Economics and Philosophy, **18** (2002) 63–87

BEYOND SYMPATHY AND EMPATHY: ADAM SMITH'S CONCEPT OF FELLOW-FEELING

Robert Sugden
University of East Anglia

When modern economists use the notions of sympathy or empathy, they often claim that their ideas have their roots in Adam Smith's *Theory of Moral Sentiments* (1759/1976), while sometimes complaining that Smith fails to distinguish clearly enough between the two concepts. Recently, Philippe Fontaine (1997) has described various forms of sympathy and empathy, and has explored their respective roles in Smith's work. My objective in this paper is to argue that Smith's analysis of how people's sentiments impinge on one another involves a concept of *fellow-feeling* that is distinct from both sympathy and empathy. Unlike sympathy and empathy, fellow-feeling does not fit into the ontological framework of rational choice theory – which may explain why it tends to be overlooked by modern readers of Smith.

In Section 1, I examine how sympathy and empathy are understood in rational choice theory. In Sections 2 and 3, I present Smith's analysis of fellow-feeling and approval, and show how radically this differs from modern theories of sympathy and empathy. In Section 4, I suggest that Smith's theoretical approach can help to explain how social relations have subjective value for human beings, and in so doing, can add an

The ideas presented in this paper have developed in discussions with many people, but particularly Michael Bacharach, Nicholas Bardsley, Luigino Bruni, Robin Cubitt, Margaret Gilbert, Benedetto Gui, Shaun Hargreaves Heap, Martin Hollis, Judith Mehta, Timothy O'Hagan, Chris Starmer and Jung-Sic Yang. Earlier versions of this paper were presented to a Gerst Program conference on community and responsibility at Duke University, and to a seminar on economics and interpersonal relations organized by the University of Padova. I am grateful for comments from participants at those meetings and from two referees (one of whom identified himself as Stephen Darwall). My work has been supported by the Leverhulme Trust.

affective dimension to theories of team thinking. Finally, in Section 5, I consider how Smith's account of fellow-feeling fits with his understanding of economics and of rationality.

1. SYMPATHY AND EMPATHY IN RATIONAL CHOICE THEORY

To understand the ways in which the concepts of sympathy and empathy are used in the modern theory of rational choice, it is necessary to consider the conceptual framework within which that theory is constructed. And to understand why this framework is as it is, we need to look at the history of rational choice theory and of welfare economics.[1]

Rational choice theory and welfare economics are in a direct line of descent from the utilitarianism of Jeremy Bentham. It is sometimes suggested that David Hume and Smith belong to the same utilitarian tradition,[2] but I take there to be a fundamental divide between, on the one hand, Hume and Smith and, on the other, Bentham and his economist successors. While the Scottish writers are sceptical about the powers of reason and emphasize the diversity of natural human sentiments, the Benthamite tradition appeals to universal principles of rationality and favours stylized models of human psychology.

In classical utilitarianism, the fundamental concept is *utility*. Bentham advocates a *principle of utility* which 'approves or disapproves of every action whatsoever, according to the tendency which it appears to have to augment or diminish the happiness of the party whose interest is in question'. Utility is defined as:

> that property in any object, whereby it tends to produce benefit, advantage, pleasure, good, or happiness (all this in the present case comes to the same thing) or (what again comes to the same thing) to prevent the happening of mischief, pain, evil, or unhappiness to the party whose interest is considered: if that party be the community in general, then the happiness of the community: if a particular individual, then the happiness of that individual.

Bentham immediately goes on to say that the 'interest of the community' must be understood as 'the sum of the interests of the several members who compose it', and that a thing promotes the interest of an individual 'when it tends to add to the sum total of his pleasures' (1789/1970, p. 12). If, as Bentham clearly intends, rational decision making is to be defined as the maximization of pleasure, there must be a common currency of

[1] The following discussion draws on Hollis and Sugden (1993).

[2] For example, Rawls (1971, pp. 22–3) treats Hume's *Treatise of Human Nature* (1740/ 1978) and Smith's *Theory of Moral Sentiments* as founding texts of classical utilitarianism. See also note 9 below.

pleasure (and its negative, pain), to which all sentiments relevant to rational decision making can be reduced.

Notice that rationality in decision making is defined relative to the 'party whose interest is considered'. Two definitions of this party – two ways of *framing* the problem of rational decision making – are particularly salient for utilitarianism: the party as 'a particular individual', and the party as 'the community in general'. A decision is rational with respect to the *individual frame* if it maximizes the sum of pleasures for the particular individual in question; it is rational with respect to the *community frame* if it maximizes the sum of pleasures for all individuals in the relevant community. This is not to say that, according to utilitarians, rationality requires each person to use the individual frame when he decides how to act in his private sphere of life.[3] Many utilitarians have held that a truly rational individual would frame all decisions in terms of the interests of the widest possible community. However, utilitarian economists have generally assumed that the actual decisions of individuals as economic agents are approximately rational with respect to the individual frame.

This utilitarian framework was retained by the founders of neoclassical economics. These economists gave their theories more mathematical structure by defining a *utility function* for each individual, which assigns a numerical index of utility to every bundle of consumption goods; individual rationality was then defined as the maximization of the value of this function subject to a feasibility constraint. The welfare of a community was defined as the sum of the utility indices of its component individuals.

The most significant disruption to this tradition came at the beginning of the twentieth century when there was a move, initiated by Vilfredo Pareto (1909/1972), to jettison the psychological assumptions of utilitarianism. The 'Paretian turn' was gradually accepted by the economics profession, culminating in Paul Samuelson's (1947) revealed preference theory and in Leonard Savage's (1954) axiomatic formulation of expected utility theory. As a result of these innovations, the concept of the utility function has been retained, along with the idea that all of an individual's choice-relevant attitudes to goods can be represented in a single dimension; but the utility function has been re-interpreted as a representation of the individual's *preferences* over consumption bundles. The common currency of pleasure has been replaced by the common currency of preference. 'Preference' has been given various interpreta-

[3] Lyons (1973, p. 20) argues that Bentham's own position is that 'one ought to promote the happiness of . . . those subject to one's direction, influence, or control'; thus, roughly speaking, governments ought to consider the happiness of the communities they govern, and private individuals should consider their own happiness.

tions, but always with the core idea that an individual's preferences are reflected in or revealed in her choices. Thus, the individual's preferences can be understood as whatever she takes to be choice-relevant reasons, all things considered, or as the psychological dispositions that prompt her to make whatever choices she makes, or (in the most austere versions of the theory) simply as re-descriptions of those choices.

Sympathy and empathy were brought into rational choice theory to deal with two distinct problems. The first problem, which confronted both the utilitarian and Paretian versions of rational choice theory, was that of explaining non-selfish behaviour. Economists have generally preferred, whenever possible, to explain human behaviour in terms of rational self-interest; but even in the domain of 'economic' decisions, narrowly construed, we can find behaviour that is difficult to explain without invoking some kind of non-selfish motivation. (Charitable donations provide an example.) How can such motivations be introduced into rational choice theory? Taking the sparse ontology of the theory as given, the most obvious way to represent non-selfish motivation is as *a kind of preference*. Hence the idea, now well-established in rational choice theory, of assuming what are variously called *altruistic*, *benevolent* or *sympathetic* preferences: Joe is sympathetic to Jane to the extent that he has an intrinsic preference that her preferences are satisfied. Or, equivalently: Joe is sympathetic to Jane to the extent that her utility is an argument in his utility function.[4]

The second problem confronting rational choice theory is an unavoidable consequence of taking the Paretian turn. It is the problem of making sense of interpersonal comparisons of utility, when 'utility' is interpreted as a representation of preferences and not as a measure of pleasure. The Paretian analogue of the sum of individuals' utilities is a *Bergson-Samuelson social welfare ordering*, which ranks all *social states* – that is, all possible states of affairs for society. Formally, a social welfare ordering ranks social states in the same way that a preference ordering ranks consumption bundles for a given individual. (Just as, given certain assumptions, a preference ordering can be represented by a utility function, so a social welfare ordering can be represented by a *social welfare function*.) Intuitively, the social welfare ordering represents the viewpoint of the community as a whole, while a preference ordering represents the viewpoint of a particular individual. But what exactly does it mean to talk about the viewpoint of the community as a whole? An individual's preference ordering is revealed in her consumption choices. Where is the social welfare ordering revealed?

John Harsanyi's (1955) answer to this question has been very

[4] This concept of sympathy was introduced to utilitarian rational choice theory by Edgeworth (1881), two decades before the Paretian turn was made.

influential. Harsanyi distinguishes between an individual's *subjective preferences* – the preferences that are revealed in her actual choices – and her *ethical preferences*, interpreted as her judgements about the welfare of the community as a whole. The idea is that a person's ethical preferences are those preferences that would be revealed in her choices among social states, were she impartially taking account of every individual's subjective preferences. Harsanyi models this hypothetical choice situation by assuming that the chooser does not know which of the actual individuals in the community she is, but knows only that she has the same probability of being each of them. Let us call the constraints on information in this situation *Harsanyi's veil of ignorance*.

In order to apply expected utility theory to this decision problem, Harsanyi has to assume that the chooser has preferences over objects of the form (x, i), where x is any social state and i is any person; for her to prefer some (x, i) to some (y, j) is for her to judge that she would prefer being person i in social state x to being person j in social state y. Preferences of this kind are *empathetic preferences*. Harsanyi takes it as axiomatic that, in judging whether it is better to be some person i in some social state x, or the same person i in a different social state y, the chooser's empathetic preferences must coincide with person i's subjective preferences between x and y: if the chooser is to imagine *being* person i, she must suppose that she takes on i's subjective preferences.[5]

Harsanyi shows that, given the axioms of expected utility theory, ethical preferences must take the form of a social welfare ordering. Further, this ordering can be represented by a social welfare function in which the index of social welfare is the sum of indices of individual utility, and in which the index of utility for each individual is a representation of that individual's subjective preferences. Thus, by using empathetic preferences, Harsanyi constructs a form of utilitarianism that is compatible with the Paretian turn.

Working in the tradition of social contract theory, Ken Binmore (1994, 1998) proposes a variation of Harsanyi's approach to interpersonal comparisons. Binmore distinguishes between the *game of life* and the *game of morals*. The game of life represents real human interaction, in which each individual acts on his own (subjective) preferences. For an outcome of the game of life to be stable, it must be a Nash equilibrium: each individual must maximize his own utility, given the behaviour of the others. If, however, the game of life has more than one equilibrium,

[5] Whether the idea of such preferences is coherent is a matter of dispute. Rawls (1971, pp. 173–5) denies the meaningfulness of preferences between *being* one person, with all of that person's character, desires and purposes, and *being* another. In order for the chooser to have a preference that is *hers*, she must have a standpoint of her own, and Harsanyi's framework does not allow this. I agree with Rawls; but my object here is to explain Harsanyi's construction, not to defend it.

there is a problem of equilibrium selection, which has to be solved by convention. Binmore proposes a particular convention, intended as a model of how in fact equilibrium selection problems are typically solved. This convention is to select whichever equilibrium would be chosen by the players of the game of life, were they required to reach agreement on an equilibrium while located behind Harsanyi's veil of ignorance, acting on empathetic preferences of the kind assumed by Harsanyi. The imaginary process of reaching agreement behind this veil of ignorance is the game of morals. Thus Binmore, like Harsanyi, distinguishes between a domain of real behaviour in which individuals act on subjective preferences, and a domain of moral reasoning in which they imagine acting on empathetic preferences.

In both Harsanyi's and Binmore's constructions, empathy and sympathy are fundamentally different concepts. Sympathy is revealed in an individual's *actual* choices, and so is a property of subjective preferences. Empathy impacts only on ethical preferences. Here is Binmore's summary of the distinction:

> Adam sympathizes with Eve when he so identifies with her aims that her welfare appears as an argument in his utility function. . . . The extreme example is the love a mother has for her baby. Adam empathizes with Eve when he puts himself in her position to see things from her point of view. Empathy is not the same as sympathy because Adam can identify with Eve without caring for her at all. For example, a gunfighter may use his empathetic powers to predict an opponent's next move without losing the urge to kill him. (1998, p. 12)

Notice how the conceptual framework of rational choice theory constrains what can be said about sympathy and empathy. In Binmore's account, the distinguishing characteristic of sympathy is that it is registered in the sympathizer's utility function – that is, that Adam's choices are affected by his sympathy for Eve. There is no way of saying that Adam's *feelings* are affected by his perception of Eve's *feelings*, without also saying that Adam is motivated to perform actions which benefit Eve. Once the Paretian turn has been taken, this feature of rational choice theory is unavoidable, because feeling, like all other psychological concepts, has been stripped out of the conceptual scheme. But the utilitarian version of rational choice theory faces a similar problem, as a result of its one-dimensional psychology. In the utilitarian scheme, the only way that Adam's feelings can be affected by his perception of Eve's feelings is for him to gain pleasure from his perception of Eve's pleasure. Since individual rationality is understood as the maximization of pleasure, Adam's sympathy for Eve must also be a motive for action for him.

There is a corresponding flattening of the idea of empathy.

Intuitively, the idea of empathy seems to signify one person's *understanding* of another. But the only mental attitudes that rational choice theory admits are preferences (and, in the presence of uncertainty, beliefs). Thus, the limit of Adam's understanding of Eve is reached when Adam has full knowledge of Eve's preferences and beliefs. Binmore's example of the gunfighter illustrates the point. The gunfighter is interested only in predicting his opponent's actions. In order to predict the behaviour of a rational opponent, it is sufficient to know his preferences and beliefs. 'Empathetic power', in Binmore's sense, can be nothing more than an ability to *discover* another person's preferences and beliefs. There is no way of representing the intuitive idea of Adam's *entering into* or *going along with* (it is hard to avoid saying *sympathizing with*) Eve's feelings, without also asserting that Adam *cares for* Eve, that Adam is motivated to confer benefits on Eve. The same limitation appears in Harsanyi's and Binmore's assumption that, when Adam empathetically identifies with Eve, he imaginatively takes on all Eve's actual preferences, whatever they may be. There is no room in Harsanyi's and Binmore's conceptual scheme for a notion of empathetic understanding that could allow Adam, with full knowledge of Eve's preferences, to go along with some of her feelings but not others.

Within modern rational choice theory, then, the distinction between sympathy and empathy is categorical; and these two theoretical concepts seem to exhaust the possibilities for representing positive relationships between the mental states of different people. From this perspective, arguments which do not recognize a sharp distinction between sympathy and empathy can seem merely confused. Consider Binmore's discussion of the 'Adam Smith problem' – the alleged inconsistency between *The Theory of Moral Sentiments* and *The Wealth of Nations* (Smith, 1776/1976). Binmore claims that, in order to recognize the mutual consistency of these books, we need to recognize that Smith's definition of sympathy is similar to the modern definition of empathy:

> Commentators on the Adam Smith problem are largely agreed that we must look to the modern distinction between empathy and sympathy in order to achieve a reconciliation between his two books . . . However, modern apologists are too ready to forgive Adam Smith for failing to honor his own definition whenever he offers a serious argument. Instead, he repeatedly falls into the trap of appealing to sympathy in the sense that it is understood [in rational choice theory]. That is to say, he implicitly assumes that the welfare of others appears as an argument in our *personal* utility functions.

Binmore goes on to say that if we are to understand Smith, we must begin by 'clearing away the confusion between the concepts of empathy and sympathy that he shared with David Hume' (1998, p. 368).

But perhaps the confusion results from trying to read Smith through the lens of modern rational choice theory. Perhaps Smith has a model of inter-relationships between individuals' mental states which cannot be represented in the framework of that theory, but which is nonetheless coherent. And perhaps Smith's model represents significant features of the real world, which the modern theory has edited out. Let us see.

2. THE PLEASURE OF MUTUAL SYMPATHY

The most famous words in *The Theory of Moral Sentiments* are probably those of the opening sentence: 'How selfish soever man may be supposed, there are evidently some principles in his nature, which interest him in the fortune of others, and render their happiness necessary to him, though he derives nothing from it except the pleasure of seeing it' (p. 9).[6] The fame of this sentence is, I think, unfortunate. For a modern economist, the idea that one person has an interest in the fortune of others, or derives pleasure from the pleasure of others, immediately suggests a model of altruistic preferences. But this would misrepresent Smith's intentions.

Almost equally famous is Smith's highly-coloured opening example. He presents a supposedly typical human response to the knowledge that a fellow-man is being tortured:

> By the imagination we place ourselves in his situation, we conceive ourselves enduring all the same torments, we enter as it were into his body, and become in some measure the same person with him, and thence form some idea of his sensations, and even feel something which, though weaker in degree, is not altogether unlike them. His agonies, when they are thus brought home to ourselves, when we have thus adopted and made them our own, begin at last to affect us, and we then tremble and shudder at the thought of what he feels. (p. 9)

This is followed by another, similar example:

> That this is the source of our fellow-feeling for the misery of others, that it is by changing places in fancy with the sufferer, that we come either to conceive or to be affected by what he feels, may be demonstrated by many obvious observations . . . When we see a stroke aimed and just ready to fall upon the leg or arm of another person, we naturally shrink and draw back our own leg or our own arm; and when it does fall, we feel it in some measure, and are hurt by it as well as the sufferer. (p. 10)

Are these responses sympathy or empathy?

It is clear that Smith's spectator is *identifying* with the victim and

[6] All unattributed citations are to Smith (1759/1976).

https://doi.org/10.1017/S0266267102001086 Published online by Cambridge University Press

imagining *an experience of* pain. Through this act of imagination, the spectator is cognitively able to attribute particular feelings of pain to the other person. So far, this is empathy in the modern sense. But in addition, this imagining of pain is a source of *real* – not just imaginary – pain to the spectator. Is this sympathy in the modern sense? Not necessarily: we are not entitled to infer that the spectator is motivated to act to benefit the victim. Particularly in Smith's second example, the spectator's imagining of the victim's pain is presented as an involuntary psychological response, specific to a particular moment in time and to a particular type of feeling (even to a particular part of the spectator's body). What effect this response has on the spectator's actions is left open by Smith's account. Smith is writing about *affective states*, about how one person's affective state influences another's, not about *preferences*. The concepts he is using do not belong to the ontology of rational choice theory.

To avoid confusion, I shall use Smith's term *fellow-feeling* to represent interdependencies of feeling of the kind shown in these examples. That is: fellow-feeling is to be understood as one person's lively consciousness of some affective state of another person, where that consciousness itself has similar affective qualities – pleasurable if the other person's state is pleasurable, painful if it is painful.

The real distinctiveness of Smith's account emerges in his discussion 'of the pleasure of mutual sympathy' (the title of the second chapter of *The Theory of Moral Sentiments*). He proposes that human beings derive pleasure from all forms of fellow-feeling. Suppose that Jane experiences some pleasure or pain, and that Joe has fellow-feeling for this. Joe's fellow-feeling consists in a qualitatively similar, but perhaps much weaker, imaginatively experienced pleasure or pain. But, according to Smith, an additional psychological mechanism comes into play, which gives pleasure both to Joe and to Jane, *irrespective of whether Jane's original feeling was pleasure or pain*. Jane's consciousness of Joe's fellow-feeling for her is a source of pleasure to her; and Joe's consciousness of his own fellow-feeling for Jane is a source of pleasure to him.

From a theoretical point of view, this mechanism may seem surprising. It might seem more natural to model fellow-feeling as nothing more than a kind of *reflection* of feeling. That would lead to the implication that if Jane's original feeling was one of pain, then Joe's fellow-feeling for it would be painful for him, and Jane's consciousness of Joe's painful fellow-feeling would be painful for her. The conventional rational-choice model of altruism implies the same kind of reflection, but in the domain of preferences rather than feelings. Indeed, in a depressingly straight-faced paper about altruism within the family, published in one of the world's leading economics journals, Douglas Bernheim and Oded Stark (1988) use just this kind of model to argue that 'nice guys finish last': people who derive relatively little happiness (or 'felicity')

from consumption and who are themselves altruistic will prefer to have partners who are not altruistic towards them. Thus, the authors claim, nice guys may be rejected as potential partners because they are too altruistic. They say: 'The explanation is quite simple. An altruistic type A [i.e., man] would be depressed by his partner's low level of felicity. Since the type B [i.e., woman] cares about her partner, she would in turn be disturbed by the fact that she has made him unhappy'. In other words: if you are unhappy, other people's sympathy with your unhappiness is an additional cause of unhappiness for you.

Smith considers this kind of model, in which sympathy is simply the reflection of feeling, but rejects it as not compatible with human psychology as we know it:

> The sympathy, which my friends express with my joy, might, indeed, give me pleasure by enlivening that joy: but that which they express with my grief could give me none, if it served only to enliven that grief. Sympathy, however, enlivens joy and alleviates grief. It enlivens joy by presenting another source of satisfaction; and it alleviates grief by insinuating into the heart almost the only agreeable sensation which it is at that time capable of receiving. (p. 14)

Notice that Smith is hypothesizing 'another source of satisfaction', distinct from the pleasures and pains that are constitutive of fellow-feeling. This satisfaction derives from the *correspondence of sentiments* between oneself and another: 'this correspondence of the sentiments of others with our own appears to be a cause of pleasure, and the want of it a cause of pain, which cannot be accounted for [by a theory of reflected feelings]' (p. 14).

It is not entirely clear whether Smith thinks this pleasure can be induced by the mere *knowledge* that one's own sentiments are aligned with those of another person, or whether he thinks there has to be a lively *consciousness* of this alignment, based on imaginative identification with the other. The chapter title, 'Of the pleasure of mutual sympathy', implies the latter, with the implication that this pleasure arises only from consciousness of fellow-feeling. However, some of Smith's examples suggest the former, broader interpretation. As I shall show shortly, it is important for his argument that dissonance between our sentiments and those of others is a source of pain. In this kind of case, we clearly do *not* have a lively consciousness of the other person's sentiments, based on imaginative identification. Instead – at least in the cases that interest Smith – we are aware of a divergence between the other person's actual sentiments and those sentiments that, had she had them, we could have had fellow-feeling for. I suggest that the best reading of Smith is that our awareness of *any* correspondence of our sentiments with those of others is a potential source of pleasure, and that our awareness of *any*

dissonance is a potential source of pain. That consciousness of fellow-feeling is pleasurable is an instance of this more general hypothesis.

So, if Jane is in a state of grief, Joe's fellow-feeling for her consists in his consciousness of her pain, which is painful to him too. But by virtue of this fellow-feeling, there is a correspondence of sentiments between Joe and Jane, and their consciousness of this correspondence is a source of pleasure to them both. Smith thinks that the pleasure derived from the correspondence of sentiments usually outweighs any pains of fellow-feeling. Thus, we are *pleased* when we are able to feel sympathy for the painful feelings of others. With what I believe to be psychological acuteness, Smith points to the obverse of this phenomenon: the unease and irritation we feel when we find we cannot sympathize with someone else's apparent sentiments of distress (p. 16).

The pleasures of mutual fellow-feeling can be enjoyed in any joint activity between people whose sentiments are suitably aligned. Smith gives an example of one person reading aloud to another, which will strike a chord with many parents:

> When we have read a book or poem so often that we can no longer find any amusement in reading it by ourselves, we can still take pleasure in reading it to a companion. To him it has all the graces of novelty; we enter into the surprise and admiration which it naturally excites in him, but which it is no longer capable of exciting in us; we consider all the ideas which it presents rather in the light in which they appear to him, than in that in which they appear to ourselves, and we are amused by sympathy with his amusement which thus enlivens our own.

Conversely, if the sentiments of the reader and the listener do not correspond, the jointness of their activity is a source of pain: 'we should be vexed if he [the listener] did not seem to be entertained with it, and we could no longer take any pleasure in reading it to him' (p. 14). Smith is offering a theoretical account of the subjectively-experienced difference between doing something alone and doing the same thing *together with others*. The account depends on hypotheses about causal relationships between affective mental states, and makes no reference to preference or choice. Thus, it cannot be expressed in the language of a theory of rational choice from which all references to affective states have been stripped out.

3. PROPRIETY

For Smith, the psychology of fellow-feeling and the correspondence of sentiments is tightly linked with that of approval and disapproval; and approval and disapproval form the basis of our sense of morality. Smith's first, rough formulation of the link between fellow-feeling and

approval is that we approve of other people's sentiments just to the extent that we 'go along with' them – that is, to the extent that we have fellow-feeling for them:

> When the original passions of the person principally concerned are in perfect concord with the sympathetic emotions of the spectator, they necessarily appear to this last just and proper, and suitable to their objects; and, on the contrary, when, upon bringing the case home to himself, he finds that they do not coincide with what he feels, they necessarily appear to him unjust and improper, and unsuitable to the causes which excite them. To approve of the passions of another, therefore, as suitable to their objects, is the same thing as to observe that we entirely sympathize with them; and not to approve of them as such, is the same thing as to observe that we do not entirely sympathize with them. (p. 16)

He then adds that we can also approve or disapprove of another person's sentiments by recognizing that we are *capable of* going along with them, even if, because of the particular circumstances of the case, we do not actually do so. For example, if we merely hear that a stranger has suffered some serious misfortune, and observe actions of his that express intense grief, we might not actually *feel* sympathetic pain. Still, we may recognize that, were that person's circumstances to be brought close to us, we would experience fellow-feeling. Such recognition – our 'consciousness of conditional sympathy' – is enough to ground our approval of the stranger's sentiments as suitable to their object, as showing *propriety* (pp. 17–18).

It is crucial to Smith's analysis of approval that the imaginative identification with others that constitutes fellow-feeling is only partial – in contrast to the total identification presupposed by Harsanyi's model of empathy. That is, when Joe imaginatively changes places with Jane, he takes with him enough of his own characteristics to be able to pose the question of whether, in Jane's circumstances, *his* sentiments would be the same as Jane's. Disapproval is possible only because the answer to this question can be negative.

Smith offers many examples in which his representative person (the generic 'we'), imaginatively identifying with another, does *not* go along with the other's sentiments. He uses these examples to support hypotheses about how the psychological mechanisms of fellow-feeling work. There is no suggestion that we are necessarily at fault when we do not enter into the other person's sentiments. Rather, the suggestion is that (at least from our point of view), the other's sentiments are at fault. Thus, if the other's passion is too intense for us to go along with, we call it weakness (in the case of grief) or fury (in the case of resentment); if it is not intense enough, we call it insensibility or lack of spirit (p. 27). A simple example: 'We are even put out of humour if a companion laughs

louder or longer at a joke than we think it deserves; that is, than we feel that we ourselves could laugh at it' (pp. 16–17). The presence or absence of fellow-feeling for varying responses to humorous situations is the basis for our judgements about the propriety of degrees of laughter.

Many commentators have remarked on the delicacy – some would say, the ambiguity or equivocation – of Smith's account of identification. Exactly which personal characteristics do I take with me when I imagine myself as another person? Recall the passage about the blow aimed at the other person's leg. Here Smith seems to be suggesting that my fellow-feeling with the victim depends on my identifying *my leg* with *his leg*. Yet he also wants to allow that a man can have fellow-feeling for the pain that a woman experiences in childbirth, 'though it is impossible that he should conceive himself as suffering her pains in his own proper person and character'. In the same passage as he presents the childbirth example, he says (in the context of sympathy with grief): '[I]n order to enter into your grief . . . I consider what I should suffer if I was really you, and I not only change circumstances with you, but I change persons and characters' (p. 317).[7] How can this be reconciled with Smith's earlier claim that excesses of grief do not elicit sympathy, but are put down to the 'weakness' of the griever? Isn't the other person's weakness – his particular susceptibility to grief – an aspect of his character?

The best I can do to dissolve this tension is to point out that when Smith writes about 'changing persons and characters', his object is to deflect the criticism that sympathy is founded in self-love. It is not to explain when we do and do not sympathize with others. He is saying that *if and when* we enter into another person's sentiments, the standpoint from which we imaginatively experience those sentiments is that of the other person: we see the world from (what we take to be) the other person's point of view. Thus, our imagined feelings are not a form of self-love. But that is not to say that other people's sentiments, reflecting as they do the peculiarities of those people's characters, *necessarily* elicit our sympathy. Before we can enter into another person's sentiments, we have to recognize them as sentiments that *we* – in some sense that allows us to retain our own identities – would feel in that person's circumstances.

In another apparently ambiguous passage, Smith says that we can feel sympathy for a person who has lost his reason, even when that person appears to be happy in his own way, 'insensible of his own misery'. In this case:

[7] Fontaine (1997, pp. 266–7) uses this passage to support his claim that Smith's core concept of identification involves imaginatively *becoming* the other person. I recognize that Fontaine's is a natural reading of the passage; but given the central role that Smith's account of approval plays in the *Theory of Moral Sentiments*, I think we must assume that Smith intended identification to be only partial.

> The compassion of the spectator must arise altogether from the consideration of what he himself would feel if he was reduced to the same unhappy situation, and, what perhaps is impossible, was at the same time able to regard it with his present reason and judgment. (p. 12)

Reading Smith from the vantage point of rational choice theory or analytical philosophy, it would be easy to interpret the clause 'what perhaps is impossible' as an admission of incoherence. But we must remember that Smith is not telling us how rational agents ought to go about the process of imaginative identification. He is reporting what he believes to be facts about human psychology. In a certain sense, Smith thinks our compassion for the insane person is misplaced. But precisely because compassion is misplaced in that situation, the fact that we feel compassion is valuable evidence about the psychology of identification.[8]

Whatever we make of the details of Smith's theory of identification, it is clear he is assuming that, as a matter of psychological fact, human beings, knowing what the sentiments of others are, sometimes go along with those sentiments and sometimes do not. The difference between these two cases is the origin of our judgements of approval and disapproval.

Approval, Smith says, works like a mirror. If a human being could come to adulthood without having any contact with fellow humans, he would have no conception of his own sentiments as objects of thought. But society provides each of us with a mirror 'in the countenance and behaviour of those he lives with, which always mark when they enter into, and when they disapprove of his sentiments' (p. 110). In this way, we become conscious of our own sentiments, and of other people's approval and disapproval of them. Given Smith's analysis of approval, becoming conscious of other people's approval just *is* becoming conscious of an (actual or conditional) correspondence of sentiments. Thus, we receive pleasure from the consciousness that others approve of us, and pain from the consciousness of other people's disapproval:

> Nature, when she formed man for society, endowed him with an original desire to please, and an original aversion to offend his brethren. She taught him to feel pleasure in their favourable, and pain in their unfavourable regard. She rendered their approbation most flattering and most agreeable to him for its own sake; and their disapprobation most mortifying and most offensive. (p. 116)

[8] Modern cognitive psychologists recognize that 'anomalous' situations, in which human judgements deviate systematically from normative standards of truth or validity, can often give useful information about how judgements are made in 'normal' cases. Kahneman (1996, p. 252) gives as an example the human tendency to over-estimate distances in fog; this is evidence that our perceptual mechanisms interpret visual blur as a signal of distance.

The psychological mechanisms of approval and disapproval tend to induce norms of propriety of sentiment within any group of interacting people. Because we desire approval, we earn subjective rewards for changing our sentimental repertoires in ways which bring them into line with prevailing norms, and we incur subjective penalties for changes which deviate from those norms. These inducements lead us, consciously or unconsciously, to adapt our sentiments so as to align them with whatever norms of propriety are approved of by others. This social process – what Smith calls 'the great school of self-command' (p. 145) – imparts a tendency for people who live together in a society to develop similar affective responses to similar stimuli, and to subscribe to norms which give approval to those responses. Thus, in what Smith would think of as a well-ordered society, there is a close correspondence between people's actual sentiments (still more, between their *expressions* of sentiment) and the sentiments that others can go along with. But the process of reaching this equilibrium is not one of impartial empathy, as understood by Harsanyi: it involves changes in *actual* sentiments, as well as changes in fellow-feeling.

These psychological mechanisms provide the building blocks for Smith's explanation of the human sense of morality. *The Theory of Moral Sentiments* is a study of spontaneous order. Smith tries to show that the complex order we observe in the world of morality is an unintended consequence of the interactions of many individuals, each of whom acts on simpler principles of fellow-feeling. For Smith, norms of propriety *are* moral sentiments: 'the general rules of morality . . . are ultimately founded upon experience of what, in particular instances, our moral faculties, our natural sense of merit and propriety, approve, or disapprove of' (p. 159).

I am inclined to say that *all* Smith's moral sentiments are norms of propriety. In Smith's formal system, there is a distinction between our sense of the propriety of an action and our sense of its merit. An action has propriety to the extent that the motivating sentiment of the actor is in proportion to the cause that has excited it. (Thus, the person who laughs too loudly acts with impropriety: his motivating sentiment, of amusement, is disproportionate to the humour of the joke.) An action has merit or demerit to the extent that it is deserving of reward or punishment (p. 18). But our sense of the merit of an action depends on our sympathy with the gratitude of those who are benefited by it; and our sense of its demerit depends on our sympathy with the resentment of those who are harmed by it (pp. 67–9). In effect, our moral sentiments of merit and demerit are norms of propriety with respect to gratitude and resentment.

For Smith, the ideal standard of moral sentiment is to be found in the judgements of the *impartial spectator*. It is important to recognize that the

impartiality of Smith's spectator is not at all the same as the impartial empathy of Harsanyi's ethical preferences.[9] Harsanyi's construction adopts the viewpoint of someone who identifies equally with every individual's pleasures and pains, or who takes an impartial interest in the satisfaction of every individual's desires; by imagining ourselves as such an ideal empathizer, we are supposed to be able to aggregate the preferences of distinct individuals into a single measure of social welfare. But Smith's impartial spectator is not an aggregating device. Rather, he represents, in an idealized form, the *correspondence* of sentiments that is induced by social interaction: he represents the mirror of social approval.

To take the viewpoint of the impartial spectator is to bring one's own sentiments into correspondence with what other people can go along with. Thus, the impartial spectator feels the whole range of natural sentiments, to the extent that they are susceptible to fellow-feeling: he does not feel only those pleasures and pains that would enter into a utilitarian calculus. For example, it is important for Smith that we have a natural sentiment of resentment – the sense of humiliation and anger that we experience when we are conscious that some other person has injured us in a manner that we feel to be illegitimate, and which prompts us to seek redress and revenge. Resentment, he argues, is particularly susceptible to fellow-feeling. In Smith's theory, moral sentiments concerning justice are norms of propriety for resentment (pp. 74–91). Justice is not based on a utilitarian calculus of pleasure and pain, but on the psychology of resentment.

The impartial spectator has fellow-feeling for each sentiment in its due proportion. What are those due proportions? Just those proportions that people in general can most easily go along with. Harsanyi's ideally impartial empathizer would take account of each person's feelings at their 'true' values, that is, as experienced by the person who feels them. In contrast, Smith's ideal spectator has fellow-feeling for other people's sentiments, not in proportion to the intensity with which those sentiments are actually felt, but in proportion to their general tendency to induce fellow-feeling in other people. If some classes of pleasures and pains are more susceptible to fellow-feeling than others, the judgements of Smith's impartial spectator will be subject to what a utilitarian would regard as systematic biases. And, according to Smith, some pleasures

[9] This difference is sometimes overlooked. For example, Rawls (1971, pp. 183–92; see also p. 263) considers the claim that a social system is right 'if an ideally sympathetic and impartial spectator would approve it more strongly than any other institution feasible in the circumstances'; the impartial spectator is 'equally responsive to the desires and satisfactions of everyone affected by the social system'. Rawls seems to imply that this Harsanyi-like position is the one taken by Smith and Hume.

and pains *are* more susceptible to fellow-feeling than others. For example, he argues that the 'appetites of the body' for food and sex, although felt intensely at first hand, tend not to induce much fellow-feeling: people find it hard imaginatively to identify with other people's bodily appetites. Similarly, Smith thinks, it is hard to sympathize with physical (as contrasted with emotional) pain. Thus, on Smith's account, public expressions of bodily appetites and of physical pain tend to evoke disapproval (pp. 27–31).[10]

It should not be surprising that Smith's morality of impartial fellow-feeling does not always correspond with Harsanyi's morality of impartial empathy. The fundamental difference, I suggest, is that Smith is offering a naturalistic theory of morality, while Harsanyi is offering a rationalistic one. Smith, unlike Harsanyi, is constrained by the facts of human psychology – by what human beings do and do not have fellow-feeling for, by what they do and do not tend to approve. In a rationalistic theory of morality, it would be odd to propose that moral judgements should reflect the vagaries of human capacities for imaginative identification. But if our moral sentiments are in fact generated by the interplay of fellow-feeling, we must expect them to incorporate whatever systematic biases fellow-feeling really is subject to. How could it be otherwise?

4. THE BOND OF SOCIETY

In his final book, *Trust within Reason*, Martin Hollis tries to find what he calls 'the bond of society' – the body of principles which can explain how societies cohere. The book is built around an elaborate allegory of the 'Enlightenment Trail' – a beautiful walk on which Adam and Eve set out together. By telling us of a succession of side paths leading to pubs of varying qualities, Hollis makes his trail into an embodiment of the backward induction paradox of rational choice theory: only if Adam and Eve trust one another in a way that that theory seems incapable of explaining will they complete the walk and reach that highly desirable pub, *The Triumph of Reason* (Hollis, 1998, pp. 14–18). Surprisingly, however, Hollis never asks why Adam and Eve wanted to walk together (or drink together) in the first place. The fact that human beings so often

[10] This feature of Smith's construction is, I think, overlooked in Darwall's (1999, pp. 141–4) insightful interpretation of the ideal spectator. Darwall says that to take the impartial spectator's viewpoint is to project into a particular person's situation and imaginatively to feel what 'any of us' would feel in that situation, viewing it as that person does. This account leaves room for approval and disapproval by maintaining a crucial space between what the person actually feels and what the impartial spectator imaginatively feels. However, it downplays the distinction between what any of us would feel directly and what any of us would have fellow-feeling for.

choose to do things together is, if not quite a puzzle from the perspective of rational choice theory, at least a regularity that that theory cannot explain. Smith's account of fellow-feeling may perhaps be able to explain it, and in so doing, tell us something about the bond of society.

Imagine that Adam and Eve have driven together to the car park at which the Enlightenment Trail starts. So far, there might be a simple economic explanation of why they have travelled together: by using the same car, let us say, they can share costs. But suppose that, leading from the car park, there is not just one trail but two. One leads into mountains, the other along a seashore. Although the views and challenges of the two trails are very different, each offers an excellent hike. Adam is indifferent between the two, and so is Eve. They might well toss a coin to decide which trail they hike together. But alternatively, Adam might take one trail and Eve the other. Is there anything to be gained by their walking together?

Suppose Adam and Eve care for each other in the way that rational choice theory allows – the way, that is, that Bernheim and Stark's couples do. Then Adam will derive utility from the knowledge that Eve is enjoying her hike, as Eve will do from her knowledge of Adam's enjoyment. In this sense, and given that they both enjoy themselves, both gain from their caring for one another. But this gain can be achieved just as well if they take different trails as if they take the same one; it seems that nothing is added by their walking *together*. If Bernheim and Stark's model is intended to represent how people's caring for one another impacts on their choices, something must be missing. It is surely a characteristic feature of human friendship that friends like to engage in activities together. These are often activities which, on the face of it, might equally well be pursued individually – eating, drinking, watching films, taking walks. But for friends, apparently, added value is created by doing such things together. Where does this added value come from?

Smith's theory offers an answer that I find convincing: the added value arises from the consciousness of fellow-feeling. Thus, two people hiking together can gain pleasure from enjoying the same views, facing the same challenges, and enduring the same discomforts. For this source of pleasure to be tapped, it is necessary that those people's responses to the experiences of the hike are sufficiently aligned. It is no fun to walk through an old-growth forest with someone who thinks one tree is the same as another, or to feel physically exhausted in the company of someone who is not even pleasantly tired. What is required of good hiking companions, I suggest, is not that each prefers that the other's preferences are satisfied, but that they have fellow-feeling with respect to those sentiments – both pleasurable and painful – that are likely to be induced by the experiences of hiking. The psychological mechanisms are

those that Smith describes in the example of reading aloud to a companion: consciousness of fellow-feeling is a source of pleasure in its own right.

In the terminology adopted by Benedetto Gui (1996), fellow-feeling is an essential part of the technology by which *relational goods* – that is, social relations that have subjective but non-instrumental value to the participants – are produced. As a model of a relational good, Gui (2000) offers the 'atmosphere' of a hairdresser's shop, created by friendly interactions between the hairdresser and the customers. This relational good is distinct from the service of having one's hair cut, which conceivably (and with suitable developments in robotics) might be supplied without human contact at all. Although the conversation in such settings may seem trivial and the friendly relations generated may be transitory, I think we can discern in them the significance of fellow-feeling and of the correspondence of sentiments. Think of how, in friendly conversation between casual acquaintances, people try to find topics on which they have common opinions or beliefs.[11] Think, too, of how much easier it is for two strangers to begin a conversation when they can be confident that they have some sentiment in common – say, because they were both hoping to travel on the same inexplicably cancelled train, or because they are both caught in the same violent snowstorm. Human social life is lubricated by the exchange of expressions of corresponding sentiments.

Smith's model of the connections between fellow-feeling, approval and morality may also help to explain why such exchanges – even the apparently inconsequential exchanges of the hairdresser's shop – are important for our sense of well-being. On most plausible accounts, we derive subjective well-being from the sense that our lives are going well for us, that we are being reasonably successful in our pursuit of what we take to be worthwhile goals. But if Smith is right, our sense of what is worthwhile is in part founded on and maintained by the perception of other people's approval: our consciousness of the correspondence of our sentiments with those of others helps us to maintain the sense that our own goals are worth pursuing.

Hollis's (1998, pp. 126–42) answer to his own question about how societies cohere is that people who are fellow members of a community reason collectively, as if each of them was a component part of a single

[11] Smith allows that, among friends, differences of opinion, say about art, literature or philosophy, may be entertaining. But that is only because such topics 'ought all of them to be matters of great indifference to us both; so that, though our opinions may be opposite, our affections may still be very nearly the same' (p. 21). For casual acquaintances, who are not confident of a bedrock of corresponding sentiments, this luxury may not be available.

collective agent. This idea, variously described as *plural subjects*, *we-thinking* or *team thinking*, is alien to the conventional theory of rational choice, in which only individual persons can have the status of decision-making agents. It has been discussed in philosophy by Margaret Gilbert (1989), Susan Hurley (1989), Raimo Tuomela (1995) and Hollis himself, and in economics by Michael Bacharach (1993, 1999) and myself (Sugden, 1993, 2000). I believe that Smith's analysis of fellow-feeling is complementary with the main lines of thought in this literature.

In this literature, there is an emphasis on the conceptual, rational and cognitive aspects of team thinking, rather than the affective. Thus, for example, Gilbert uses the methods of analytical philosophy to elucidate the *concept* of a plural subject. On her analysis, a plural subject comes into existence when the individuals who are to participate in it openly 'express to each other willingness to be part of a plural subject of a certain goal' (1989, p. 17). She argues that it is a conceptual truth that, in expressing such willingness, individuals accept commitments to uphold, in the relevant circumstances, whatever preferences, beliefs or attitudes the plural subject has taken on (1989, p. 162). This kind of approach bypasses questions about whether individuals subjectively *recognize* these commitments, and whether and how they are *motivated* to act on them.

Similarly, Bacharach's and my analyses of team thinking are primarily concerned with its cognitive dimensions. We treat team thinking as a distinctive form of rational choice. The basic idea is that the members of a team recognize some objective as being that of the team; in deciding which action to take as an individual, each member considers which *combination* of actions by team members would best achieve the team's objective, and then performs her part of that combination. The question of what motivates individuals to act in this way is treated as external to the model of team rationality. I have argued that this explanatory strategy is consistent with that followed in conventional rational choice theory, in which preferences are taken as given and the motivation to act on them is not explained (Sugden, 2000). Although true, that doesn't answer the question.

An empirically based analysis of fellow-feeling and of the correspondence of sentiments may be able to fill these gaps by explaining the affective qualities of team thinking. On Smith's account, it is a fact of human psychology that people who repeatedly interact with one another tend to develop and express common sentiments. It is also a fact that such common sentiments tend to become the objects of common approval within the group of interacting people. Thus, the observed failure of any one member of a social group to uphold the attitudes of that group will cause pain or unease to other members (this is just the negative of the pleasure of mutual sympathy); and it will be disapproved

of. So the desire for approval can motivate people to uphold what, on Gilbert's account, are their commitments. The same desire can motivate people to act according to the dictates of what, on Bacharach's and my accounts, is the rationality of team thinking.

For readers who have been trained in conventional rational choice theory, Bacharach's and my theories of team thinking have a particularly unsettling feature: they allow a single individual simultaneously to subscribe to two or more systems of preference, one for each 'team' to which he belongs, but say nothing about the rationality of deciding which of these systems of preferences to act on in any particular situation. Smith's naturalistic approach may help here too, by explaining how people can come to have such multiple preferences. Someone who interacts in several distinct social groups – say, those of her peer group and her family – will be exposed to distinct processes, each of which is tending to induce its own norms of propriety. So long as these spheres of social interaction do not impinge much on one another, there seems to be nothing in the psychology of fellow-feeling to prevent the same person from approving some norm as a member of one social group while also approving a conflicting norm as a member of another. (To adapt one of Smith's examples, a joke may warrant laughter among a group of workmates, or of teenagers hanging out together, while every one of the group would deem it to be unacceptably coarse in the presence of a partner or date.) We have all surely experienced conflicting motivational pulls at the interfaces between the different groups to which we belong.

Smith's official position, it must be said, downplays these possibilities of motivational conflict. He recognizes the 'natural desire to please' as a fundamental property of human psychology, and describes how, when we are young, this naturally leads us 'fondly [to] pursue the impossible and absurd project of gaining the goodwill and approbation of every body'. This project creates just the kind of motivational conflict I have been discussing. However, Smith claims that as mature adults, we recognize the futility of trying to please everyone, and instead consider how our actions would appear to an impartial spectator who had no particular relation to any of the specific people whose approval we are naturally inclined to seek. The viewpoint of the impartial spectator gives us a 'tribunal within the breast' which can sometimes support us in holding out against the general opinion of mankind. Yet, Smith says, if we enquire into the standing of this tribunal, we find that its jurisdiction 'is in a great measure derived from the authority of that very tribunal [i.e., the sentiments of others], whose decisions it so often and so justly reverses' (p. 129). I think there is a fundamental tension here between Smith the social theorist, looking for a naturalistic explanation of actual human sentiments, and Smith the moralist, committed to the virtues of

benevolence, justice and self-command.[12] The moralist in Smith would like to be able to claim that, for all of us, the judgements of the impartial spectators in our respective breasts are the same, irrespective of the experiences to which we have been exposed; but as a social theorist, he explains the impartial spectator as a construct that each of us makes from his own experience. In looking for a coherent reading of Smith, we may sometimes have to choose whether to give priority to his social theory or to his morality, to his assumptions or to his conclusions. My interest is in the social theory.

5. THE INVISIBLE HAND

I have presented a theory of sociality, which I have attributed to Smith. In presenting this theory, I have kept entirely within the domain of sentiments, making no reference to preferences or to choices – the central concepts of rational choice theory. But neither have I made any reference to any of the staple concepts of classical economics, such as production, exchange and the division of labour. Can it possibly be right to suggest that Smith, of all thinkers, would conceive of sociality without economics?

To resolve this paradox, we must understand the strategy of Smith's accounts of spontaneous order. These accounts work in two different ways, which for Smith are complementary. His more usual *bottom-up* method is to start by investigating the facts about human motivation, as we actually observe and experience them. He then draws theoretical implications about the consequences of interactions between individuals who are so motivated, and compares these implications with observations of social life. But he also uses a *top-down* method, which starts from certain presuppositions about the kind of order we should expect to find in human affairs. These presuppositions are deist or functionalist: the universe is governed by natural laws, set in motion by a benevolent creator (sometimes represented as a supreme being, sometimes as 'nature' or as a female person, 'Nature'). He works on the hypothesis – compatible with the biological knowledge of his time – that the 'two great purposes of nature' are 'the support of the individual, and the propagation of the species' (p. 87).

To use one of Smith's favourite metaphors, consider the problem of

[12] This is a manifestation of a more general tension that Chazan (1998) identifies in eighteenth-century philosophical thought. On one side (represented for Chazan by Hume) is the idea that responsiveness to the sentiments of others is natural to human beings, and is a necessary condition for the development of morality and self-esteem. On the other (represented by Rousseau) is an ideal of moral self-sufficiency, liable to corruption by an undue concern about others' opinions. Smith seems to have a foot in each camp.

trying to understand the mechanism of a watch. The bottom-up method is to look at the individual components and to investigate how they interact with one another. The top-down method is to understand the watch as a machine designed to keep track of time. In understanding the workings of a watch, we are likely to do best by using these two methods in combination. Smith thinks that the same is true for understanding the workings of society (p. 87). Thus, while the bottom-up approach takes human psychology as given, the top-down approach allows us to see that the psychological principles on which we act are not arbitrary; they are part of an overall order which has a purpose. Crucially, however, that purpose is Nature's, not some aggregation of the purposes of human beings. What we take to be *our* purposes are components of the order itself, just like the springs of the watch: they are part of the mechanism by which Nature's purpose is achieved.

Smith believes that we are naturally endowed with desires for the achievement of Nature's purposes with respect to our species – that is, for the preservation of human life, and for the propagation of our species, viewed impersonally. But 'it has not been intrusted to the slow and uncertain determinations of our reason' to work out how we as individuals can most effectively further those purposes. They are too important to be left to reason:

> With regard to all those ends which, upon account of their peculiar importance, may be regarded, if such an expression is allowable, as the favourite ends of nature, she has constantly in this manner not only endowed mankind with an appetite for the end which she proposes, but likewise with an appetite for the means by which alone this end can be brought about, for their own sakes, and independent of their tendency to produce it. (p. 77)

Smith gives the obvious examples of hunger, thirst, pain and sexual desire – natural passions which direct us towards those actions that in fact tend to promote our survival and reproduction, without our needing to be conscious of Nature's purpose in so directing us. Similarly, he argues, we have natural passions which direct us towards those forms of behaviour that are necessary to sustain social organization.

To the question of why social organization is necessary – that is, what purpose of Nature's it serves – Smith's answer is economic: man, he says, 'can subsist only in society'. Social organization is necessary to generate the physical security and material wealth that allow the survival and growth of human populations (p. 85). In this sense, then, social cooperation *is* a matter of economics. But precisely because social cooperation is so important for human survival, we have to be so constituted by Nature that we *directly* desire to participate in society, and not merely desire the ends that society achieves.

In part, we are oriented towards social life by innate desires. We are also so constituted that, in the course of behaviour that we are naturally inclined to engage in, further socially-oriented desires reliably develop within us. On Smith's account, we have innate tendencies for fellow-feeling, for feeling pleasure in the correspondence of sentiments, for gratitude and for resentment. His theory explains how, as a result of psychological processes activated by simple human interactions, these innate tendencies induce the more complex moral sentiments of benevolence and justice. Those tendencies and moral sentiments motivate us to participate in society, and to abide by the constraints that social life imposes on us.

In this way, Smith offers an explanation of human sociality that does not depend on assumptions about the instrumental benefits, economic or political, that individuals gain from society. Nor does it depend on assumptions about rational choice. Instead, it depends on assumptions about the natural psychology of fellow-feeling. Society does in fact provide us with instrumental benefits, and as rational beings we can recognize the value of these. But we should be more humble than to suppose that our sociality is a product of our rationality:

> When by natural principles we are led to advance those ends, which a refined and enlightened reason would recommend to us, we are very apt to impute to that reason, as to their efficient cause, the sentiments and actions by which we advance those ends, and to imagine that to be the wisdom of man, which in reality is the wisdom of God. (p. 87)

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

Bentham, Jeremy. 1789/1970. *An Introduction to the Principles of Morals and Legislation*. Athlone Press

Bernheim, R. Douglas and Oded Stark. 1988. 'Altruism within the family reconsidered: do nice guys finish last?' *American Economic Review*, 78:1034–45

Binmore, Ken. 1994. *Game Theory and the Social Contract. Volume 1: Playing Fair*. MIT Press

Binmore, Ken. 1998. *Game Theory and the Social Contract. Volume 2: Just Playing*. MIT Press

Chazan, Pauline. 1998. *The Moral Self*. Routledge

Darwall, Stephen. 1999. 'Sympathetic liberalism: recent work on Adam Smith'. *Philosophy and Public Affairs*, 28:139–64

Edgeworth, Francis. 1881. *Mathematical Psychics*. Kegan Paul

Fontaine, Philippe. 1997. 'Identification and economic behavior: sympathy and empathy in historical perspective'. *Economics and Philosophy*, 13:261–80

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

Gui, Benedetto. 1996. 'On relational goods: strategic implications of investment in relationships'. *International Journal of Social Economics*, 23:260–78

Gui, Benedetto. 2000. 'Beyond transactions: on the interpersonal dimension of economic reality'. *Annals of Public and Cooperative Economics*, 71:139–69

Harsanyi, John. 1955. 'Cardinal welfare, individualistic ethics and interpersonal comparisons of utility'. *Journal of Political Economy*, 63:309–21
Hollis, Martin. 1998. *Trust within Reason*. Cambridge University Press
Hollis, Martin and Robert Sugden. 1993. 'Rationality in action'. *Mind*, 102:1–35
Hume, David. 1740/1978. *A Treatise of Human Nature*. Oxford University Press
Hurley, Susan. 1989. *Natural Reasons*. Oxford University Press
Kahneman, Daniel. 1996. 'Comment' [on paper by Charles Plott]. In *The Rational Foundations of Economic Behaviour*, pp. 251–4. Kenneth Arrow *et al.* (eds.). Macmillan
Lyons, David. 1973. *In the Interest of the Governed*. Oxford University Press
Pareto, Vilfredo. 1909/ 1972. *Manual of Political Economy*. Trans. A. S. Schweir. Macmillan
Rawls, John. 1971. *A Theory of Justice*. Harvard University Press
Samuelson, Paul. 1947. *Foundations of Economic Analysis*. Harvard University Press
Savage, Leonard. 1954. *The Foundations of Statistics*. John Wiley
Smith, Adam. 1759/1976. *The Theory of Moral Sentiments*. Oxford University Press
Smith, Adam. 1776/1976. *An Inquiry into the Nature and Causes of the Wealth of Nations*. Oxford University Press
Sugden, Robert. 1993. 'Thinking as a team: towards an explanation of non-selfish behavior'. *Social Philosophy and Policy*, 10:69–89
Sugden, Robert. 2000. 'Team preferences'. *Economics and Philosophy*, 16:175–204
Tuomela, Raimo. 1995. *The Importance of Us*. Stanford University Press.