Intersubjectivity, *The Theory of Moral Sentiments* and the Prisoners' Dilemma

Vivienne Brown
The Open University
v.w.brown@open.ac.uk

published in *The Adam Smith Review* (2011) 6: 172-190, (Ed.) Fonna Forman-Barzilai.

I

One of the remarkable transformations in Adam Smith scholarship since the publication of the Glasgow Edition of the Works and Correspondence of Adam Smith initiated in the 1970s is the high and growing level of interest in *The Theory of Moral Sentiments* (TMS, 1976). It might have been expected that philosophers would be drawn to this work, but economists and other social scientists have been discovering its insights and have applied them to a range of social and economic issues. Coinciding with new developments in behavioural economics, game theory, institutional economics and evolutionary biology – and perhaps even contributing to some of them – this new interest in TMS has also stimulated attempts to rethink and even attempt to overcome some of the institutionalized disciplinary boundaries between the human sciences.

In this essay I would like to take up just one aspect of this wider application of the TMS to social and economic issues—the simultaneous one-shot Prisoners' Dilemma game. This game was first formally developed in 1949/50 although some scholars argue that there are important precursors in the writings of, for example, Hobbes and Hume.[1] The game has been taken to express in stark form the conflict between individual and collective interest: individual rationality construes cooperative play as self-sacrificial and so rational players do not cooperate, even though both players are worse off not cooperating than if they had both cooperated. As an *n*-person game it has come to represent the class of 'social dilemmas' in which society loses out because the application of individual rationality leads individuals not to cooperate, even though all would benefit from cooperation. Important examples of social dilemmas are given by the arms race, degradation of the environment, and human contribution to climate change. The basic two-person Prisoners' Dilemma game has thus come to symbolise the wider significance of social dilemmas in an increasingly interconnected world.

The Prisoners' Dilemma is also controversial amongst game theorists. Critics argue that game-theoretic reasoning fails both normatively and explanatorily for the simultaneous one-shot Prisoners' Dilemma, in that it recommends strategies that lead to a Pareto-inefficient outcome and it fails to explain why a significant proportion of players do in fact cooperate (eg Bacharach 2006, Gold and Sugden 2007). Defenders respond that such criticism misses the point of the game which is that it illustrates how individual rationality can lead to Pareto-inefficiency. They thus insist that cooperative play is irrational in a simultaneous one-shot Prisoners' Dilemma even though it leads to beneficial results if both players cooperate (eg Binmore 1994). It is into this controversy that TMS has been introduced by the critics in order to provide refinements on the way that preferences and choices are modelled in the Prisoners' Dilemma game, and to show thereby how cooperation might be explained as rational in terms of this more rounded view of human behaviour and motivation (eg V.L. Smith 2010). If defenders of the Prisoners' Dilemma game are correct, however, such applications of TMS miss their mark and suggest that TMS is being put to service in a losing argument.

These are the issues that I wish to take up in this essay. I consider the debate between critics and defenders of the traditional logic of the simultaneous one-shot Prisoners' Dilemma game together with the place of TMS in this debate. I argue that what is needed for this debate is a new approach to individual practical reasoning and that existing applications of TMS to the Prisoners' Dilemma have neglected a fundamental aspect of TMS which can help to provide conceptual resources for developing such an approach. I thus enrol TMS on the side of the critics, in that I try to show that cooperation can be rational, but I also aim to develop a new and better response to the defenders of the Prisoners' Dilemma. I thus hope to provide a new way of dissolving what otherwise seems to be a dichotomy between individual and collective interests.

Section II describes the Prisoners' Dilemma game and critically examine how the insights of TMS have been brought to bear in explaining how cooperation might be conceived as rational behaviour, and Section III argues that the resources of TMS suggest a new approach that answers to the criticisms raised in Section II. Section IV concludes the essay.

II

In this essay I shall focus on a 2-person simultaneous one-shot Prisoners' Dilemma. The game is often illustrated in terms of two prisoners taken in for interrogation in connection with a crime that they are suspected of having committed together, but the logic of the game is quite general and that is how I'd like to consider it here.

In this simple version the game is played just once and with two players, whom I'm calling Player 1 and Player 2. The aim of the game for each player is to maximize the payoff. Each player has a choice of two strategies which I'm calling *cooperate* and *not cooperate*. As there are 2 possible strategies for each of the 2 players, there are 4 possible pairs of strategies altogether: (*cooperate*, *cooperate*), (*not cooperate*, *cooperate*), (*cooperate*, *not cooperate*) and (*not cooperate*, *not cooperate*), where the first strategy listed in the pair is that of Player 1 and the second is that of Player 2. The payoffs are symmetrical for the two players and give the value of each player's chosen strategy given the other player's chosen strategy. The standard assumptions of rationality and common knowledge apply, so each of the players knows all the payoffs and knows that the other player knows them, and each of the players knows that the other knows that he knows them, and so on. The players choose their strategies simultaneously. The structure of payoffs is given in Figure 1, with Player 1 as the row player and Player 2 as the column player; and some illustrative values are provided in Figure 2:

Figure 1: The Prisoners' Dilemma (payoff structure)

		Player 2	
		cooperate	*not cooperate*
Player 1	*cooperate*	b, b	d, a
	not cooperate	a, d	c, c

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

Figure 2: The Prisoners' Dilemma (illustrative payoffs)

		Player 2	
		cooperate	*not cooperate*
Player 1	*cooperate*	3, 3	1, 4
	not cooperate	4, 1	2, 2

According to a form of reasoning employed in game theory and decision theory, known as the principle of dominance, Player 1 reasons as follows: 'if Player 2 chooses *cooperate*, my payoff for *not cooperate* is higher (4) than for *cooperate* (3); and if Player 2 chooses *not cooperate*, my payoff for *not cooperate* is higher (2) than for *cooperate* (1). Thus, whichever strategy Player 2 chooses, my payoff to *not cooperate* is higher than my payoff to *cooperate*. Therefore I choose the strategy *not cooperate*.' Player 1's strategy *not cooperate* is said to be the (here, strictly) dominant strategy because it yields a better (that is, not just a no-worse) payoff regardless of which strategy Player 2 chooses. Playing *cooperate* is thus construed as 'self-sacrificial'. As the game is symmetrical, Player 2 reasons along the same lines and chooses the strategy *not cooperate*. Thus, *not cooperate* is the strictly dominant strategy for both players: no rational player would choose to cooperate. If both play *not cooperate* the outcome is the strategy pair (*not cooperate*, *not cooperate*), with payoffs (2, 2). This is the Nash equilibrium at which each player's chosen strategy is a best reply to the other player's chosen strategy (neither would regret the strategy chosen, given the other player's chosen strategy). If both players had chosen (*cooperate*, *cooperate*), however, the payoffs would have been (3, 3). In the Prisoners' Dilemma game, the Nash equilibrium is not Pareto-efficient: reasoning individually according to the dominance principle does not result in the best outcome for the players.

This reasoning is unaffected by a player's 'trust' in the other to cooperate, or by prior communication between the players. Even if a player trusts the other to cooperate, or even if there is prior communication or a prior agreement with the other player to cooperate, it is still the case that *not cooperate* yields a higher payoff to that player. Neither player therefore has any reason to trust the other to cooperate, and prior communication is merely 'cheap talk' as it has no effect on which strategy gives the higher payoff and hence no effect on the conclusion to play *not cooperate*.

Applications of this game are manifold. Construed in terms of an *n*-person game, we can see examples of it in terms of the arms race, environmental resource use, environmental recycling, and voting behaviour. These are situations in which players benefit if all play cooperatively, but an individual player benefits even more – so the reasoning goes – if all others play cooperatively and he is the only one not to. The trouble is that if all players reason and act in that way, the outcome is worse for everyone.

The Prisoners' Dilemma game has been widely debated on account of this apparent conflict between individual and collective interests. How can it be that rational behaviour leads to a Pareto-inefficient outcome whereas it is irrational behaviour that leads to Pareto efficiency? Furthermore, there is significant evidence that a substantial proportion of people do behave cooperatively. There is everyday evidence that individuals do cooperate to some degree in many contexts; for example, many people do take the trouble to vote, many people do take the trouble to recycle their rubbish, and so on. There is also a large experimental literature on the Prisoners' Dilemma. Here the consensus seems to be that players cooperate about half the time in simultaneous one-shot Prisoners' Dilemma games (Camerer 2003: 46). Must we conclude that these cooperators are behaving irrationally?

This simultaneous one-shot version of the game is both the simplest version of the game and the hardest in terms of explaining why the experimental evidence shows a substantial level of cooperation. It is harder to explain cooperation in a simultaneous one-shot game than in a repeated game because punishment and reputation effects, which can explain cooperation in repeated play as rational, have no bearing if the game is played just once since their effectiveness relies on there being future rounds of play. Defenders of the traditional logic of the game argue that cooperative play in a simultaneous one-shot game is irrational, and that players' cooperation in such games is to be explained either by a lack of experience at playing such a game or by the players' importing norms of cooperation from everyday life where threats of future punishment provide incentives to current cooperation (Binmore 1994, 2006). For critics, the challenge is to explain how cooperative play in the simultaneous one-shot Prisoners' Dilemma can be justified or explained in some way. They argue that the standard approach misconstrues the nature of human preferences and motivations, and hence mis-specifies the payoffs of the game. Traditionally, payoffs in game theory have been construed in terms of material payoffs that answer to players' self-interest. Critics of this traditional approach to game theory argue that this is too narrow in overlooking the relevance of non-material payoffs and the significance of attitudes towards other players' payoffs (eg Camerer, 2003, ch. 2 for survey). They re-theorize the objects of preference to include not only material benefits to a player, but also non-material items involving social or moral considerations such as positive or negative feelings associated with cooperating or not, sympathetic feelings for other players, and moral or psychological factors concerning fairness or social norms. Players may thus be other-regarding as well as self-regarding, and utility maximization becomes maximization of 'social utility' (or 'social preferences') such that players attach value to a range of factors, including even cooperation itself. The resulting payoff transformation in the case of the Prisoners' Dilemma implies that *cooperate* can dominate *not cooperate*, thus making *cooperate* the dominant strategy. For such maximizers of social utility, it is rational to cooperate when such payoff transformation makes *cooperate* the dominant strategy.

It is in developing this payoff transformation approach to the Prisoners' Dilemma that some theorists have found inspiration in the rich account of social and moral life in TMS, and in the account of sympathy and the moral sentiments presented there. Robert Frank (1988, 2004, 2007) focuses on the moral sentiments and argues that moral sentiments such as guilt provide 'precommitment' for cooperative behaviour by changing the structure of payoffs. For example, a person who would feel guilty from not cooperating would experience a lower real payoff to *not cooperate* than would otherwise be the case because the negative feeling of guilt would have to be set against the material payoff. If this guilt effect were sufficiently substantial it would result in a lower payoff to *not cooperate* than to *cooperate*. This is illustrated in Figure 3 where the payoff to *not cooperate* is reduced by 2 as compared with Figure 2:

Figure 3: Payoff transformation: reduced payoff from *not cooperate* arising from (negative) moral sentiment of guilt

Player 2

cooperate *not cooperate*

Player 1	*cooperate*	3, 3	1, 2
	not cooperate	2, 1	0, 0

Applying the principle of dominance each player reasons as follows: ‘if the other Player chooses *cooperate*, my payoff for *cooperate* is higher (3) than for *not cooperate* (2); and if the other Player chooses *not cooperate*, my payoff for *cooperate* is higher (1) than for *not cooperate* (0). Thus, whichever strategy the other Player chooses, my payoff to *cooperate* is higher than my payoff to *not cooperate*. Therefore I choose the strategy *cooperate*.’ Payoff transformation thus changes the dominant strategy from *not cooperate* to *cooperate*, so it now becomes rational to cooperate. The outcome (*cooperate*, *cooperate*) is the new Nash equilibrium.

Thus, Frank argues, moral emotions facilitate mutual cooperation even in a simultaneous one-shot Prisoners’ Dilemma, and in this are consistent with, perhaps even necessary for, the pursuit of self-interest. In addition to the importance of moral sentiments, Frank also argues that the forging of sympathetic bonds, for example in pre-play communication, is important in communicating intentions and influencing cooperative behaviour. In an experiment that he reports, almost 74% of people cooperated in a one-shot Prisoners’ Dilemma when they had spent 30 minutes talking to each other prior to the game (Frank 2007: 206). This is in contrast to the standard game-theoretic view that prior communication is ‘cheap talk’ in not affecting the rational strategy of not cooperating.

David Sally (2000) draws on the TMS to argue that the Prisoners’ Dilemma game is changed by how it is perceived by the players. If it is perceived as a game played with others who are similar, familiar or valued positively – with others for whom sympathy is felt – then it is played with an enlarged sense of self and a correspondingly broader range of payoffs. Whether or not Sally’s notion of sympathy is the same as Adam Smith’s – Sally goes on to identify sympathy in terms of ‘reading the mind of another’ (pp. 570, 572) – what he draws from TMS is a socialised picture of human communication that lends psychological support to his argument that the payoffs of the game need to register an ‘enlarged self-interest’ that includes others’ interests too (eg pp. 571, 574, 575). Sally argues that the process of interpersonal identification ‘loosens the boundary of the self’ and leads to ‘an expansion of the self’ so that there is ‘some overlap’ with another person (pp. 571, 574). This overlap of the sympathetic self with another implies a ‘sympathetic utility function’ such that ‘other-oriented action may be motivated by the same basic self-concern as most other actions. Simply put, I help another because the other is part of myself’ (pp. 599–604, 574). This implies that the payoff to *cooperate* is increased because in helping another I help myself. As with Frank, changing the preferences leads to a different payoff matrix (pp. 570, 575). This is illustrated in Figure 4 where the payoff to *cooperate* is increased by 2 as compared with Figure 2:

Figure 4: Payoff transformation: additional benefit from *cooperate* arising from other-regarding sympathy

		Player 2	
		cooperate	*not cooperate*
Player 1	*cooperate*	5, 5	3, 4
	not cooperate	4, 3	2, 2

Again, with payoff transformation *cooperate* is the dominant strategy and so individual reasoning according to the dominance principle results in the (*cooperate*, *cooperate*) outcome. Each player finds that, whichever strategy is chosen by the other player, *cooperate* has a higher payoff than *not cooperate* (5>4; 3>2). Again, the outcome (*cooperate*, *cooperate*) is the new Nash equilibrium.

Both Frank and Sally thus provide instances of payoff transformation inspired by Smith's TMS. According to these interpretations, adding more objects of preference – in connection with mutual sympathy, moral emotions or social norms – supports cooperation where payoff transformation makes *cooperate* the dominant strategy. These interpretations thus introduce additional behavioural complexity into game theory and emphasise the importance of non-material items in players' utility functions. They might also shed some light on experimental results where players are working to an expanded set of preferences: if players are cooperating because their non-material together with their material preferences result in *cooperate* being the dominant strategy, then such players are being rational in cooperating. Postulating that players are maximizing social utility preserves the traditional logic of the game whilst providing an explanation of apparent experimental anomalies.

But the theoretical weakness of such payoff transformation is that it does not answer to the challenge of the Prisoners' Dilemma game because payoff transformation changes the game so that it is no longer an instance of the Prisoners' Dilemma (Binmore 1994: 102–14). Arguing that players are rational in cooperating when *cooperate* is the dominant strategy thus does not bear on the question whether it is rational to cooperate in a Prisoners' Dilemma game; and integrating moral sentiments and sympathy into the Nash equilibrium, as Sally explicitly recommends (2000: 575), cannot constitute a challenge to thinking about the game in terms of the Nash equilibrium. The payoff transformation approach to the Prisoners' Dilemma therefore does not challenge the traditional logic of the game, and incorporating Smith's insights from the TMS does not alter that.

This is also evident in another influential adoption of TMS in explaining cooperative behaviour. Herbert Gintis et al. (2005a, b) draw on the insights of TMS to explain cooperation but here it is recognised that doing so goes beyond the Prisoners' Dilemma. Gintis et al. argue that Smithian sympathy is important in understanding cooperation in human evolution and that this can be modelled using game theory. They argue that cooperation can be explained by the presence of players who are 'strong reciprocators', that is, players who are altruistically predisposed to cooperate with others whom they expect to cooperate, but who will punish, even at a cost to themselves, those who do not cooperate in reiterated play. The altruism of such 'conditional cooperators' is thus held to explain cooperation even in simultaneous one-shot games. This reasoning accords with the standard logic of the Prisoners' Dilemma that it is self-sacrificial to cooperate, even if the other player is expected to cooperate, but Gintis et al. argue that some players are predisposed to such 'altruistic behaviour'. These theorists, however, recognise that this changes the game, in this case from a Prisoners' Dilemma game to a coordination game with two equilibria (*cooperate*, *cooperate*; *not-cooperate*, *not-cooperate*) (Fehr and Fischbacher 2005).[2] In such a coordination game, if strong reciprocators believe that the other player will cooperate, then (*cooperate, cooperate*) is the equilibrium outcome.

As with Frank and Sally, Gintis et al.'s attempt to explain cooperation accepts the standard logic of the Prisoners' Dilemma game that it is self-sacrificial to cooperate if *not-cooperate* is the dominant strategy. To challenge the standard logic of the game that it is self-sacrificial to cooperate we need to show how a *different form of individual reasoning* could lead players to cooperate. [3]

III

How might we show this? It seems to me that this is where the conception of human agency in TMS has something important to teach us that is not picked up by Frank, Sally or Gintis et al.. TMS is generally understood primarily in terms of its account of sympathy, moral sentiments and the making of moral judgments. From this perspective its account of human sympathy is essentially an exploration of the sympathetic dimension of social relations and the ways in which this influences the nature of morality and the making of moral judgments. But the investigation into social life in TMS also presents a distinctive view of the nature of individual human agency itself. A core aspect of TMS is that the nature of individual human agency is constituted by the spectatoriality of social life. In TMS all are spectators to each other: each person is a spectator to others, and these others are at the same time spectators to that person. Human agency is thus construed in terms of an overarching spectatoriality in which mankind lives 'in the eyes of the world', 'in the view of the public', and 'open to the eyes of all mankind' (TMS I.iii.1.15, II.iii.3.2; I.iii.2.1, V.2.10; I.iii.2.1). The notion of spectatorship might seem to be a passive one in that it denotes onlookers rather than agents, but in TMS it is an active notion in denoting the way that all are simultaneously spectators and spectated upon. All human life – from the immediate family, to circles of friends and acquaintances, to public life – is presented as an arena within which individuals live in the eye of others.

This omnipresent spectatoriality has important implications for the theorisation of the individual in that it provides an *intersubjective* account of individuality.[4] According to this spectatorial model, an individual human being is incomplete without others since, if human beings live 'in the eyes of the world', without the eyes of others the life that is lived is not fully human. Without those eyes there is a lack of something that is constitutive of human life. We can see the importance of this intersubjective individuality in three main areas of TMS.

First, social emulation and material self-betterment is presented in terms of the need to appear well in the eyes of the world. Smith asks from whence 'arises that emulation which runs through all the different ranks of men, and what are the advantages which we propose by that great purpose of human life which we call bettering our condition?' (I.iii.2.1). The answer is: 'To be observed, to be attended to, to be taken notice of'. It is 'vanity', Smith says, 'not the ease, or the pleasure, which interests us'. But, he adds, 'vanity is always founded upon the belief of our being the object of attention and approbation'. Social emulation and material self-betterment thus presuppose intersubjectively constituted individuality.

Second, sympathetic responses to others are construed in terms of spectatorial relations among intersubjectively constituted individuals. Fellow-feeling or sympathetic responses by spectators are presented as a necessary part of socialised human nature, a fundamental component of the pleasures of social life and human interaction. Spectators experience pleasure in sympathising with others and in having those others, as their spectators, sympathise with them in turn; and because of this human need to experience the sympathy of spectators, those being looked upon try to modulate their feelings to elicit the sympathy of their spectators, and they do this by imagining how their spectators view them. The outcome is a series of sympathetic and modulated feelings which provide the basis for the sociality and relational pleasures of human life.

Third, morality is made possible by society since it is society that provides the initial moral looking-glass by which people are able to evaluate themselves. Without this moral mirror of society, morality and critical self-awareness would not be possible (III.1.3). People would be as unaware of the moral qualities of their mind and character as they would of the aesthetic qualities of their facial features. It is by viewing themselves 'with the eyes of other people' that they are able to judge themselves (III.1.2, 4, 5; III.2.3). The possibility of moral agency is thus constituted intersubjectively, both in first-person and third-person cases. In first-person cases, the person imagines the extent to which an impartial spectator to himself would be able to share and, hence, approve his sentiments and conduct. To the extent that the agent is able to imagine that such a spectator would share and hence approve his sentiments and conduct, to that extent he is able to judge favourably of his own conduct (III.1.6). Moral judgment is construed in terms of a spectatorial relation in which the imagined spectator is invested with a degree of impartiality that would otherwise be unavailable to the agent in scrutinising himself. This model of moral judgment is applied in the case of third-person judgments; a spectator judges another to the extent that he can share the sentiments of the other and hence sympathise with what he imagines are those sentiments (I.i.3, 4). To the extent that the spectator is able to exercise impartial judgment upon himself, rather than simply reflect conventional mores, to that extent he is also able to exercise considered moral judgment on others. There is thus both a social gaze as well as a moral gaze.[5] The social gaze reflects conventional values including the ambitious emulation of spectators in society. The moral gaze embodies impartial moral sentiments and is the foundation of moral judgment proper.[6] The portrayal of the impartial spectator as a metaphor of moral judgment thus builds upon the portrayal of the spectatoriality of human life; it is only in being imagined as impartial that the moral judge is different, not in being a spectator to oneself.

An intersubjective conception of individuality is presupposed in these accounts of self-awareness, sentiments and judgment. This intersubjectivity is an existential characteristic of individuals in TMS and so it is registered at a more fundamental level than the formation of specific sentiments and sympathies. It is thus present whether or not agents experience particular moral sentiments on particular occasions (such as Frank's sentiment of guilt in situations of non-cooperation), whether or not agents feel that other players are especially valued (as in Sally's account of enlarged self-interest and expansion of the boundaries of the self for especially valued others), and whether or not agents are predisoposed towards altruism (as in Gintis et al.'s strong reciprocity).The intersubjectivity that I am identifying would still be present even in cases of inappropriate sentiments, low valuation of other players, or absence of any predisposition to altruism; that is, it would be present even in the absence of what scholars have suggested to be the specific preference adjustments that would sustain payoff transformation in the case of the simultaneous one-shot Prisoners' Dilemma. It is thus independent of any particular preference formation or moral (or immoral) qualities, and for this reason it pertains even in cases of the narrowest and most selfish self-interest. It is in this sense that I characterise such intersubjectivity as an existential characteristic of human individuality in TMS.

It seems to me that this conception of intersubjectivity introduces the possibility of theorising individuality in a different way. This conception of intersubjectivity is in contrast with the classical liberal notion of the individual agent who lives in a sort of external relation to society and to others. This may be illustrated by comparing the classical liberal metaphor of the individual with that of Smith's spectatorial metaphor. The classical liberal metaphor is that an individual agent occupies a 'space' of thought and action that is intimately his own and which is to be protected against others. In clarification of this, it might be said that individual preferences and self-images – as given within the domain of the 'private' – are in fact influenced by the individual's association with others, and that it is only within the 'public sphere' that the image of the separable individual really comes into play as one who claims ownership of that individual space. By contrast, however, Smith's spectatoriality of human life erodes notions of separateness or spatial distance in that individuals' spectatorially-based conceptions of themselves are not independent of how they imagine others are viewing them. It nibbles away at the distinction between the private and the public sphere by making both susceptible to the eyes of others, whether the eyes are of family members or acquaintances, or indeed 'the eyes of the world' or 'the eyes of all mankind'. Thus the classical liberal notion of 'association with others' is displaced by the notion of spectatoriality which erodes somewhat the distinctions of inside and outside, private and public, which the classical liberal notion of individuality tends to rely upon.[7] By this means Smith's account in TMS facilitates an understanding of individuality that repositions it within the inter-relatedness of all human life, the most private and personal as well as the most public. This is not to detract from another aspect of classical liberal thinking, which is evident in Smith's writing, that the 'freedom' of the individual is something to be respected and preserved. Furthermore, Smith's account of moral judgment and moral behaviour in TMS is premised upon freedom of individual choice (Brown 2009). Yet, the individual whose freedom is so important to Smith's analysis is one who is constituted intersubjectively in TMS.[8]

Smith's account of intersubjectivity thus provides conceptual resources for moving beyond the dichotomy of the 'individual' and the 'collective' that runs through much of liberal thought and much of game theory too. It holds out the promise of developing a conceptual space for understanding a form of individual agency that is compatible with cooperation in pursuing individual interests; and this holds whether those individual interests are construed in terms of self-regarding or other-regarding behaviours.

This intersubjective approach to individual agency suggests that for game-theoretic purposes we need a different mode of individual *reasoning*, not a different set of individual *preferences*. In developing such a mode of reasoning, I adapt the notion of a schema of instrumental practical reasoning as presented in Gold and Sugden's edited version of Michael Bacharach's uncompleted manuscript, *Beyond Individual Choice* (Gold and Sugden 2006, 2007; Bacharach 2006).[9] Such a schema of instrumental practical reasoning generates normative statements of what should be done as the conclusion derived from a series of premises; the validity of such a schema is construed in terms of its success in generating the best outcome for the reasoning player. Schema 1 gives a schema of practical reasoning in the case of individual rationality construed in terms of what 'I' as an agent should do in trying to achieve the best outcome (Gold and Sugden 2006: 156–7). As I present it, this schema also includes the possibility of broad payoffs (material plus non-material payoffs) as well as narrow payoffs (material payoffs), thus including the possibility of a social payoff function including, for example, other-regarding factors. I include these here to show that the schema of reasoning is independent of the actual content of preferences.

Schema 1: Individual Rationality

(1) I must choose either *A* or *B*.
(2) If I choose *A*, the outcome will be O_1 (narrow or broad payoff).
(3) If I choose *B*, the outcome will be O_2 (narrow or broad payoff).
(4) I want to achieve O_1 more than I want to achieve O_2

I should choose *A*.

Schema 1, individual rationality, is set up in terms of what 'I' should do to achieve 'my' objective: 'I' am an agent in pursuit of 'my' objective *O*, (where O_1 is more preferred than O_2), and this holds whether or not objectives include non-material or other-regarding ones. It is immaterial whether objectives are formed as a result of social influences, affective relations, or moral criteria. As choosing *A* yields the more preferred outcome for 'me', then *A* is what 'I' should choose.

If this reasoning is applied to the Prisoners' Dilemma, the agent aims to maximize the payoff, *P* (narrow or broad), and the reasoning is consistent with the principle of dominance. This is shown in Schema 2:

Schema 2: Individual Rationality in the Prisoners' Dilemma

(1) I must choose either *cooperate* or *not cooperate*.
(2) If the other player chooses *cooperate*, the outcome will be P_1 (narrow or broad payoff) if I don't cooperate and P_2 (narrow or broad payoff) if I cooperate.
(3) If the other player chooses *not cooperate*, the outcome will be P_3 (narrow or broad payoff) if I don't cooperate and P_4 (narrow or broad payoff) if I cooperate.
(4) I want to achieve P_1 more than I want to achieve P_2, and I want to achieve P_3 more than I want to achieve P_4

I should choose *not cooperate*.

The conclusion that 'I' should choose *not cooperate* holds even if the payoffs are construed broadly to include non-material or other-regarding benefits, as long as the payoff structure is consistent with the Prisoners' Dilemma. Cooperative behaviour would thus require changing *P* to make *cooperate* the dominant strategy; this is Frank's and Sally's approach.

But Smith's account of intersubjectivity is suggestive of a different form of individual reasoning because it has a different conception of individual agency. In TMS individual agency is formed 'in the eye of others' and this implies that individuality presupposes omnipresent others whatever the agent's preferences or dispositions: for each individual 'I' there is always another 'I' (or other 'I's) already present. When individual agents aim to maximize on individual objectives, each maximizing 'I' acknowledges the presence of 'You', who is another maximizing 'I' (or other maximizing 'I's). Thus when reasoning to maximize on individual objectives, the individual 'I' includes 'You' within the practical reasoning employed. Schema 3 attempts to capture this alternative mode of individual reasoning for the Prisoners' Dilemma where individual rationality explicitly takes account of intersubjectivity:

Schema 3: Intersubjective Rationality in the Prisoners' Dilemma

(1) I must choose *cooperate* or *not cooperate*, and you must choose *cooperate* or *not cooperate*.
(2) I want the best outcome for myself, and you want the best outcome for yourself.
(3) The outcome for me of my choice of strategy depends upon your choice of strategy, just as the outcome for you of your choice of strategy depends upon my choice of strategy.
(4) The best outcome for me, given that you are a maximizing agent, is when you and I cooperate; and the best outcome for you, given that I am a maximizing agent, is when you and I cooperate.

I should choose *cooperate*, and so should you.

Schema 3 provides an example of intersubjective rationality in the case of the Prisoners' Dilemma. The agent who reasons and maximizes here is an individual 'I'. Nonetheless, this 'I' acknowledges that there is a 'You', who also has to choose a strategy (premise 1) and who also aims to maximize on individual objectives (premise 2). With this inclusion of 'You', the 'I' acknowledges not only the constraints on what either can achieve (premise 3) but also the existence of some congruence of individual interests between them (premise 4). Intersubjective rationality therefore concludes that 'I should choose *cooperate*, and so should you', because that is what yields the best that a player can achieve individually given the presence of another maximizing 'I'.

Schema 3, intersubjective rationality, makes possible what I term 'instrumental cooperation'. As with Schemata 1 and 2, Schema 3 is posed in terms of individual reasoning and individual maximization, but unlike Schemata 1 and 2 it can exploit the individually beneficial potential for cooperation. Importantly, this shift in agential focus construes instrumental cooperation in terms of realising the potential benefits afforded by congruence of individual interests, instead of construing individual cooperative behaviour as 'self-sacrificial'. Cooperative behaviour is thus retheorised as instrumentally beneficial instead of being self-sacrificial. Concluding that 'I and you' – that is, 'each of us' – should cooperate is thus not to gift the other player an advantage, but to reason to the best possible individual payoffs given the structure of the game.

According to Schema 3, then, it is rational to cooperate: cooperation is the strategy that is recommended by the instrumental practical reasoning of Schema 3. This is in contrast to the standard interpretation of the Prisoners' Dilemma where individual practical reasoning recommends non-cooperation. This difference may be illustrated by considering the issue of 'trust', where trust is the subjective probability that a player will cooperate. Intersubjective practical reasoning promotes cooperation because it facilitates *mutual trust* between the players. The reason for this is that, as intersubjective reasoning concludes that it is in the individual interest of each of the players to cooperate, both players have good reason to trust that each of them will indeed cooperate. It is because it is individually rational for each to cooperate that mutual trust between them is facilitated. By contrast, in the case of the traditional interpretation of the Prisoners' Dilemma, individual rationality concludes that it is self-sacrificial to cooperate. In concluding that it is irrational to cooperate, it promotes distrust.

Schema 3 thus rationalises instrumental cooperation and thereby promotes mutual trust between the players. In asking, 'what is the best strategy for me, given that you and I each have to choose whether to cooperate?', it seeks a solution that takes due account of the fact that each of the players is reasoning instrumentally in the presence of both constraints and some congruence of interests. This is in contrast to the traditional game-theoretic approach to individual reasoning in the Prisoners' Dilemma, which, in asking, 'what is the best strategy for me, given the other player's strategy?', fails to take into account the implications of having both players reasoning this way, a failure that leads to a Pareto-inefficient outcome. Schema 3 is thus superior to Schema 2, as evidenced by the fact that its conclusion implies higher payoffs for the players.

Normatively, intersubjective rationality thus performs better than individual application of the dominance principle in yielding better payoffs. *Explanatorily*, intersubjective rationality can provide new insights into the existing evidence on cooperative behaviour in experimental situations. One aspect of the experimental evidence that has received considerable support is that prior communication tends to result in greater cooperation, as both Frank (2007: 206) and Sally (1995: 80; 2000: 612–3) report. This is inexplicable in rational terms according to standard game theory according to which such communication is mere 'cheap talk' that does nothing to dislodge *not cooperate* as the rational strategy. The significance of this prior communication may be explained, not in terms of preference adjustment, but in terms of promoting intersubjective rationality which leads players to the conclusion that 'each of us' should cooperate. Experimenters also argue that players are more likely to cooperate if they believe the other player will cooperate, even though this also goes against the standard interpretation, as Frank (2007: 206) and Fehr and Fischbacher (2005: 165) report. This result could be explained in terms of players assuming that the other player is also intersubjectively rational. Experiments and wider social observations also note the importance of norms of cooperation in influencing outcomes in Prisoners' Dilemma games. A difficulty with making sense of this in terms of the traditional understanding of the game is that such norms would have to outweigh what is deemed to be the self-sacrificial nature of cooperation, but if such norms are understood as conducing towards intersubjective rationality they can readily be theorised in explaining cooperative behaviour.

In positing socialised agents as beings that live outside themselves to some degree – in the eyes of the world that are looking on them – Adam Smith's TMS provides a model of agency that supplies conceptual resources for supporting the intersubjective rationality of Schema 3.[10] Smith's spectatorial account of human agency illustrates how individual human agency internalises the presence of another 'I': individuals view themselves from the standpoint of others as well as viewing others from their own standpoint, a multiperspectival viewpoint that facilitates intersubjective rationality. Intersubjectivity is thus a characteristic of human individuality; it makes morality possible, just as it makes immorality possible too, according to Smith's account.

Instrumental cooperation in the Prisoners' Dilemma is independent of issues of fairness or morality. Neither Frank's emphasis on the function of moral sentiments, nor Sally's emphasis on the significance of sympathy for those other players who are especially valued, has any essential part in the intersubjectivity being explored here. Schema 3 works to resolve the Prisoners' Dilemma by introducing a different mode of individual reasoning. It is therefore not reliant on any moral considerations or sentiments, or on any sympathetic feelings, that would change the value of the payoffs. Indeed, the instrumentally cooperative outcome of the Prisoners' Dilemma game might be contrary to the public interest or contrary to morality: for example, if they were to cooperate, the prisoners in the original Prisoners' Dilemma game would be acting so as to minimize punishment for their crime. Instrumental cooperation is thus not intrinsically moral. The approach of many theorists, including Frank, Sally and Gintis et al., tends to regard cooperative behaviour as intrinsically moral or prosocial in some sense. But the account of instrumental cooperation developed in this essay shows that cooperation is beneficial for the players in the sense of achieving an outcome that is Pareto-superior to the Nash equilibrium in terms of players' payoffs. This is separate from the question whether it is beneficial or 'good' in some other or wider sense for the players' payoffs to be thus increased. Cooperation amongst criminals or, say, amongst collusive oligopolists, is clearly a different matter from cooperation amongst citizens to vote and do their recycling: agents may cooperate to further many different kinds of ends, not all of which are socially beneficial or even legal.

Of course, this is not to suggest that there are not many socially beneficial forms of cooperation which are motivated by moral or social concerns or by feelings of friendship and sympathy; such cooperation is clearly important in human life, but if it is explained by payoff transformation it doesn't address the theoretical problems posed by the Prisoners' Dilemma. Neither is it to underestimate the importance of societal norms in predisposing people to one conception of agency, or one mode of reasoning, over another. The analysis of modes of reasoning presented in this essay thus does not detract from the importance of such norms in social and economic life.[11] Notwithstanding these wider considerations, however, what the essay does argue is that, in the specific case of the simultaneous one-shot Prisoners' Dilemma, achievement of the cooperative outcome can be an instance of instrumental cooperation motivated by the pursuit of individual interest, whether or not pursuit of that individual interest might be thought to involve self-regarding or other-regarding behaviours, or to involve material or non-material objectives, or to conduce towards a moral, amoral or immoral orientation. I suggest that recognition that cooperation can be individually rational provides a challenge to arguments which purport to show that cooperation by individual agents in the simultaneous one-shot Prisoners' Dilemma is self-sacrificial and hence irrational. Given the importance of such dilemmas in an increasingly interdependent world, this conclusion seems worth taking seriously.[12]

IV

In this essay I've argued that the resources of TMS suggest, or at least are consistent with, a new mode of individual practical reasoning along the lines of Schema 3's intersubjective rationality. In contrast with the notion of individual agency that is characteristic of game theory and decision theory, this notion of the individual agent incorporates an intersubjectivity that recognises the presence of another 'I'. According to the intersubjective mode of individual practical reasoning presented in Schema 3, it is because the individual 'I' is premised upon the presence of another 'I' that it is able to pursue individual objectives by means of instrumental cooperation. This intersubjective mode of reasoning thus dissolves the dichotomy between 'individual interest' and 'collective interest' that has characterised much of the debate about the Prisoners' Dilemma, thus enabling recognition of the way that some individual players can exploit congruent individual interests by adopting cooperative strategies.

The intersubjective mode of individual practical reasoning presented in Schema 3 does not rely on moral principles or on moral motivation for agents: even the most selfish, although not only the selfish, can reason their way to instrumentally cooperative choices. Its aim is efficiency (for the individuals concerned, not necessarily for society) not morality; and that is why it provides resources for challenging the traditional logic of the Prisoners' Dilemma. This is not inconsistent with Smith's approach either. In many respects Smith was not optimistic about the prospects for moral excellence. He inclined to the view that the basic safety and well-being of society are more likely to be assured if (in addition to the public works he argues for) society relies upon the pursuit of individual interest, suitably restrained by the laws of justice and conventional rules of decent behaviour, rather than on voluntary benevolence or the demanding morality of the impartial spectator.

Acknowledgements

Earlier versions of this essay were presented at: ‘Adam Smith on the Conditions of a Moral Society’, an interdisciplinary conference commemorating the 250th anniversary of the publication of Adam Smith’s *Theory of Moral Sentiments*, University of Oslo, Norway, August 2009; ‘Workshop on *The Theory of Moral Sentiments* 1759 on its 250th anniversary’, Institute of Advanced Study, University of Bologna, Italy, December 2009; the Cambridge Seminar in the History of Economic Analysis, Clare Hall, Cambridge, UK, February 2010; and HET conference, Kingston University, UK, September 2010. I am grateful for discussion on these occasions and in particular I thank Nahid Aslanbeigui, Luigino Bruni, James Konow, John O’Neill and Roberto Scazzieri. I also thank Lisa Herzog for discussion and comments. This essay was completed whilst I was an Academic Visitor at the Faculty of Philosophy, Oxford; I am grateful for the hospitality provided. I am indebted to an anonymous referee for very helpful comments.

[1] For a history of the game see Poundstone (1992) and http://ask.metafilter.com/126323/Prisoners-Dilemma-citation.

[2] See also Rabin (1993) for the argument that the Prisoners’ Dilemma should be understood as a coordination game.

[3] Schick (1997) argues that, as both selfishness and altruism are consistent with the Prisoners’ Dilemma (hence the possibility of an altruists’ dilemma), ‘the problem is rationality’ (p. 96). But it seems to me that Schick forgets this when he later argues that sociality or friendship may be (even, is) sufficient to overcome the dilemma. His later argument is that if a player is influenced by the other player’s *wanting* him to cooperate, the player will cooperate: ‘Suppose that Adam and Eve are social vis-à-vis each other. Each then wants to do what the other wants him or her to do. Each knowing what the other wants, each will here take S [*cooperate*], and that will yield (S, S), which is the cooperative outcome. That is, if Adam and Eve are social, they will both cooperate, and this though rationality directs them both to T [*not cooperate*]. … sociality implies cooperation. And in a one-round Prisoners’ Dilemma, social people cooperate and rational people don’t.’ (pp. 121–2). But, as with Frank’s and Sally’s attempts to integrate Smith’s sympathy, Schick’s sociality yields this outcome by payoff transformation so that *cooperate* becomes the dominant strategy because it includes the payoff to pleasing the other. I am indebted to Raino Malnes for drawing my attention to this work by Schick.

[4] This interpretation builds on that of Brown (1994). I am grateful to Christel Fricke and Dagfinn Føllesdal for stimulating my interest in the intersubjectivity of TMS by inviting me to the Adam Smith – Edmund Husserl workshops, CSMN, 2007 and 2008.

[5] This distinction is introduced in Brown (1997).

[6] The initial explanation of third-person moral judgments in Part I thus needs to be understood in the light of the later explanation of first-person moral judgments in Part III.

[7] This raises some questions as to whether what is taken as the canonic liberal representation of individuality is actually true to the writings of what are taken to be the canonic liberal philosophers; but that is another story.

[8] See Brown (1994) ch. 8, for a discussion of some issues relating to Smith and freedom.

[9] Bacharach argues in favour of a collective/team notion of agency and rationality as a means of solving a number of problems in game theory, including the Prisoners' Dilemma; and so Gold and Sugden contrast Schema 1's 'Individual Rationality' with schemata showing 'Team Rationality' in terms of the agent 'We' construed as a team. I do not address Bacharach's arguments in this essay.

[10] This is not to suggest that the intersubjectivity of TMS is the only model of individuality that could support intersubjective rationality, although it was in thinking about TMS that I first developed the intersubjective rationality of Schema 3.

[11] For the argument that competitive economic efficiency (as specified in the First Fundamental Theorem of Welfare Economics) requires moral normative constraints, see Schultz (2001).

[12] It does not imply, of course, that public policy responses to such dilemmas are thereby made redundant. The aim of this essay is only to argue against the proposition that cooperation in the specified circumstances is self-sacrificial and, hence, irrational.

Bibliography

Bacharach, M. (2006) *Beyond Individual Choice: Teams and Frames in Game Theory*, N. Gold and R. Sugden (eds), Princeton University Press.

Binmore, K. (1994) *Playing Fair*, vol. 1 of *Game Theory and the Social Contract*, MIT Press.

—— (2006) 'Why do people cooperate?', *Politics, Philosophy and Economics*, 5: 81–96.

Brown, V. (1994) *Adam Smith's Discourse: Canonicity, Commerce and Conscience*, Routledge.

—— (1997) 'Dialogism, the gaze and the emergence of economic discourse', *New Literary History*, 28: 697–710.

—— (2009) 'Agency and discourse: revisiting the Adam Smith problem', in the *Elgar Companion to Adam Smith*, Jeffrey T. Young (ed.), Edward Elgar, pp. 52–72.

Camerer, C.F. (2003) *Behavioral Game Theory: Experiments in Strategic Interaction*, Russell Sage Foundation, New York; Princeton University Press.

Fehr, E. and Fischbacher, U. (2005) 'The economics of strong reciprocity', in *Moral Sentiments and Material Interests: The Foundations of Cooperation in Economic Life*, H. Gintis et al. (eds), MIT Press, pp. 151–91.

Frank, R. (1988) *Passions Within Reason: The Strategic Role of the Emotions*, W.W. Norton.

—— (2004) ‘Introducing moral emotions into models of rational choice’, in *Feelings and Emotions: the Amsterdam Symposium*, A.S.R. Manstead, N.H. Frijda and A. Fisher (eds), pp. 422–40, Cambridge University Press.

—— (2007) ‘Cooperating through moral commitment’, in *Empathy and Fairness*, G. Bock and J. Goode (eds), Novartis Foundation Symposium 278, 2006, John Wiley, pp. 197-208, discussion pp. 208–15.

Fudenberg, D. and Tirole, J. (1991) *Game Theory*, MIT Press.

Gintis, H., Bowles, S., Boyd, R. and Fehr, E. (eds) (2005a) *Moral Sentiments and Material Interests: The Foundations of Cooperation in Economic Life*, MIT Press.

Gintis, Herbert, Bowles, S., Boyd, R. and Fehr, E (2005b) ‘Moral sentiments and material interests: origins, evidence, and consequences’, in *Moral Sentiments and Material Interests: The Foundations of Cooperation in Economic Life*, H. Gintis et al. (eds), MIT Press, pp. 3–39.

Gold, N. and Sugden, R. (2006) Introduction and Conclusion to M. Bacharach, *Beyond Individual Choice: Teams and Frames in Game Theory*, Princeton University Press.

—— (2007) ‘Theories of team agency’, in *Rationality and Commitment*, F. Peter and H.B. Schmid (eds), Oxford University Press, pp. 280–312.

Poundstone, W. (1992) *Prisoner’s Dilemma*, Doubleday.

Rabin, M. (1993) ‘Incorporating fairness into game theory and economics’, *American Economic Review*, 83: 1281–302.

Sally, D. (1995) ‘Conversation and cooperation in social dilemmas: a meta-analysis of experiments from 1958 to 1992’, *Rationality and Society*, 7: 58–92.

—— (2000) ‘A general theory of sympathy, mind-reading, and social interaction, with an application to the Prisoners’ Dilemma’, *Social Science Information*, 39: 567–634.

Schick, F. (1997) *Making Choices: A Recasting of Decision Theory*, Cambridge University Press.

Schultz, W.J. (2001) *The Moral Conditions of Economic Efficiency*, Cambridge University Press.

Smith, A. (1976) *The Theory of Moral Sentiments*, D.D. Raphael and A.L. Macfie (eds), Clarendon Press: Liberty Press imprint, 1982.

Smith, V.L. (2010) What would Adam Smith think?’, *Journal of Economic Behavior & Organization*, 73: 83–6.

http://ask.metafilter.com/126323/Prisoners-Dilemma-citation.