granted, in particular the group preferences.

The main results obtained in this paper concerning the differences between pro-group I-mode reasoning and we-mode reasoning are as follows:

1. Pro-group I-mode reasoning and we-mode reasoning yield different action recommendations. For instance in Hi-Lo, we-mode reasoning entails a specific action recommendation but pro-group I-mode reasoning does not.
2. We-mode reasoning can lead to more beneficial results than the I-mode because it eliminates suboptimal equilibria and thus creates more collective order. This result follows from Bacharach's (1999) theory of unreliable team interactions.
3. We-mode reasoning can lead to more beneficial results than the I-mode because it can prevent the emergence of collective action dilemmas by leading the agents to frame the situation as a group decision problem.
4. Even if the agents do not succeed in framing the situation as a group decision problem by creating a group preference matrix, but end up in a standard game-theoretic decision matrix, the matrix genuinely reflects their preferences if it is produced in the we-mode.

These results have implications for the design of social institutions. Because we-mode thinking yields more collective order than individualistic thinking, a rational desideratum is that social institutions be designed to encourage we-mode thinking. The results should also be taken into account in social-scientific theorizing. If we-mode reasoning and I-mode reasoning produce different rational behaviour and if, as the empirical results of Colman *et al.* (2008a, 2008b) suggest, human beings sometimes in fact reason in the we-mode, methodological individualism is not sufficient as a foundation of social sciences either in a prescriptive, rational or in a descriptive sense. Theories that rely on the assumption of methodological individualism (recall section 1) need revision or complementation. Adequate conceptualization, description and explanation of social activities (especially joint activities) of intentional and rational beings require the use of group concepts and the we-perspective, especially in its we-mode form.

REFERENCES

Aumann, R. J. and A. Heifetz, 2002. Incomplete information. In *Handbook of Game Theory with Economic Applications, volume 3*, ed. R. Aumann and S. Hart, 1665–1686. Amsterdam: Elsevier.

Bach, K. 1995. Terms of agreement. *Ethics* 105: 604–612.

Bacharach, M. 1999. Interactive team reasoning: A contribution to the theory of co-operation. *Research in Economics* 53: 117–147.

https://doi.org/10.1017/S0266267110000386 Published online by Cambridge University Press

Bacharach, M. 2006. *Beyond Individual Choice: Teams and Frames in Game Theory*, ed. N. Gold, R. Sugden. Princeton: Princeton University Press.
Binmore, K. 2009. *Rational Decisions*. Princeton: Princeton University Press.
Bowles, S. and H. Gintis 2006. Social preferences, homo economicus, and zoon politikon. In *The Oxford Handbook of Contextual Political Analysis*, ed. R. E. Goodin and C. Tilly, 172–186. Oxford: Oxford University Press.
Colman, A., B. Pulford and J. Rose 2008a. Collective rationality in interactive decisions: Evidence for team reasoning. *Acta Psychologica* 128: 387–397.
Colman, A., B. Pulford and J. Rose 2008b. Team reasoning and collective rationality: Piercing the veil of obviousness. *Acta Psychologica* 128: 409–412.
Gaertner, W. 2006. *A Primer in Social Choice Theory*. Oxford: Oxford University Press.
Gilbert, M. 1989. *On Social Facts*. London: Routledge.
Gilbert, M. 2001. Collective preferences, obligations, and rational choice. *Economics and Philosophy* 17: 109–119.
Griesinger, D. W. and J. W. Livingston 1973. Toward a model of interpersonal motivation in experimental games. *Behavioral Science* 18: 173–188.
Guala, F. 2006. Has game theory been refuted? *Journal of Philosophy* 103: 239–263.
Guala, F., L. Mittone and M. Ploner 2009. Group membership, team preferences, and expectations. CEEL Working Paper 6-09, Computable and Experimental Economics Laboratory, Trento, Italy.
Hampton, J. 1994. The failure of expected-utility theory as a theory of reason. *Economics and Philosophy* 10: 195–242.
Harsanyi, J. C. and R. Selten 1988. *A General Theory of Equilibrium Selection in Games*. Cambridge: MIT Press.
Krueger, J. I. 2008. Methodological individualism in experimental games: not so easily dismissed. *Acta Psychologica* 128: 398–401.
Lehrer, K. and C. Wagner 1981. *Rational Consensus in Science and Society*. Dordrecht: D. Reidel Publishing Company.
List, C. and P. Pettit, forthcoming. *Group Agency: The Possibility, Design and Status of Corporate Agents*.
Rawling, P. 2003. Decision theory and degree of belief. In *The Blackwell Guide to the Philosophy of the Social Sciences*, ed. S. P. Turner and P. A. Roth, 110–142. Malden: Blackwell.
Rubinstein, A. 1991. Comments on the interpretation of game theory. *Econometrica* 59: 909–924.
Sally, D. 1995. Conversation and cooperation in social dilemmas: A meta-analysis of experiments from 1958 to 1992. *Rationality and Society* 7: 58–92.
Searle, J. 1990. Collective intentions and actions. In *Intentions in Communication*, ed. P. Cohen, J. Morgan and M. E. Pollack, 401–415. Cambridge: MIT Press.
Searle, J. 1995. *The Construction of Social Reality*. London: Penguin Press.
Sugden, R. 1993. Thinking as a team: Towards an explanation of nonselfish behavior. *Social Philosophy and Policy* 10: 69–89.
Sugden, R. 2000. Team preferences. *Economics and Philosophy* 16: 175–204.
Sugden, R. 2003. The logic of team reasoning. *Philosophical Explorations* VI: 165–181.
Sugden, R. 2008. Nash equilibrium, team reasoning and cognitive hierarchy theory. *Acta Psychologica* 128: 402–404.
Tuomela, R. 1995. *The Importance of Us: A Philosophical Study of Basic Social Notions*. Stanford: Stanford University Press.
Tuomela, R. 2000. *Cooperation: A Philosophical Study*. Dordrecht: Kluwer.
Tuomela, R. 2007. *The Philosophy of Sociality: The Shared Point of View*. New York: Oxford University Press. (Paperback ed. 2010)
Van Lange, P. A. M. 2008. Collective rationality: The integrative model explains it (as) well. *Acta Psychologica* 128: 405–408.

https://doi.org/10.1017/S0266267110000386 Published online by Cambridge University Press