Economics and Philosophy, **16** (2000) 175–204 Copyright © Cambridge University Press

TEAM PREFERENCES

Robert Sugden
University of East Anglia

When my family discusses how we should spend a summer holiday, we start from certain common understandings about our preferences. We prefer self-catering accommodation to hotels, and hotels to campsites. We prefer walking and looking at scenery and wildlife to big-city sightseeing and shopping. When it comes to walks, we prefer walks of six miles or so to ones which are much shorter or much longer, and prefer well-marked but uncrowded paths to ones which are either more rugged or more popular. And so on. These common understandings greatly simplify the task of choosing between holiday destinations and activities, by allowing us quickly to eliminate many options. But what does it mean to say that *we* prefer one thing to another?

Does it mean that *each of us* – my wife, my son, my daughter, me – prefers this? Not quite. What I have called 'our' preferences are not so very different from those that I have as an individual. I believe that the same is true for my wife; and I like to think it is at least broadly true for my teenage children. But 'our' preferences are not exactly those of any one of us. My ideal walk would be somewhat longer than six miles, along rougher and less well-marked paths than we prefer as a family. Visits to gift shops rank rather lower in the family's preferences than they do in my daughter's, but much higher than they do in my son's.

The ideas presented in the paper have evolved over many years, in discussions with many people, but particularly with Michael Bacharach, Nick Bardsley, Luigino Bruni, Robin Cubitt, Jean Hampton, Shaun Hargreaves Heap, Martin Hollis, Maarten Janssen, William Kline, Judith Mehta, Chris Starmer and Bruno Verbeek. A previous version of the paper was presented at a seminar on Rationality and Intentions in Amsterdam in October 1999. I thank the participants at that seminar, especially Austin Dacey-Groth, David Gauthier, Margaret Gilbert, Anthony Laden and Chris Morris, and also Rupert Read and two anonymous referees, for comments and suggestions. My work was supported by the Leverhulme Trust.

https://doi.org/10.1017/S0266267100000213 Published online by Cambridge University Press

And so on. So, it seems, 'We prefer x to y' is not equivalent to 'Each of us prefers x to y'. Nevertheless, it is an apparently meaningful proposition which can help in some kinds of reasoning about decision problems. More specifically, when the *we* whose preference it is are reasoning about what *we* should do, it helps *us*.

In common-sense terms, I take it, there is nothing surprising about what I have said so far. But in relation to the received theory of rational choice, it is a heresy. In that theory, the only ultimate actors are individual human beings: it is individuals, not groups, which face decision problems. The question 'What should *we* choose?' simply cannot be formulated within the theory. Similarly, preferences are attributed only to individuals. It is sometimes allowed that collective preferences might be constructed by aggregating individuals' preferences in some way; but Arrow's impossibility theorem, and the failure of social choice theory to find any acceptable escape from that result, have generally been taken as showing that the aggregation approach leads to a dead end.

The prevailing view among choice theorists today, I think, is that problems of collective choice should be modelled as non-cooperative games in which the players are individuals, each with his or her own preferences. The procedures by which collective decisions are reached – for example, voting, or bargaining by offer and counter-offer – are modelled as properties of a game. Each individual acts on his own preferences within that game, aware that the other players are similarly motivated. We might want to call the result of this process a 'collective choice', but its theoretical status is simply that of an outcome of a game played by individuals, acting as individuals.

In this paper, I shall argue that the theory of choice *should* allow 'teams' of individuals to be decision-making agents and *should* allow such teams to have preferences. Further, I shall argue for an interpretation of 'team preference' in which the preferences of a team are not necessarily reducible to, or capable of being constructed out of, the preferences that govern the choices that the members of the team make as individuals.

My argument is addressed primarily to people who accept the standard theory of rational individual choice. I try to show that the concept of team preference, as I define it, is no more problematic, mysterious or question-begging than the concept of individual preference, as used in the received theory. The role that team preferences play in my theory of team agency is essentially the same as the role that individual preferences play in the standard theory of individual agency. Formally, the concept of team preference is a generalization of the standard concept of individual preference. Thus, I shall argue, someone who accepts the standard theory as a valid form of explanation of

https://doi.org/10.1017/S0266267100000213 Published online by Cambridge University Press

human behaviour should have no objection of principle to my account of team agency.

The idea that individuals might act as members of collectives has occasionally been proposed in the literature of rational choice theory. The most common starting point for such proposals has been the recognition that individually rational agents, as represented in the received theory of rational choice, can fail to find apparently obvious solutions to coordination problems. After some methodological ground-clearing (Section 1), I explain the nature of this problem (Section 2) and present the intuitive idea that it might be solved by some kind of team agency (Section 3). If we think of team agency as a means of solving coordination problems that are faced by rational individuals, it is natural to take the preferences of those individuals as fundamental, and to look for configurations of such preferences which *activate* team agency. (For example, we might define a class of coordination games and then claim that when individuals play games of this class, they respond – or perhaps, ought rationally to respond – by engaging in team reasoning.) Similarly, if team reasoning requires there to be some kind of team objective, it is natural to suppose that team objectives are constructed out of individuals' preferences. In Sections 4 and 5 I discuss various attempts to analyse team agency in terms of individuals' preferences, and argue that these analyses are inadequate as general representations of team agency in decision-making. In Section 6 I discuss another analysis, in which team agency is created when individuals openly express their willingness to participate in it, and in which this act of creation generates obligations for individuals to act on team reasons. I question the force of these obligations.

I propose a different way of thinking about team agency, in which the preferences that individuals have as members of teams are distinct from, but on a par with, the preferences that guide their private choices, and which does not involve concepts of obligation. Section 7 introduces these ideas and shows how they might be developed into a formal theory. This theory may seem to be open to two objections: it does not explain why collections of people do or do not take themselves to be teams, but simply takes the existence or non-existence of teams as given; and it does not explain why team preferences are as they are, but simply takes them as given. In Sections 8 and 9 I consider these objections. I argue that in taking these things as given, my theory does no more than the standard theory of rational choice does in taking the 'framing' of decision problems and individuals' preferences as given. This argument might be read as following a companions-in-guilt strategy. In the final section, I consider whether the similarities between my theory and the received theory really do amount to companionship in *guilt*, or whether the theoretical strategy of taking frames and preferences as given can be

https://doi.org/10.1017/S0266267100000213 Published online by Cambridge University Press

understood as part of a valid and productive methodology for economics.

1. SOME METHODOLOGICAL GROUND-CLEARING

Let me begin by saying something about the problem I take myself to be addressing and about the philosophical and methodological status that I shall be claiming on behalf of my account of team agency.

First, some remarks about ontology. Anyone who suggests that groups can act and have preferences is liable to be accused of asserting the existence of mysterious collective entities. Ultimately, the objection runs, propositions about groups are reducible to propositions about individuals, because only individuals 'really' exist. Nothing in my argument will contravene this general principle of reducibility.[1] In my account, every proposition about team agency or team preference is reducible to *some* definite proposition about individuals. However, the individual-level proposition to which it reduces is not necessarily one that is recognized in the received theory of rational choice. Team agency, as I represent it, is not reducible to individual agency *as that is represented in rational choice theory.* Instead, my object is to amend the received theory of individual agency in such a way that team agency *becomes* a coherent concept within it.

Next, some remarks about rationality. I am engaged in an enterprise whose ultimate aim is to represent and codify forms of reasoning which people in fact use, perhaps informally or even unconsciously, when making decisions as collectives. In this paper, I take a first step towards this objective by setting up an ideal type of team agency which is a generalization of the standard model of rational individual choice. It is 'ideal' in the sense of being a simplification, an abstraction, a model; but it is not presented as a model of *ideal rationality.* It will be useful just to the extent that it captures salient features of real human reasoning.

Finally, something about politics and ethics. Many people are uneasy about concepts of group agency and group preference because of certain ways in which these concepts can be used in political and moral discourse. Liberals are often shocked by Hobbes's suggestion that a political community is effectively a single person which takes on the will of the sovereign, and that in joining the community, individuals give up their private judgements and take on the sovereign's. Rousseau's conception of the General Will inspires similar unease.[2] Even my

[1] Acceptance of this principle of reducibility, or of the related principle that properties of groups supervene on properties of individuals, is a common feature of the recent literature on group agency. See, for example, Gilbert (1989, pp. 427–36) and Tuomela (1995, Chapter 5).

[2] In arguing for a conception of group agency, Hollis (1998) appeals to Rousseau's idea that

https://doi.org/10.1017/S0266267100000213 Published online by Cambridge University Press

opening example of family preferences, which I took to be benign, has a darker side: the idea of the family as a single agent can all too easily be a cover for the exploitation of some family members by others.

I feel these senses of unease too. Nevertheless, to disagree with a political or moral argument is not the same thing as to claim that it makes no sense. By developing a concept of team agency, we may be able to make sense of what would otherwise be incomprehensible arguments about how political communities and families can act as wholes, subsuming the individual identities of their members. Perhaps, having made sense of those arguments, we find them deeply objectionable; but that is no excuse for pretending that they are incoherent if they are not.

2. THE FOOTBALLERS' PROBLEM

My starting point is a problem in the theory of games, which has been discussed sporadically over many years.[3] Since part of what is at issue is how this problem should be formulated, I begin with an informal example, which I call the Footballers' Problem. Suppose that A and B are two attacking players in a football team. A has the ball, but a defender is converging on him. B has more space, so A wants to pass the ball to him. There are two directions in which B could run so as to intercept a pass from A: call these *left* and *right*. Correspondingly, there are two points on the field, *left* and *right*, to which A could pass the ball to be picked up by B. There is no time for communication, or for one player to wait to see what the other does: each must simultaneously choose left or right. Suppose that the move to the right puts B into a slightly better position. Say that the probability that the pass will result in a goal is 10 per cent if both choose left and 11 per cent if both choose right. If one chooses right and the other left, the probability is zero. What should each player do?

The answer seems obvious: each should choose right. But paradoxically, this obvious answer cannot be generated by the theory of individual rationality, as used in game theory.

Before justifying this claim, let me explain why I think the Footballers' Problem may throw some light on the foundations of social cooperation. This problem is a simple model of a situation in which there is some objective (in this case, scoring a goal) which each of a set of

the transition from the state of nature to civil society is associated with a 'remarkable change in man', which allows everyone to submit to the General Will while still remaining as free as before. Hollis seems to be suggesting that this is a transition from individual to group agency. But even while invoking Rousseau as an ally, Hollis notes the 'deep ambiguity' in the concept of the General Will, and the danger that the resolution of this ambiguity might license a totalitarian state (p. 152).

[3] See Hodgson (1967), Gauthier (1975), Regan (1980), Sugden (1991, 1993), Bacharach (1993, 1999), and Hollis (1998).

https://doi.org/10.1017/S0266267100000213 Published online by Cambridge University Press

individuals wants to achieve, and which can most effectively be achieved if those individuals coordinate their actions. Thus, we might say, it represents a situation in which social cooperation would be useful or valuable. Some of the most primitive forms of human cooperation – cooperation between adults to rear children, cooperation between hunters to kill game, cooperation between fighters in defence or predation – might be modelled in a similar way.

To this claim, it might be objected that the Footballers' Problem is special in not allowing communication between the players. In human societies, it might be argued, most coordination is achieved through the use of language, and so a model which excludes communication is representing the problem of social cooperation without representing the main means by which it can be solved. I am not convinced. It is difficult to explain *how* language can help people to solve coordination problems (as it clearly can) without appealing to prior and tacit understandings. (For example, suppose that you and I have had a meal together and each of us is proposing to pay the whole bill. There may be an exchange of expressions of the form 'Let me', 'No, let me', 'But I insist', and so on; but that hardly helps. If instead I say 'You paid last time, so now it's my turn', that may lead to coordination; but if it does, it does so by drawing attention to and activating a pre-existing understanding about turn-taking.) More fundamentally, language itself is a form of social coordination which needs to be explained, and which might plausibly be understood as emerging out of more basic tacit understandings.[4]

Table 1 shows how the Footballers' Problem would be represented in game theory. (Notice that this matrix is a *model* of the Footballers' Problem, not the problem itself. The problem itself occurs between real footballers on a field. That the problem faced by the real footballers is essentially the same as that faced by the agents in the game shown in Table 1 is a theoretical claim, which we need not accept.) The game is specified in terms of the alternative strategies open to each player ('left' and 'right') and the utility that *each player* derives from each combination of strategies. 'Utility' for each player is interpreted as a cardinal representation of his preferences (the cardinality is usually justified by appeal to the axioms of expected utility theory, and by the claim that these axioms are principles of rational consistency). In this game, there is no conflict of interest: the players' preferences over outcomes coincide perfectly, each seeking to maximize the probability that a goal is scored.

Now consider what A, as a rational agent, should do. Clearly, if A expects B to choose 'right', he should choose 'right' too. But equally, if A expects B to choose 'left', he should choose 'left' too. According to expected utility theory, A should choose 'right' if he judges the

[4] For arguments to this effect, see Lewis (1969) and Skyrms (1996).

https://doi.org/10.1017/S0266267100000213 Published online by Cambridge University Press

TABLE 1. The Footballers' Problem as represented in conventional game theory

		player B	
		left	right
player A	left	10, 10	0, 0
	right	0, 0	11, 11

probability that B will play 'right' to be greater than 10/21; and he should choose 'left' if he judges that probability to be *less* than 10/21. So what it is rational for A to do depends on what B can be expected to do. In situations like this, game theory invokes the assumption that the rationality of the players is common knowledge between them. Thus, in order to form a rational belief about what B will do, A has to take account of the fact that B is himself rational, and so will choose whatever is rational from his own point of view. But B's problem is exactly symmetrical with A's: it is rational for B to choose 'right' if A can be expected to play 'right' with a probability of 10/21 or more, but rational for B to choose 'left' if A can be expected to play 'right' with a probability of less than 10/21. We have entered an infinite regress: what it is rational for a player in a situation like A's to do depends on what it is rational for a player in a situation like A's to do.

In cases like this, one might look for a Nash equilibrium – that is, a set of beliefs which, if held by both players, would be consistent with their rationally behaving in such a way that those beliefs were confirmed. This strategy immediately runs into the problem that, in this game, there are three different Nash equilibria (that both choose 'right' with probability one, that both choose 'left' with probability one, and that both choose 'right' with probability 10/21 and 'left' with probability 11/21). For game theorists, this problem is an instance of the more general problem of *equilibrium selection*. Various writers have proposed, as a principle of equilibrium selection, the criterion of *payoff dominance*: that if one Nash equilibrium gives each player more utility than does any other, then that equilibrium is the rational solution.[5] This principle may seem natural enough. But to say that it is natural is merely to restate, in more general terms, the intuition that it is rational for each footballer to choose 'right'. We still need a justification for it. If by 'rational solution' we mean 'combination of strategies, each of which is

[5] Gauthier (1975) proposes a variant of payoff dominance, which he calls the Principle of Coordination, as a solution concept for coordination games. Gauthier's principle is weaker than payoff dominance, in that it applies only to Nash equilibria which are also Pareto optimal. The most rigorous defence of payoff dominance as a principle of equilibrium selection is given by Harsanyi and Selten (1988).

https://doi.org/10.1017/S0266267100000213 Published online by Cambridge University Press

rational for the player who chooses it', we are no nearer that justification. And if we mean something else by 'rational solution', what is that something else?

There is a more fundamental objection to the equilibrium selection approach. Within the logic of individual rationality, the fact that a particular combination of strategies is a Nash equilibrium (even a unique Nash equilibrium) gives neither player a *reason* to choose to play his part in that equilibrium. All we can say is that, *if* it were the case that each player expected every other player to play his part in a certain (strict) Nash equilibrium, *then* each player would have reason to do the same. And that provides us with no escape route from the infinite regress of reasons. The claim that it is rational for players to play their parts in some Nash equilibrium is sometimes justified by appeal to a presupposition that every game has a unique rational 'solution', that is, a prescribed strategy for each player, which is known to all rational agents. It is easy to see that if such a solution is common knowledge among rational players, it must be a Nash equilibrium. But that presupposition seems ungrounded.[6]

So, if we assume only that A and B are rational in the standard sense, and that their being so rational is common knowledge between them, we are not entitled to conclude that each will choose 'right'. Yet, intuitively, it seems obvious that 'right' is the rational choice for each of the real footballers. Apparently, something is missing from the standard theory of rational choice. But what?

3. TEAM-DIRECTED REASONING

In this paper, I am concerned with one particular departure from the standard theory, which I think is crucial for a resolution of the Footballers Problem: the idea of *team-directed reasoning*. Over the course of the paper, I shall say something about the history of this idea, and about the different ways in which this idea has been formulated.[7] But first I want to outline the basic idea, without getting bogged down in detail.

The idea is that, in relation to a specific decision problem, an individual may conceive of herself as a member of a group or team, and conceive of the decision problem, not as a problem *for her* but as a problem *for the team*. In other words, the individual frames the problem,

[6] For more on the assumption that games have uniquely rational solutions, and on why that assumption is ungrounded, see Sugden (1991).

[7] I shall focus on those conceptions of 'we-thinking' that are most closely connected with the theory of rational choice. There is a related literature in the theory of intentions which analyses the concept of collective intentions: see, for example, Searle (1990) and Bratman (1993).

https://doi.org/10.1017/S0266267100000213 Published online by Cambridge University Press

not as 'What should *I* do?', but as 'What should *we* do?' For someone who has framed a decision problem in this way, relevant advice about what to do has to be addressed, not to the individual alone, or even to each member of the team independently, but to each member of the team *as a member of the team*.

Take the case of the footballers. Imagine that the situation I have described is being analysed in a coaching session, at which A and B are both present. The coach's recommendation is that if this situation arises in actual play, A and B should each choose 'right'. Suppose A asks why these moves should be made. The coach's answer is that by each choosing 'right', A and B together maximize the probability that a goal will be scored. Now imagine that A complains that this is not an adequate answer. Of course, he wants to maximize the probability of scoring a goal, and of course he can see that this probability will be maximized if he and B both choose 'right'. But, he objects, this is not an adequate reason *for him* to choose 'right'. He needs to be convinced that, by *his* choosing 'right', *he* will maximize the probability that a goal is scored, given well-grounded beliefs about what other players will do. And to be convinced of this, he needs to be convinced that B will choose 'right' too. But just as A has not yet been given an adequate reason for choosing 'right', neither has B. Within the framework of the received theory of rational choice, this objection is entirely legitimate. But, seen outside that framework, it is obtuse.

It is obtuse, surely, because it fails to understand what a coaching session for a team is all about. The coach is not addressing A and B as separate individuals; he is addressing them collectively, as members of a team. The logic of his recommendation can be put like this: given that the team's objective is to maximize the probability that a goal is scored, the best combination of moves *by the team* is that A and B both choose 'right'. Therefore, A and B should each choose 'right'. The 'therefore' is self-explanatory to anyone who understands what it is to be a member of a team: if this combination of moves is best for the team, then this is what the team should do.

If we can understand the coaching session, we can also understand a means by which A and B can coordinate their actions on the field, even in positions for which no specific prior plans have been made. Suppose A and B face the problem of choosing between 'left' and 'right' in the course of a game, without having been given any specific instructions about what to do in this position. Each player, let us suppose, can work out which combination of moves by the two of them would be best for the team. So if their ability to reason in this way is common knowledge between them, it is almost as if they had been present together at a coaching session at which 'right' had been recommended. Each chooses whichever action is his part of the combination of actions that is best for

https://doi.org/10.1017/S0266267100000213 Published online by Cambridge University Press

the team. This is team-directed reasoning. Notice that this kind of reasoning is carried out *by individuals*, but by individuals who take themselves to be members of teams. That is why, at this stage, I prefer to speak of *team-directed* reasoning rather than *team* reasoning.

Any theory of team-directed reasoning seems to confront two difficult problems, which for short I shall call the *existence problem* and the *objectives problem*. The existence problem is to specify which teams exist. When I say that a team 'exists', I mean that its members recognize that they are members of it, and that each member's actions (in the relevant domain) are determined by team-directed reasoning. In terms of the example: how does each footballer come to know that he is a member of a particular team, and to know that his actions on the football pitch should be guided by team-directed reasoning rather than by individual rationality? The objectives problem is to specify, for any given team, what its objectives are. How does each footballer come to know that the object of the team is to score goals? If these questions seem trivial in the case of football, that is only because football is such a simple model of team-directed reasoning. I suspect that the difficulty of giving satisfactory general answers to these questions has been a main reason for the failure of the idea of team-directed reasoning to gain acceptance among theorists of rational choice.

4. COOPERATIVE UTILITARIANISM

One way of answering these questions, first suggested by D. W. Hodgson (1967) and Donald Regan (1980), is closely related to rule utilitarianism. The game-theoretic representation of the Footballers' Problem, as shown in Table 1, can be used to illustrate a general problem for act utilitarianism. Suppose that utility is additive across individuals, and that A and B are each motivated to maximize the sum of their utilities. Suppose that each acts as an independent, rational agent in pursuit of this utilitarian objective, and that it is common knowledge that each is so motivated. Then they are playing a game which is just like the game in Table 1, except that all the payoffs have been multiplied by two; and their reasoning runs into just the same infinite regress. In other words: for A and B jointly to achieve the utilitarian objective of maximizing the sum of their utilities, it is not sufficient that each acts independently on this objective, even if their so acting is common knowledge between them. However, this objective *will* be achieved if A and B act on the principles of rule utilitarianism. According to the simplest version of rule utilitarianism, each individual should consider the class of general *rules* of behaviour which could be followed by everyone, and then follow whichever of those rules would, if followed by everyone, maximize total utility. In the Footballers' Problem, the only

https://doi.org/10.1017/S0266267100000213 Published online by Cambridge University Press

rules that maximize total utility are those that prescribe 'right' to each player; so rule utilitarian players would choose 'right'.

It might be objected that (from a utilitarian perspective) the rationality of acting on the rule utilitarian principle is conditional on its being common knowledge that everyone is so motivated. For example, suppose that player A has good reason to believe that B will in fact choose 'left'. Then, as a good utilitarian, shouldn't A choose 'left' too, so as to bring about the best consequence? One way of dealing with this problem is to revise the rule utilitarian principle so that it applies only when, within some group of individuals, there is common knowledge that everyone in the group accepts the principle. This is the idea behind Regan's (1980) principle of *cooperative utilitarianism*. Regan sums up this principle like this: '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' (p. 124). Roughly, the principle distinguishes between 'cooperators' (who are willing to act on a principle of cooperation) and 'non-cooperators' (who are not). Each agent is required to be a cooperator, and to follow the rule which, if followed by all cooperators, would maximize total utility, given the expected behaviour of the non-cooperators.

Regan's proposal can be understood as involving a form of team-directed reasoning. It gives straightforward answers to the existence and objectives problems. The existence problem sets the question of which teams exist (or should exist). Regan's answer is 'the largest team possible'. In an ideal world, everyone would be a cooperative utilitarian, and they would all be members of a single team. In the world as it is, the set of all cooperative utilitarians should constitute themselves as a team. The objectives problem sets the question of what the objective of that team should be. Regan's answer is 'to produce the best possible consequences'. Regan (1980, pp. 1–3) deliberately avoids getting involved in a discussion about how the relative goodness of consequences is assessed, but it seems clear that what he has in mind is goodness in an impersonal, utilitarian sense. The proper object of the cooperators is neither to promote their own good as individuals, nor to promote the collective good of themselves as a team, but to promote the good of the world.

However, the straightforwardness of these answers is a reflection of the specificity of Regan's proposal. Regan's objective is not – as mine is – to understand the general modes of reasoning through which people in fact coordinate their actions. Instead, it is to propose how such problems *ought* to be solved *by utilitarians*. Regan's approach prescribes the maximization of the overall goodness of consequences as the uniquely rational objective for every individual and for every group of individuals. For a utilitarian, that prescription is axiomatic. But a *general*

https://doi.org/10.1017/S0266267100000213 Published online by Cambridge University Press

explanation of how people solve coordination problems cannot credibly assume that most people are utilitarians in such an austerely rationalistic sense. For many of us, it is not self-evident that either rationality or morality requires us – either individually, or as collectives delimited by kinship, locality, friendship, workplace and so on – always to forgo our own interests when this would lead to better overall consequences for the world as a whole.

Further, Regan treats the problem of measuring overall goodness as external to his proposal. Thus, in response to the question of what the objective of his cooperative utilitarian team should be, Regan's answer is formal rather than substantive. If cooperative utilitarianism is to be a practical proposition, there has to be a unique utilitarian ranking of consequences; and this has to be accessible, not merely in principle to scientific enquiry, but in practice to all cooperators. The problem of making operational the utilitarian conception of a ranking of consequences by overall goodness has been one of the central projects of welfare economics. Many economists, myself included, would say that it has never been solved.[8] This problem is an obstacle, not only to the development of Regan's utilitarian proposal, but also to the development of any theory of team agency in which teams have objectives and in which the objectives of a team are an aggregation of the preferences of its members.

5. THE CONTRACTARIAN APPROACH

Some of the objections that can be raised against Regan's utilitarian approach can be avoided by adopting a *contractarian* perspective. In this perspective, the purpose of a cooperative venture is not to promote the aggregate welfare of any group of people, but to promote the separate ends of the individuals who take part in it. A cooperative venture is initiated when a set of individuals *agree* to constitute themselves as a team, and *agree* on a common objective which each of them will then play her part in achieving. Thus, in the case of the footballers, we might say that each footballer has a preference, as an individual, that the probability of scoring a goal is maximized, and that in choosing to play in a team they all agree to take this as their common objective. Of course, the problem of the infinite regress of reasoning will not be solved if each

[8] It is now conventional to interpret 'utility' as a representation of preferences, and to treat preferences either as unexplained motivating factors which are revealed in choices, or as whatever individuals take to be all-things-considered reasons for choosing one thing rather than another. Following Pareto (1906/1972), economists have adopted this interpretation as a means of making utility into an operational concept. But given this interpretation of utility, it is questionable whether interpersonal comparisons of utility are meaningful at all. For more on this, see Scanlon (1991).

https://doi.org/10.1017/S0266267100000213 Published online by Cambridge University Press

footballer simply pursues this objective *as an individual agent*; each of them has to pursue it *as a team-directed reasoner*. That is, each must play his part in the combination of actions that is best calculated to achieve the common objective.

The idea that plural agents – the kinds of agent that I have called teams – are formed by agreement can be traced back to Hobbes's *Leviathan* (1651/1962).[9] As Margaret Gilbert (1989, p. 303) notices, the famous frontispiece of *Leviathan* represents the idea of a collective agent which is made up of individuals, each of whom acts as if part of a single body; and Hobbes's core argument is that, starting from a state of nature, it is rational for each person to join an agreement to create such a body. A similar emphasis on agreement can be found in the work of Raimo Tuomela. Tuomela proposes a conception of 'joint action' by a plural agent, and requires that for there to be genuinely joint action, the participants in that action must have agreed to perform it: 'A central thesis to be defended is that the performance of a joint action, X ... requires that the participants have explicitly or implicitly agreed to perform action X' (1995, p. 73).

If teams are formed by agreement among individuals, collective purposes are derivative from the purposes of individuals. Given this starting point, it may seem natural to suppose that plural agency requires some mechanism for aggregating individuals' preferences into a collective objective; a specification of such a mechanism would then constitute an answer to the objectives problem. Tuomela sometimes[10] seems to argue along exactly these lines. For a group of individuals to be an agent, he claims, it must have an 'authority system'; an authority system is to be understood as a procedure (or 'transformation function') for 'forming a group will ... on the basis of [the group members'] individual wills' (1995, pp. 185–191). Tuomela suggests that the transformation function is analogous with the aggregation rules considered in social choice theory (pp. 296–301). However, he does not specify this aggregation mechanism in any detail. The difficulty of coming up with a credible specification is a serious problem for this form of analysis.

Other writers who have followed a broadly contractarian approach have contented themselves with the claim that we can at least be confident in assuming that the aggregation rule will satisfy the Pareto principle. For example, this assumption is made by David Copp (1980, p. 604). It is implicit in the principle, which I proposed in Sugden (1993), that each member of a team looks for a unique rule which, if followed by

[9] For more on Hobbes's conception of collective agency, see Copp (1980).

[10] Although Tuomela's discussion of social choice theory is concerned with mechanisms which generate collective objectives by aggregating individuals' preferences, other passages in his book allow that collective objectives might be independent of individuals' preferences (e.g. 1995, p. 120).

https://doi.org/10.1017/S0266267100000213 Published online by Cambridge University Press

all members, would yield the best possible results for all members. Susan Hurley's (1989) analysis of 'collective agency' focuses on games in which there is a best rule in this sense, and argues that collective agency will lead each individual to follow that rule.[11] In general, unfortunately, there is no guarantee that such a best rule exists.

But if the aim is to explain plural agency *as it is*, rather than to prescribe how it ought to be, even the Pareto principle may be too restrictive. It does seem possible in fact for people to construe themselves as members of a plural agent whose collective objective is contrary to the unanimous preferences of those members.

Consider the training given to conscripts joining an army. I take it that one of the normal aims of such training is to inculcate a sense of plural agency, but not one that is supposed to be founded on consent. The collective objective that the conscripts are trained to pursue may have little or no connection with the conscripts' preferences as individuals, and this fact may be common knowledge among them.

Here is another example. Common experience suggests that plural agency can come into existence gradually, without anything explicit being said. (Think of how, in the case of friendships and romantic relationships, two people become a 'we'.) If that is right, then what the members of a plural agent take to be its objectives at any given time may be the product of its history, and may be influenced by shared conceptions of salience. Imagine a group of long-standing friends who have always pursued their friendship in a particular context – say, by meeting in pubs and sampling different beers. Collectively, they act on the objective of seeking out curious pubs and little-known beers. Perhaps, over time, each of them privately loses interest in pubs and beers, but the friendship continues. Because the group's existence depends on a body of unspoken assumptions, it is difficult for anyone to raise the question of whether the group should take up a new pursuit. Even if the group's activities gradually shift to reflect changes in its members' preferences, this process may be subject to a great deal of inertia. There is, then, nothing inherently inconsistent in the possibility that every member of the group has an individual preference for y over x

[11] Hurley (1989, pp. 136–70) argues that if 'it is a good thing for [a certain form of] collective agency to exist', then it is rational for individuals to participate in that agency. Thus, her claim is not merely that, *if* the players of a game take themselves to be a team, *then* the team's objective should respect the Pareto principle. Hurley seems to be endorsing the stronger principle that, whenever a set of individuals share a common preference that can most effectively be achieved by team-directed reasoning, it is rational for each of them to take themselves to be members of the necessary team. Thus, for example, it is irrational for either player to defect in the Prisoner's Dilemma. This argument, I suggest, rests on too restrictive a conception of rationality. For more on this, see Sugden (1993).

https://doi.org/10.1017/S0266267100000213 Published online by Cambridge University Press

(say, each prefers wine bars to pubs) while the group acts on an objective that ranks x above y.

6. OBLIGATIONS TO TEAMS

Gilbert (1989) proposes a concept of a 'plural subject' which rests on something similar to, but not quite equivalent to, agreement. In order for a joint action to be performed by a plural subject, it is necessary that 'the participants express to each other willingness to be part of a plural subject of a certain goal' (1989, p. 17; see also pp. 182–4). In other words, plural agents are created out of individual ones when those individuals openly declare their willingness to participate in that creation. The 'openness' here is intended to signify that everyone's having expressed willingness to participate is common knowledge in roughly the sense of David Lewis (1969, pp. 52–60): the nature of each person's expression is such that everyone has reason to believe that it has been made, everyone has reason to believe that everyone has reason to believe that it has been made, and so on.

Gilbert does not require formal acts of agreement. She argues that for plural agency to be activated, it is sufficient that the constituent individuals openly go along with 'We …' statements which presuppose such agency. Thus, as in the case of gradually emerging friendships and romantic relationships, 'we-ness' can be activated by subtle but significant exchanges of words or signals which express willingness to create a plural subject.[12] Similarly, a young teenager might be understood to be a member of her family by virtue of her openly going along with statements which take the family to be a 'we' (Gilbert, 1989, pp. 171–2). Gilbert also allows the understandings which underlie plural agency to be elicited under coercion (1999, p. 254). And she does not presuppose that the preferences, beliefs or attitudes of a plural subject are reducible to the preferences, beliefs or attitudes that the members hold as individuals. Thus, she would have no difficulty with the examples I presented at the end of Section 5.

However, her acceptance of such examples comes at a cost. She construes plural agency as imposing *obligations* on individuals. Participation in a plural agent is taken to involve, as a matter of conceptual necessity, the acceptance of certain commitments. Essentially, the commitment is to uphold – at least, in situations in which the agency of the plural subject is active – whatever preferences, beliefs or attitudes that plural subject has taken on (Gilbert, 1989, p. 162). Tuomela's account of joint action has similar implications (1995, pp. 73–6). This approach gives rise to tensions, which are illustrated by the relationships between

[12] Tuomela's concept of 'implicit agreement' has similar connotations.

https://doi.org/10.1017/S0266267100000213 Published online by Cambridge University Press

the teenager and her family, and between the conscript and his unit. It does seem right to say that the teenager can take herself to be a member of her family, and that the conscript can take himself to be a member of his unit, without there having been any prior acts of agreement. But at the same time, it is hard not to feel uneasy about the assertion that these individuals have *obligations* to participate in the plural agency of their groups. The more we weaken the concept of agreement so as to accommodate the range of types of plural agency that exist in the world, the less plausible it becomes to claim that the residual notion of 'agreement' generates obligations.

Why do Gilbert and Tuomela need their concepts of obligation? The answer, I think, is that they are looking for a criterion to identify cases in which a person has reason to act as a member of a team. If a person is capable of reasoning *either* as an individual, pursuing her private preferences, *or* as a team member, doing her part in the team's pursuit of its collective objective, it may seem natural to look for a meta-principle which determines which kind of reasoning should be used in any given situation. Such a meta-principle, were it to exist, might explain how each member of a team can have confidence in the other members' playing their parts. To see why, imagine that such a principle does exist. It shows that, in specific circumstances, each team member has reason to act as such. Thus, whenever it is common knowledge that those circumstances hold, each team member has reason to believe that each other member has reason to act as a team member; each has reason to believe that each other has reason to believe that each other has reason to act as a team member; and so on. Any meta-principle that is capable of activating sufficient reasons for action is thereby capable of generating rational obligations – that is, the obligation to act on sufficient reasons.

7. A THEORY OF TEAM AGENCY

At the end of Section 3, I set out two apparent problems for a theory of team agency: the existence problem and the objectives problem. I have discussed a range of attempts to resolve those problems, and have argued that the answers they generate are too restrictive. I shall now propose a radical method of avoiding the difficulties encountered in trying to solve the two problems.[13]

The difficulties have arisen because contributors to the literature have made one or other (or sometimes both) of two presuppositions. The first of these will seem natural to anyone who has learned to think

[13] To my knowledge, the only decision theorist to have proposed taking this step is Bacharach (1999). The present paper is complementary with Bacharach's. His paper presents a much more formal and general analysis of team agency than I do here, but offers less supporting philosophical argument.

https://doi.org/10.1017/S0266267100000213 Published online by Cambridge University Press

within the framework of conventional rational choice theory. It is that an account of collective decision-making is incomplete unless it is grounded on propositions about individuals' preferences. Or, more precisely: unless it is grounded on propositions about those preferences – I shall call them *individual-directed preferences* – that govern the choices that people make when they act as individuals. Thus, it seems, we have not fully explained the concept of a team objective unless we have explained how that objective can be constructed, given information about the individual-directed preferences of the team members. Similarly, we have not fully explained the idea that a person may act as a member of a team unless we have explained his choosing to act in this way as an implication of his (and perhaps of other people's) individual-directed preferences. This presupposition imposes severe constraints on possible answers to the existence and objectives problems.

The second presupposition will perhaps seem more natural to philosophers than to economists. It is that a theory of rational choice should be grounded on an account of *good reasons*. Thus, a theory of team agency should explain when it is and is not rational for individuals to take themselves to be members of teams. The implication is that the existence and objectives problems have to be solved by appeal to good reasons: merely empirical 'solutions' do not count.

My purpose in this paper is to argue that these presuppositions are unwarranted. That is, a theory of team agency may legitimately take both the existence of teams and their objectives as matters for empirical (rather than rational) explanation. There is no requirement that the existence and objectives of teams must be explained either in terms of individual-directed preferences or in terms of good reasons.

I shall now set out just such a theory. As a first step, I shall outline what the components of the theory are, and how they fit together. Then I shall present the components in more detail. Finally, I shall defend the theory.

The first component of the theory is an account of the structure of *team-directed reasoning*. This is a mode of reasoning which a person might use when choosing what to do when interacting with other people. My claim is that team-directed reasoning is a reasonably adequate model of a certain kind of reasoning that people in fact use *when they take themselves to be acting as members of teams*. One of the salient features of this mode of reasoning is that it generates recommendations for action that are not conditional on the actor's beliefs about what the other individuals will do. In this respect, team-directed reasoning is quite different from the strategic reasoning that is modelled in conventional game theory. The lack of conditionality in the conclusions of team-directed reasoning allows an escape from the infinite regress which traps strategic reasoning in cases such as the Footballers' Problem.

https://doi.org/10.1017/S0266267100000213 Published online by Cambridge University Press

Team-directed reasoning is carried out *by individual agents.* I intend to *use* this theoretical concept to model the reasoning of people who take themselves to be members of teams, but 'taking oneself to be a member of a team' is not defined within the formal structure of team-directed reasoning. In principle, then, an individual might engage in team-directed reasoning without believing that anyone else was doing so. However, my concern is with people who *do* take themselves to be members of teams. Thus, my theory needs its second component: an account of what it is to take oneself to be a member of a team. Since what is being accounted for is *an individual's perception* of his place in a team, the analysis is still of deliberations which are carried out by a single individual; but now the individual's beliefs about other people matter. To take oneself to be a member of a team, one has to hold certain beliefs about the other members of that team.

The final component of the theory is an account of what it is for a team to exist. It is only at this stage that the analysis brings together the reasoning of different individuals. Roughly, my account is that a team exists to the extent that its members take themselves to be members of it. It is at this stage that we may be able to speak of a team as an agent in its own right, and as having preferences.

These preliminaries over, I present the theory itself. Consider a set of individuals, $A_1, \ldots, A_n$, who interact in some way that can be described by a *game form*. A game form can be defined in the following way: For each individual A_i there is a set S_i of alternative *strategies*, from which that individual must choose one and only one. For every possible *array* of chosen strategies (one chosen by each individual), there is an *outcome*. An outcome is to be interpreted simply as a description of what would happen if the relevant array of strategies was chosen. No information about preferences or utilities is incorporated into the description. (To include such information would be incoherent, since preferences are to be understood as preferences *over outcomes*, and utility is to be understood as a representation of preference.)

If to each array of strategies we were to assign, in place of an outcome, an *array* of utility indices, one for each individual, we would have a normal-form game (such as the representation of the Footballers' Problem in Table 1). But suppose instead that we define a *team-directed utility* function t(.) which, to each outcome x, assigns a *single* utility index t(x), to be called 'team-directed utility'; this is to be interpreted as a representation of *team-directed preferences* over the relevant outcomes.[14]

[14] For ease of exposition, and to maintain as much similarity as possible with conventional game theory, I am assuming here that team-directed preferences satisfy the axioms of expected utility theory and so can be represented by a cardinal utility function. But this assumption is not fundamental to my proposal.

https://doi.org/10.1017/S0266267100000213 Published online by Cambridge University Press

TABLE 2. The Footballers' Problem as a team-directed decision problem

		player B	
		left	right
player A	left	10	0
	right	0	11

For the moment, I suspend the question of what 'team-directed preference' means. The entity that is defined by the set of players, the sets of strategies and the indices assigned by a team-directed utility function will be called a *team-directed decision problem*. Table 2 represents the Footballers' Problem as a team-directed decision problem. There are still two players, each of whom has to make a separate decision between 'left' and 'right', but now there is only one scale of preference on which outcomes are ranked: the scale of team-directed preferences.

Let G be any game form. Let t(.) be a team-directed utility function. Now suppose there exists an array of strategies (s_1*, …, s_n*) such that each s_j* is an element of S_j and such that, in terms of t(.), the team-directed utility generated by this combination is strictly greater than that generated by any other combination. Then each A_i engages in *team-directed reasoning* with respect to G and t(.) if she chooses s_i* in virtue of the fact that (s_1*, …, s_n*) uniquely maximizes team-directed utility.[15] Notice that, although team-directed reasoning is carried out by individuals, it is not an instance of individual reasoning as that is represented in the standard theory of rational choice. The two kinds of reasoning are different in structure. In the standard theory, the individual appraises alternative *actions by her* in relation to some objective (her preferences), given her beliefs about the actions that other individuals will choose. An individual who engages in team-directed reasoning appraises alternative *arrays of actions by members of the team* in relation to some objective (team-directed preferences, as represented by t(.)).[16]

[15] If two or more different combinations of strategies yield exactly the same utility for the team, this decision rule fails to determine what each individual should do. Bacharach (1993) and Sugden (1995) suggest some ways in which this difficulty can be overcome if individuals recognize a sufficiently rich 'framing' or 'labelling' of their decision problem.

[16] Bacharach (1999) proposes an account of team-directed reasoning in which the mode of reasoning I have just described is a special case. One important feature of Bacharach's analysis is that individual members of a team are not necessarily aware that their interaction calls for team thinking. (For example, suppose that A, B and C are the members of a team; A and B know that the objectives of the team require a meeting of the three of them, but C does not know this.) On Bacharach's analysis, the problem to be solved by team-directed reasoning is to find the array of strategies which, if followed by all 'aware' members, contributes most to the team's objectives, given the behaviour of the

https://doi.org/10.1017/S0266267100000213 Published online by Cambridge University Press

Team-directed reasoning, as I have presented it so far, is a purely formal construct. It describes the structure of a mode of reasoning that an individual might use to determine what to do when interacting with others. Such reasoning could conceivably be used by one individual without any expectation that other individuals were reasoning in a similar way. Indeed, if t(.) is interpreted as a measure of the common good, unilateral team-directed reasoning by one individual could be interpreted in Kantian terms, as his acting on a maxim that he could will to be a general law.

However, the idea of acting *as a member of a team* seems to require more than unilateral team-directed reasoning. Take the case of the footballers. On my analysis, each player's reasoning takes the form: 'We are both members of the team; the team's objective is to win; therefore I should choose ''right''. Team-directed reasoning explains the logic of the final 'therefore'. But I still need to explain what it is for a player to believe that he is a member of a team.

To act as a member of a team is to do one's part in the array of actions which, taken together, best achieve the team's objectives. But to construe one's own action as a part of this larger whole, one must have some confidence that the other parts will come about. More fundamentally, to construe oneself as a member of a team, one must have some confidence that the other members of that team construe themselves as members too. Within my model, what is expected of a person by virtue of her being a member of a team is that she engages in team-directed reasoning. Thus, to take oneself to be a member of a team is to engage in such reasoning oneself, while holding certain beliefs about the use of such reasoning by others.

At this point, I need more definitions. I shall say that an individual A_i has *first-order team confidence* with respect to some game form G and some team-directed utility function t(.) if she believes that every other individual (that is, every other individual in the set $\{A_1, \ldots, An\}$) engages in team-directed reasoning with respect to G and t(.). To have first-order team confidence is to be confident that the others will do their part in whatever array of actions is uniquely optimal with respect to t(.). A_i has *second-order team confidence* with respect to G and t(.) if she believes that every other individual has first-order team confidence with respect to G and t(.); and so on. If for every finite positive integer m, an individual has mth-order team confidence with respect to G and t(.), then she has *full team confidence* with respect to G and t(.). Thus, for A_i to have full

'unaware' members. Bacharach also considers games in which some but not all players are members of the same team. I leave these complications aside to focus on what I see as the central philosophical issues.

https://doi.org/10.1017/S0266267100000213 Published online by Cambridge University Press

team confidence is for her to believe that, as far as everyone else is concerned, everything is in place for team agency to be operative.

Accordingly, I define team agency as an ideal type in the following way: Consider any game form G for individuals A_1, …., A_n, and any team-directed utility function t(.). Suppose the following two conditions are satisfied. First, each individual A_i engages in team-directed reasoning with respect to G and t(.). Second, each individual A_i has full team confidence with respect to G and t(.). Then *team agency* exists with respect to the *team* {A_1, …, A_n}, the game form G, and the *team utility function* t(.); A_1, …, A_n are *members* of the team; t(.) represents *team preferences*; and the team engages in *team reasoning*.

Notice the transition here from 'team-directed' concepts to 'team' concepts. Team-directed reasoning is something that one individual can engage in, independently of any others. Similarly, team-directed preferences are preferences that can be held by any individual, independently of any others. In contrast, team reasoning, team preferences and team agency are properties of *a set of* individuals, and require a network of common beliefs.

Notice also that my definition of team agency includes the special case in which n = 1. In this case, there is only one relevant individual, A_1. The game form reduces to the choice problem, faced by this individual alone, of choosing one strategy from the set S_1, in the knowledge that each strategy leads to a determinate outcome. Team-directed reasoning by this individual simply amounts to his choosing the strategy (provided there is one) which leads to the most-preferred outcome, as assessed by the preferences of this one-person team. Since there are no relevant individuals other than A_1, the requirement of full team confidence is vacuous. Thus, in the one-individual case (and leaving aside the special problems created by indifference), team agency reduces to individual agency, as that is represented in the standard theory of rational choice; and team preferences reduce to the individual-directed preferences that, in the standard theory, lie behind choices.

Nothing in this account of team agency purports to tell people when they *ought* – whether morally or rationally – to act as members of teams. Consider some individual A_i. Suppose he has full team confidence with respect to some G and t(.). My analysis prescribes what he should rationally choose *if* he takes himself to be acting as a member of the team – that is, *if* he engages in team-directed reasoning. But whether he does so take himself remains open. For example, suppose that Bill is a conscript in a disciplined army unit. Everyone in the unit has been trained to act as a team member with respect to certain military objectives, and to rely on the others to do the same. Bill is sure that all the other members of the unit will act in this way, and that they will rely on him to do so too. Can we say that rationality requires Bill to act in the

https://doi.org/10.1017/S0266267100000213 Published online by Cambridge University Press

same way? Surely not. Without being irrational, he may choose to ignore the team and act on his own individual-directed preferences.

This signals a difference between team agency, as I have defined it, and the analyses discussed in Sections 5 and 6. My account of team agency does not require that the individuals who participate in it agree to do so, or openly express their willingness to do so. What is required instead is that there is *confidence* among the members of the team that each of them will engage in team-directed reasoning with respect to a common set of team preferences. For brevity, I shall now contrast team confidence with agreement, but what I shall say about agreement applies also to 'open expression of willingness'.

Agreement is one way in which team confidence might be generated. Whether a particular act of agreement actually generates the confidence necessary for team agency is an empirical question. Nevertheless, one might claim, as a conceptual truth, that agreements generate certain kinds of *intentions*: a person who has sincerely agreed to do something has thereby formed an intention to do it. And intentions might be construed as entailing reasons for action. But there are other ways of generating team confidence which do not, as a matter of conceptual necessity, generate reasons for action. Mutual confidence in the use of a particular mode of reasoning can be brought about merely by mutual observation of behaviour, combined with inferences about how other people reason. (For example, as I drive along the roads of Britain, I observe the behaviour of other drivers; I discover regularities in this behaviour; I make inferences about the reasoning which underlies it; I use these inferences to predict how drivers whom I have never met before will behave; and I stake my life on the truth of these predictions. Other drivers do the same about me.) Mutual confidence, then, is ultimately an empirical and not a rational concept.

I can no longer postpone the question: What do I mean by 'team preference'? This question can be answered at two levels: the level of the team, and the level of the team members.

At the level of the team, team preference is a ranking of outcomes which is revealed in the team's decisions. To see this, suppose that for some team $\{A_1, \ldots, A_n\}$, team agency exists with respect to some game form G and some team utility function t(.). This implies that every individual A_i has the same team-directed preferences over the possible outcomes (that is, the team preferences that are represented by t(.)), and engages in team-directed reasoning, relative to this common set of team-directed preferences. Thus (except in the case in which team-directed reasoning does not prescribe a unique combination of strategies), the combined effect of the choices of the members of the team will be to bring about the outcome which, of those that are feasible, is most highly ranked in terms of the team's preferences. So it is *as if* the team were a

https://doi.org/10.1017/S0266267100000213 Published online by Cambridge University Press

single agent, choosing among feasible outcomes according to *its* preferences. In this sense, it is meaningful to talk about the team as an agent in its own right. There is nothing ontologically mysterious about this. Team agency, so understood, supervenes on a particular kind of individual agency, namely the agency of individuals who engage in team-directed reasoning and who have certain sets of beliefs and of team-directed preferences in common.

At the level of the team members, a team preference is a team-directed preference which is common to all those members, and which governs the team-directed reasoning of each of them. So the relevant question is: What do I mean by team-directed preference? For an individual who engages in team-directed reasoning, her team-directed preferences constitute a ranking of outcomes which she uses, by way of that reasoning, to determine which strategy she chooses. That is just about all that needs to be said.

At first sight, this interpretation of team-directed preferences may seem circular: I have defined team-directed reasoning in terms of team-directed preferences, and then I have said nothing more than that a person's team-directed preferences are rankings which guide her team-directed reasoning. But, as I shall argue in Section 8, this is no more circular than is the standard interpretation of preferences in the received theory of rational choice.

8. PREFERENCE

So what is that standard interpretation? Within the received theory, to say that an individual prefers some state of affairs x to another state of affairs y is to describe a mental state of that person, which disposes her to choose actions which lead to x rather than actions which lead to y. (The notion of *disposition* is significant, because one can have preferences over options that are not in fact feasible. Thus, though I cannot afford to buy either a Scottish island or a Premiership football team, I can say that I would prefer to have the island – meaning that, were I able to choose between them, that is what I would take.)

On some revealed-preference accounts, preference is nothing more than a disposition that a person may come to have, for whatever reason or for none, which *prompts her* to choose actions of one kind rather than actions of another. However, such an interpretation of preference seems not to acknowledge the sense in which the theory of rational choice is a theory *of reasoning*.[17] It would be more faithful to the practice of rational choice theory to say that a person's preferences are whatever she takes to be choice-relevant reasons, all things considered.

[17] David Gauthier and Chris Morris persuaded me of the validity of this objection to interpreting preferences merely as dispositions.

https://doi.org/10.1017/S0266267100000213 Published online by Cambridge University Press

In practical economic applications of rational choice theory, a distinction is made between the domain in which preferences are defined (which I shall call the domain of *outcomes*) and the domain of decision-making in which economic explanation operates (the domain of *actions*). Individuals are modelled as instrumentally rational to the extent that they choose those actions that are best calculated to achieve preferred outcomes. For example, in consumer theory, an individual's preferences are defined over the alternative bundles of goods that she might ultimately consume, while the theory aims to explain her actions in trading goods and money in markets. The individual's actions within markets (say, buying a jar of coffee for £1.50 at Tesco rather than for £1.55 at Sainsbury) are explained instrumentally, as the means of achieving the most preferred bundle of consumption goods. Similarly, in game theory, each individual's preferences are defined over the alternative outcomes of a game; which outcome occurs is determined by the combination of strategy choices made by all the players in the game. The individual's action within the game – her choice of strategy – is explained instrumentally, as a means to achieving the most preferred outcome, given her beliefs about the strategy choices of the other players.

If we are to understand these connections between preferences and actions as the results of *reasoning*, we have to suppose that individuals take their preferences to be reasons for them to act in the corresponding ways. But the modern theory of rational choice does not try to explain why people have the preferences they do: the chain of instrumental explanation stops at preferences. Preferences, then, just are whatever people take to be reasons for choosing one action rather than another. The theory does not endorse these as *good* reasons.

Notice that the standard account of preferences depends *on a particular theory* about the relationship between preference and action. This theory, which I shall call the theory of *individual-directed reasoning*, is that an individual chooses whichever action, among those that are feasible to her, leads to the outcome she most prefers, given her beliefs about what other people will do. The theory provides an essential link between the idea of a preference as (what the agent takes to be) an all-things-considered, choice-relevant reason and choice itself. Thus, when we say that a person prefers outcome x to outcome y, we mean that she takes herself to have reason to make those choices among actions that, *according to the theory*, are implied by a preference for x over y (that is, those choices which she believes will bring about x rather than y). To say this is not to lapse into circularity: given the theory, and given a particular line of demarcation between outcomes and actions (such as that provided by consumer theory, or that provided by normal-form game theory), the concept of preference is well-defined.

My analysis of team agency follows just the same logic, except that it

https://doi.org/10.1017/S0266267100000213 Published online by Cambridge University Press

uses a more general theory of the relationship between preference and action. That theory is what I have called team-directed reasoning. Formally, the theory of individual-directed reasoning is a special case of the theory of team-directed reasoning – the case in which the team has only one member.

When I say that a person who engages in team-directed reasoning has a team-directed preference for x over y, I mean that she takes herself to have reason to make those choices among actions that, according to the theory of team-directed reasoning, are implied by that preference. This analysis is a straightforward generalization of the standard analysis of individual-directed reasoning. Against a charge of circularity or emptiness, the two analyses stand or fall together.

9. FRAMES

The standard account of preference has another important feature, but one which is often overlooked: preferences are defined relative to particular conceptions of, or *framings* of, decision problems.[18] This feature is inevitable if the theory, when applied to the real world, is to have empirical content.

If the theory is not to be empty, some states of affairs which are in fact distinguishable from one another have to be treated as 'the same' for the purposes of the theory. To see why that is so, consider some individual who, in some specific choice problem P, has to choose between an action which leads to x and an action which leads to y. Suppose she chooses the former. Given the standard theory of rational choice, we are entitled to infer that she prefers x to y.[19] But that, in itself, merely records that she takes herself to have reason to choose what she in fact chooses. If we are to be able to draw any new inference about her choices, we have to be able to find some *different* choice problem P′ in which the options include actions which lead to x and to y. We can then infer that, in the new problem, she will not choose the action which leads to y. Clearly, for such a new problem to exist, each of the outcomes x and y must be treated as 'the same' when it occurs in P′ as when it occurs in P. But if P and P′ can be described as distinct entities, there must be *some* difference between 'x in P' and 'x in P′', and between 'y in P' and 'y in

[18] There is a thin strand of economic literature which does recognize the frame-relativity of preferences. See, for example, Broome (1991, Chapter 5), Bacharach (1993), and Sugden (1995).

[19] Here again, I leave aside the possibility of indifference. It is well known that indifference creates serious conceptual problems for any revealed-preference interpretation of the theory of rational choice. The usual way round these problems is to define preferences over some continuous space of outcomes and then to make assumptions (e.g. increasingness and continuity) which limit the range of cases in which indifference can occur.

https://doi.org/10.1017/S0266267100000213 Published online by Cambridge University Press

P''. The theory of rational choice is given content, in any specific application, by means of background assumptions about how finely outcomes may be individuated.

For example, the standard version of consumer theory defines outcomes for a given consumer as vectors of quantities of different goods. This definition allows many different market actions by the consumer to lead to the same outcome. Consider the case in which the same good (the same, that is, in terms of the theory's classification of goods) is offered by two different suppliers at the same price. Buying a given quantity of the good at this price is deemed to lead to the same outcome for the consumer, irrespective of who supplies it. The implicit assumption is not that the two suppliers are *indistinguishable*, but that the identity of the supplier does not enter into the consumer's conceptualization of the good. In other words, it is being assumed that the consumer takes her decision problem to be, or frames that problem as, a choice among final consumption bundles.

Further thought about this kind of example soon shows that questions about how individuals frame their decision problems do not always have easy answers. Consider two types of canned drink, sold in the same supermarket. Their ingredients are the same and, in blind tasting tests, people cannot distinguish between them. However, one type carries a highly-advertised brand name, while the other does not. Is there one good here or two? From the viewpoint of economics, there is no objective answer to this question, independent of consumers' subjective perceptions: what matters is whether consumers *take them to be* the same. Before we can make use of the theory of rational choice, then, we have to make some assumptions about individuals' conceptions of the decision problems they face.

Here is another example. Economic theories of the capital and insurance markets generally assume that people maximize expected utility and are risk-averse. However, the assumption of universal risk aversion appears to be inconsistent with the existence and profitability of the gambling industry. Economists who want to defend the risk-aversion assumption sometimes draw a distinction between 'wealth-oriented' and 'pleasure-oriented' decision-making under risk, and claim that risk-aversion is characteristic only of wealth-oriented decision-making, such as (it is asserted) occurs in capital and insurance markets. Gambling, in contrast, is classed as a pleasure-oriented recreational activity, akin to a consumption good or service; the price of this activity is the expected money loss from engaging in it.[20] Thus, the standard theory of rational

[20] For an example of this argument, see Hirshleifer and Riley (1992, pp. 26–8). I leave open the question of how convincing this particular argument is; some decision theorists would see it as an *ad hoc* stratagem to insulate an established theory from falsifying evidence.

https://doi.org/10.1017/S0266267100000213 Published online by Cambridge University Press

individual choice can be applied to gambling in at least two different ways, depending on how individuals are assumed to conceptualize acts of gambling. Nothing in the theory itself tells us which frame we should use when applying it.

Now consider the theory of team agency, as I have presented it. My theory retains the core idea of the standard theory of rational individual choice, that an action is rational for an agent to the extent that it leads to (or can be expected to lead to) preferred outcomes for that agent. The extra generality comes through allowing teams to be agents for the purposes of the theory, so that a given action can be judged to be rational or irrational from a range of different perspectives, each corresponding with a different potential team. In terms of my very first example: I may prefer ten-mile walks to six-mile ones when I adopt the perspective of myself as a single-member team, but prefer six-mile walks when I adopt the perspective of myself as a member of a family team. Or in terms of the Footballers' Problem: for the problem as I originally stated it, as a problem for real footballers, the rationality of team agency does indeed prescribe that each footballer chooses 'right'. But that is because the footballers *are* (and take themselves to be) members of the same team, and because that team has the objective of scoring goals. The theory does not imply that in *every* coordination game in which strategies and individual-directed preferences take the form shown in Table 1 (that is, in the Footballers' Problem as it would be represented in game theory), 'right' is uniquely rational for each player.

Thus, the implications of my theory may be indeterminate if the unit of agency has not been specified. Is this an objection to the theory?

My reply is that this is no more of an objection to my theory of team agency than it is an objection to the standard theory that preferences are defined relative to frames. In the standard theory, too, a given action can be judged rational or irrational from a range of different perspectives. The implications of this theory, too, may be indeterminate if the frame has not been specified. (Think of the action of buying the branded canned drink when the unbranded drink is on sale at a cheaper price.) Once again, my theory of team agency and the standard theory of rational choice stand or fall together.

10. CONCLUSION

As a rhetorical device, the companions-in-guilt strategy is often effective; and I hope it has been so in this paper. But what has it actually established?

I have argued that the theory of choice should allow teams of individuals to be collective decision-making agents and to have preferences. Two problems – the 'existence problem' of specifying which teams

https://doi.org/10.1017/S0266267100000213 Published online by Cambridge University Press

exist, and the 'objectives problem' of specifying what the objectives of teams are – have been seen as obstacles to the development of theories of team agency. But if the argument of this paper is correct, these are not problems for the theory of team agency *specifically*. They are manifestations of two more general problems – that of explaining how people conceptualize or frame decision problems, and that of explaining the origins of preferences. And these are problem for the theory of choice *in general*.

Obviously, realizing this does not make those problems disappear; but it may make them easier to tackle. When the problems have arisen in the analysis of team agency, they have often seemed more daunting than they really are, because decision theorists have thought that solutions have to take a very special form: they have supposed that the existence and objectives of teams must be derived from individual-directed preferences. The companions-in-guilt argument shows what is wrong with that supposition. If team agency is on a par with the kind of individual agency that is represented in the standard theory, then there is no reason to expect team-directed concepts to be reducible to individual-directed ones.

Instead, we might look at how economists already deal with analogues of the existence and objectives problems when applying the conventional theory of rational individual choice. I have argued that the conventional theory has no empirical content *on its own*, in the absence of any auxiliary hypotheses about how individuals conceptualize their decision problems. However, it serves as a template for the construction of more specific theories which *do* have empirical content.

Take the case of consumer theory. The core of this theory is the standard analysis of rational choice. But consumer theory also includes the fundamental auxiliary hypothesis that each individual conceptualizes her actions in markets as means to achieving preferred bundles of consumption goods. In addition, various hypotheses are advanced about what count as consumer goods, and about the general properties of consumers' preferences (for example, that preferences are continuous and convex, and that larger bundles are preferred to smaller ones) which go well beyond the implications of formal rationality. In particular applications of the theory, there might be further hypotheses, for example that particular pairs of goods are substitutes while others are complements.

Additional hypotheses of this kind are justified in various ways. Some are modelling conventions which have proved useful in generating successful theories; some are empirical generalizations, which are supported by various mixes of evidence, common-sense introspection, and psychological, sociological and biological theorizing; and some (continuity may be an example) are just there to make the analysis more

https://doi.org/10.1017/S0266267100000213 Published online by Cambridge University Press

tractable. Although these hypotheses are components of a theory which is intended to represent rational choices, they are not themselves hypotheses *of rationality*.

I have proposed a theory of team agency which generalizes the conventional theory of rational individual choice. Just like the latter theory, the theory I have proposed does not have empirical content as it stands. What it gives us is a template for constructing more specific theories which, by virtue of additional hypotheses, can have such content. The additional hypotheses that are needed are not hypotheses about the nature of rationality, but about how individuals frame their decision problems, which teams they take themselves to belong to, and what they take the objectives of those teams to be. There is nothing outlandish about what is required here. On the contrary, there is a tradition of theoretical and experimental research in social psychology which develops and tests hypotheses of just those kinds in order to explain 'group identity' (see, for example, Brewer and Kramer, 1986 and Brewer and Gardner, 1996).

So the relationship between my proposal and the conventional theory is not really that of companions *in guilt*. I have proposed a strategy for theorizing about teams which parallels the strategy that has generated our current theories of individual choice. To the extent that the latter theories can be judged successful, there is some reason to hope that my proposed strategy may succeed too.

REFERENCES

Bacharach, Michael. 1993. Variable universe games. In *Frontiers of Game Theory*, pp. 255–75. Ken Binmore, Alan Kirman and P. Tani (eds.). MIT Press

Bacharach, Michael. 1999. Interactive team reasoning: a contribution to the theory of cooperation. *Research in Economics*, 53:117–47

Bratman, Michael E. 1993. Shared intention. *Ethics*, 104:97–113

Brewer, Marilynn B. and W. Gardner. 1996. Who is this 'we'? Levels of collective identity and self representation. *Journal of Personality and Social Psychology*, 71:83–93

Brewer, Marilyn B. and R. M. Kramer. 1986. Choice behavior in social dilemmas. *Journal of Personality and Social Psychology*, 50:543–49

Broome, John. 1991. *Weighing Goods*. Blackwell

Copp, David. 1980. Hobbes on artificial persons and collective actions. *The Philosophical Review*, 89:579–606

Gauthier, David. 1975. Coordination. *Dialogue*, 14:195–221

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

Gilbert, Margaret. 1999. Reconsidering the 'actual contract' theory of political obligation. *Ethics*, 109:236–60

Harsanyi, John and Reinhard Selten. 1988. *A General Theory of Equilibrium Selection in Games*. Harvard University Press

Hirshleifer, Jack and John G. Riley. 1992. *The Analytics of Uncertainty and Information*. Cambridge University Press

Hobbes, Thomas. 1651/1962. *Leviathan*. Macmillan

Hodgson, D. H. 1967. *Consequences of Utilitarianism*. Clarendon Press

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

https://doi.org/10.1017/S0266267100000213 Published online by Cambridge University Press

Hurley, Susan L. 1989. *Natural Reasons*. Oxford University Press
Lewis, David. 1969. *Convention: A Philosophical Study*. Harvard University Press
Pareto, Vilfredo. 1906/1972. *Manual of Political Economy*. Trans. A. S. Schweir. Macmillan
Regan, Donald. 1980. *Utilitarianism and Cooperation*. Clarendon Press
Scanlon, Thomas M. 1991. The moral basis of interpersonal comparisons. In *Interpersonal Comparisons of Well-Being*, pp. 17–44. Jon Elster and John E. Roemer (eds.). Cambridge University Press
Searle, John R. 1990. Collective intentions and actions. In *Intentions in Communication*, pp. 410–15. Philip R. Cohen, Jerry Morgan and Martha E. Pollack (eds.). MIT Press
Skyrms, Brian. 1996. *Evolution of the Social Contract*. Cambridge University Press
Sugden, Robert. 1991. 'Rational choice: a survey of contributions from economics and philosophy. *Economic Journal*, 101:751–85
Sugden, Robert. 1993. Thinking as a team: toward an explanation of nonselfish behavior. *Social Philosophy and Policy*, 10:69–89
Sugden, Robert. 1995. A theory of focal points. *Economic Journal*, 105:1269–302
Tuomela, Raimo. 1995. *The Importance of Us*. Stanford University Press

https://doi.org/10.1017/S0266267100000213 Published online by Cambridge University Press