Rules, Norms, Commitments

Fabienne Peter and Kai Spiekermann

f.peter@warwick.ac.uk

k.p.spiekermann@warwick.ac.uk

Draft in preparation for the *Sage Handbook of Philosophy of Social Science*. London: Sage, forthcoming 2009.

October 2008

1. Introduction

In Christopher Nolan's film *The Dark Knight*, the character of the Joker claims that "the only sensible way to live in this world is without rules". The Joker, impressively impersonated by the late actor Heath Ledger, is portrayed as a powerful and malicious figure, determined to sow chaos by defying all those rules which influence and direct the lives of the people he interacts with. The Joker breaks the conventions of daily city life, mocks the mafia's code of honour, and defies moral principles such as the principle that one should not harm innocents. As a result of the destructive force of the Joker's conduct, the social structure as it is defined by a multitude of rules and normally upheld by those whose actions follow these rules starts to falter. The film suggests that this produces an outcome that is considerably worse than the previous stalemate between organized crime and those seeking to enforce the law. The film thus contradicts the Joker and seems to endorse the widely held view that while not all rules are beneficial, some rules are necessary for, and perhaps even constitutive of, social life. On

this widely held view, rules and rule-based behaviour are ubiquitous and following rules makes sense.

Even if this broad view is accepted, there is still a range of challenging questions that face philosophers of social science. First, what are rules? Are rules primarily solutions to coordination or cooperation problems, or is there more to them? Some have argued that rules do not just regulate existing behaviour, but help create new social facts. Others insist that rules are necessary for any meaningful action. Second, what motives do individuals have to follow rules? Is rule-following itself, not just the avoidance of sanctions, in the individual's self-interest? Do individuals follow rules for reasons other than self-interest, maybe for reasons that relate to the normativity of certain rules? And, finally, how do rules emerge and what determines their stability? Pointing to the fact – functionalistically – that a rule would have a beneficial effect if followed by all individuals does not explain how rules come into being in the first place, and how they can persist.

This chapter looks at different approaches to answering this question. In the next section, we provide an overview from a historical perspective. In subsequent sections, we discuss in some more detail three approaches which dominate contemporary debates about rules and rule-following: rational choice theory, including classical game theory, evolutionary theories, including evolutionary game theory, and approaches based on collective intentions.

Before we start, however, it will be helpful to settle some terminological issues. What is a rule? As we shall use the term, the concept of a rule includes the possibility of making mistakes, of failing to conform to a rule (Winch 1958: 32; Coleman 1990: 242; Pettit 1993: 82). Seen in this light, the concept of a rule is necessarily prescriptive. It involves expectations that oneself and other people might have towards one's behaviour. Among rules thus interpreted, we distinguish between conventions on the one hand and social and moral norms on the other. Conventions come with expectations about how they are appropriately followed. In the US and the UK, for example, a conventional "continental breakfast" involves bread or pastries, butter and jam. If a hotel menu described its "continental breakfast" as consisting of delicious rice dumplings, customers are likely to be puzzled (but possibly quite

pleased). Social and moral norms, by contrast, are prescriptive in a stronger sense. They involve not just empirical, but also normative expectations: one ought to conform to them. Such norms may, for example, rule that certain types of foods ought not to be eaten, at least not during certain periods.

As we shall use the term here, both conventions and norms are thus rules, but they are different kinds of rules. According to H. L. A. Hart's influential terminology, only norms are rules of obligation. As rules of obligation, social norms come with a normative expectation to conform, and there are sanctions against those who deviate. Norms are also deemed more important – more socially beneficial – than other rules. A third feature of norms understood as rules of obligations is that acting in conformity with them may often be in conflict with what narrow self-interest dictates.[1] This last feature raises the question of commitment. We call commitment the willingness to make a particular rule the motive for one's action. In the case of social and moral norms, acting from commitment is acting from the normative reasons formulated by these rules. The rules in question may, however, also refer to conventions or to personal plans. Examples are "always drive on the left side of the road", or "exercise every day".

2. The Ontology of Rules and Rule-Following

It is widely accepted that there are two types of rules. There are rules which regulate existing forms of behaviour and there are rules which create the very possibility of some forms of behaviour. John Searle (1995, p. 27) calls the former "regulative rules" and the latter

[1] For this account, see Hart (1961: 84 – 85); see also Ullman-Margalit (1979), Tuomela (1984), Miller (2001), and Gilbert (2006), among others. Note that there is some disagreement in the literature about whether or not conventions should be interpreted as rules. While many philosophers do interpret conventions as rules (e.g. Lewis 1969), others argue that conventions are not prescriptive and hence not rules (e.g. Gilbert 1989, Searle 1995).

"constitutive rules".[2] An example for a regulative rule is the convention to drive on one particular side of the road. This convention does not create the possibility for driving, and driving does not depend on the existence of such a rule. The rule merely solves some problems that might arise with driving. The rules that define the game of chess, by contrast, enable a form of behaviour that does not exist without these rules. The general form of constitutive rules, in Searle's distinction, is "X counts as Y" or "X counts as Y in context C" (Searle 1995, p. 28). That is to say, a constitutive rule states that some object X has, at least in some contexts, some particular properties which are not reducible to its physical properties. Searle's favourite example is how certain pieces of paper count as money in a particular country. In the context of chess, the set of rules specifies how a particular set of moves of figures on a chequered board count as a board game.

It may not always be possible to sharply distinguish between the two types of rules. Anthony Giddens (1984, pp. 19ff), for example, has argued that constitutive rules, such as the rules of chess, tend to involve regulative elements too, and, conversely, regulative rules tend to contain constitutive elements. An example is the convention to keep offices open from nine to five, which may be part of what constitutes "work", as opposed to "leisure". But the distinction is nevertheless helpful to shed light on different approaches to the question of what rules are and what their significance is. Approaches prominent in sociology – functionalism, structuralism and interpretivism – have tended to focus on constitutive rules (or on the constitutive element in regulative rules) in their explorations of "social" as opposed to "natural" facts. Rational choice theory, by contrast, has highlighted regulative rules (or the regulative element in constitutive rules). David Lewis' study of conventions (1969), David Gauthier's study of moral principles (1986), and Jon Elster's study of commitment (1979; see also 2000) all take particular interest in the instrumental role that rules and rule-following might play in individuals' pursuit of their interests.

[2] For related distinctions, see Rawls (1955) and Giddens (1984). Giddens distinguishes between "constraining" and "enabling" rules.

Let us start with a debate from sociology. The early days of sociology were marked by the struggle for an independent discipline devoted to the study of social phenomena – as distinct from biological, psychological, or economic phenomena. Emile Durkheim (1938) famously argued that there are independent social facts, and that it is sociology's task to study these facts. Rules such as family norms or dress codes are examples of social facts. According to Durkheim, such social rules exist externally to individuals and have coercive power over them. In this view, the regularities observed in individuals' behaviour, and even the individuals' subjective desires to conform to social rules, are not the result of their subjective wills but produced by social rules (1938, p. 2). And these social rules exist because of the function they fulfil for society as a whole. Durkheim (1938, p. 96) discusses the example of norms of blame and punishment. Instead of seeing them as being brought about by "the intensity of the collective sentiments which the crime offends", he argues that these sentiments are better interpreted as enabling the very individual intentions which sustain the punishment of crime.

Max Weber defended the contrasting view, which takes the subjective meaning individuals attach to their actions as starting-point, not objective social structures. Weber distinguishes action ("Handeln") from behaviour ("Verhalten"); the former is behaviour to which the agent attaches subjective meaning. According to Weber, it is the study of "social action" – meaningful behaviour that "takes account of the behaviour of others and is thereby oriented in its course" (Weber 1947, p. 88) – that should define the core of the discipline. As Weber (1947, p. 88) puts it: "[s]ociology … is a science which attempts the interpretative understanding of social action in order thereby to arrive at a causal explanation of its course and effects."

Weber gave much room to social action that results from a commitment to social rules. But he was also adamant that not all observed behavioural regularities derive from the commitment towards social rules. He stressed how action can be oriented by social rules by being directed against them; actions can thus be rule-oriented without being determined by them. In addition, in a move that foreshadowed rational choice theory, Weber argued that

people's self-interested action, including action that settled on some conventional course of action, could also bring about social regularities (Weber 1947, pp. 120ff).

Many followed Weber's view that the concept of "social action" defines the subject matter of social science. Some, however, argued that Weber put too much emphasis on subjective meaning to capture the causal impact of social structures, including social rules, on individual action. According to Talcott Parsons, sociology had to overcome the "utilitarian dilemma" which affects individualist theories of action (Parsons 1937, p. 64). As James Bohman puts it, this dilemma is the following: "*either* the actor is independent in choosing his or her end, in which case the ends are random rather than rational; *or*, the actor is not independent, in which case the causes are external rather than voluntary" (Bohman 1993, p. 33, his emphasis), for example determined by biological factors. Parsons sought a theory of action that avoided both horns of the dilemma. He did so by combining insights from Weber and Durkheim. Parsons argued that "Weber … missed the important distinction … between motivation considered as a real process in time and atemporal complexes of meanings as such" (Parsons 1937, p. 636). The former refers to the actual subjective meaning an agent attaches to his or her actions and captures a particular relation between ends, means, and conditions. The latter sense of meaning, however, can be detached from the motivations of the particular agent, and refers to systems of ideas and the value frames embodied in social structures. It is a property of objective social structures and accounts for their normative force. According to Parsons' structural-functionalist theory, all actions take place in a "frame of reference" (Parsons 1937, p. 733) that logically connects "ends, means, conditions, and norms". As a result, individual actions, while endowed with subjective meaning, are not random. Nor are they directly determined by social norms. But by defining the frame of reference, social structures give individual actions a normative orientation: "action is ... the process of alteration of the conditional elements in the direction of conformity with norms" (1937, p. 732). In his later work, when pressed to explain how the norms that are part of the action frame of reference remain not merely external to the motivation of individuals but

actually come to influence individual action, Parsons added the idea that individuals internalize norms in the process of social action.

Many have objected to Parsons' version of structural-functionalism that it leans too far to the objectivist side and neglects the role of rational agents. Rational agents are able to give an account of their reasons for acting in a particular way. They are aware how rules constrain their actions, and there are different ways in which they might factor in these constraints in their instrumental or ends-oriented deliberations. Parsons' theory of social action fails to take into account the role individual agency plays in the production and reproduction of social structures and social rules.[3] There is a tendency in Parsons' theory of social action to treat individual agents as "'judgmental dopes' who passively assimilate the rules and roles they are socialized into and merely act out the value orientations of their culture and its institutions" (Bohman 1993, p. 37). As such, it rests on an implausibly strong and rigid notion of individual commitment to social norms.

The opposite stance is taken by defenders of rational choice theory (RCT; see: Rational Choice Theory). The standard version of the theory shifts the emphasis away from Weber's concept of social action to individual action when explaining social phenomena. It also brackets the influence the normative content of rules might have on individual action. Rational choice theorists embrace the first horn of Parsons' utilitarian dilemma. They deal with the problem of randomness that troubled Parsons by using a stripped-down conception of rationality. According to this conception, rationality demands that individual preferences are well-ordered and that individual's actions are describable as an attempt to best satisfy such well-ordered preferences. Following Savage's (1954) seminal work, if an agent's preferences meet certain axioms of consistency, the preferences can be represented as utilities over consequences and subjective probabilities over events. A rational agent in Savage's sense acts as if maximising expected utility. While Savage developed his theory for parametric decisions of a single agent, game theory has drawn on his decision theory to develop theories of strategic interactions between agents (see: Game Theories), which form

[3] It neglects, in other words, what Giddens (1984, p. 191) calls "structuration" – the interplay between individual agency and social structures; see also Coleman (1990).

an important part of RCT. Rational choice theorists explain observed behavioural regularities as being the result of such utility maximizing behaviour. Applied to rule-following behaviour, RCT implies that a convention or a social norm is observed because the action that the rule recommends happens to be the one that maximizes individual utility. Game theory in particular has enabled much headway in answering the question whether complying with a rule is individually rational.

RCT's success comes at a price, however. The standard model of RCT highlights how certain rules are compatible with individual rational action. This perspective does not have much room for rule-following behaviour as such. Because the standard model identifies the content of individual preferences by whatever the individual subjectively values, it is particularly affected by the following dilemma (McClennen 2004, p. 223). Suppose that a particular rule requires an individual to do A in circumstances C. If there is a better alternative B, then B is the rational choice – not following the rule. If there is no better alternative in C, then A should be chosen. But again, the recommendation is not to follow the rule, but to do A because it best satisfies individual preferences.

To put the point differently, the standard model of rational choice theory neglects the possibility of individuals acting from commitment to particular rules (Sen 1977). If a convention merely helps individuals to coordinate by identifying one among several alternative courses of actions judged equally good, this seems not much of a problem. In most cases of rule-following, however, there is something at stake if one rather than another rule is followed. In addition, rule-following behaviour – as opposed to behaviour that is merely in accordance with a rule – seems ubiquitous. Neglecting commitment is thus not satisfactory.

The contemporary literature is marked by three responses to this problem. The first, adopted by many rational choice theorists, is to explore the possibility of extending the standard model to incorporate rule-following (see sections 3 and 5). Evolutionary game theorists adopt a second response (see section 4). They model individual actions as based on strategies and study the survival chances of different strategies. Social norms are interpreted as strategies and models of evolutionary game theory aim to analyse the emergence and

evolutionary stability of different strategies or norms. Finally, there are also some philosophers who are not convinced that revising rational choice theory or moving to evolutionary game theory will solve the problem. These philosophers advocate alternative theories of practical reason (e.g. Anderson 2000; 2001; see section 5).

A rather different perspective on rules and rule-following is taken by those defending an interpretative approach to the social sciences. As a version of anti-naturalism, this approach holds that the methods of the natural sciences cannot be used to study the social realm. Interpretivism does not endeavour to link rules to the explanation of regularities in people's observed behaviour. What makes social rules important, according to this approach, is how they relate to what counts as meaningful action. The most influential account of this kind has been given by Peter Winch (1958).[4]

Winch, like Parsons, starts with Max Weber's concept of social action and treats it as the fundamental concept for the social sciences. Recall that for Weber, action is subjectively meaningful behaviour – actions are performed for a reason. And action is social if it is oriented towards others. In a first step, Winch focuses on the question of what constitutes meaning. He claims that all meaningful action is action that follows a rule. He defends this claim with the help of a broad conception of commitment. Winch (1958, p. 50) argues that meaningful action is committed action, in the following sense. It commits the agent to act in a similar way in a similar situation in the future. And committed action, thus interpreted, is rule-bound: "I can only be committed in the future by what I do now if my present act is the *application of a rule* (1958. p. 50; his emphasis). He grants that interpreting commitment in this way "is most obviously appropriate where we are dealing with actions which have an immediate social significance" (1958, p. 50). One of his examples is the norm of promise-keeping. But his broad conception of commitment is not limited to that; one can also be committed to private actions, Winch argues, such as when one places a bookmark with the intention to continue reading at the marked passage.

[4] Other important contributors to this tradition include, for example, Taylor (1971) and Habermas (1984).

In a second step, Winch links this account of meaningful action to Wittgenstein's insights about language and rule-following. Winch claims that because rule-following is a social concept, all meaningful action is social. His argument is, briefly, the following. The concept of rule-following is related to the concept of making mistakes. Without the possibility of failure, the idea of following a rule does not make sense. This makes rule-following an evaluative concept and links our actions to the actions and expectations of other people (Winch 1958, p. 32). Drawing on Wittgenstein's private language argument, Winch argues that although it is possible to formulate and follow rules that apply only to one's own behaviour, the very concept of rule-following as something that qualifies appropriate behaviour relies on the possibility of external checks – on other people being able to recognize one's behaviour as following rules and evaluate its appropriateness. Individuals can only develop a sense of rules that apply to their private behaviour if they have experienced behaviour governed by established social rules, Winch claims. It is for this reason that rules point to a social setting. Winch concludes that Weber was wrong to distinguish between meaningful behaviour and meaningful behaviour that is social: "all meaningful behaviour must be social, since it can be meaningful only if governed by rules, and rules presuppose a social setting" (1958, p. 116). In Winch's view, therefore, even more than in Weber's, social action is the fundamental concept of the social sciences. This amounts to an anti-individualist view, according to which individuals non-causally depend on each other (Pettit 2000, p. 70).

Anticipating objections to his interpretation of the concept of social action, Winch stresses that such rule-following behaviour need not be conscious. It is present whenever it makes sense to distinguish between a right and a wrong way of doing things (Winch, 1958, p. 58). In addition, Winch insists that rule-following behaviour needs to be interpreted broadly. An example that he discusses is the anarchist. Even the anarchist can act meaningfully; the relevant rule in this case may be "break all rules". According to Winch, this distinguishes the anarchist from the "berserk lunatic", whose behaviour is indeed pointless and thus does not qualify as social action (Winch 1958, p. 53). Heath Ledger's Joker in the *Dark Knight* vividly illustrates the difference between the two.

Yet, some important objections to this approach to rules and rule-following remain. A first objection is that Winch's broad conception of commitment includes too much. There can be meaningful behaviour that does not rest on a distinction between doing things correctly or incorrectly and hence does not follow a rule. The objection can be stated in terms of Weber's account of meaningful behaviour: there is subjectively meaningful behaviour that is not influenced by customs or the social order. In defence of Winch it can be argued that all meaningful behaviour – action done for a reason – involves concept possession. Insofar as concept possession covers the correct and incorrect usage of the concept, Winch is right that all meaningful behaviour is rule-based in this sense (Gilbert 1989, p. 71; Pettit 2000).

The problem with this defence is that it leads to a very thin interpretation of rule-following. Hence even if Winch can be defended along this line, there is a further main objection against Winch which states that his broad conception of commitment demands too little. Winch fails to explain the special influence that some social rules have on human behaviour – e.g. some social or moral norms. The commitment to such norms does not follow from mere concept possession. So the question remains why people are deeply committed to some rules, both with regard to their own behaviour but also with regard to the behaviour of other people, while quite indifferent about others. Winch's argument that meaning necessarily depends on social rules is thus incomplete because it only shows that meaning is something that must be shareable, but not why and how groups of people establish certain rules of obligation that orient the actions of their members (Gilbert 1989, pp. 93 and 400; see also section 5 below).

Philip Pettit (1993; 2002) proposes to combine elements of rational choice and sociological approaches in order to get to an answer to the question as to what gives certain rules, such as social norms, their resilience (see also Elster 1989). As Pettit puts the question (2002, p. 309): "what ensures that in suitable circumstances those norms can be relied on to emerge and persist?" His answer to this question draws on the following two ontological claims. The first concerns the effect of structural regularities on individual agency. The kind of collectivism that Durkheim endorses implies that there are social regularities such as

particular cultural norms or the incidence of suicide which are not causally or logically continuous with regularities in the intentional actions of individuals. Such "socio-structural" regularities, as Pettit calls them, have the power to "override" individual agency in Durkheim's theory. Pettit's more moderate claim, compatible with ontological individualism, is that while it can often make sense to invoke structural regularities in explaining social phenomena, such regularities need not be seen as undermining the status of individual intentional agents. The second claim concerns the relation between individual agents. Pettit rejects the atomist ideal, endorsed in the standard model of rational choice theory. According to atomism, the actions of individual intentional agents may be causally affected by the actions of other agents, but their status as agent does not depend on others. Instead of atomism, Pettit endorses the holist view that individual intentional agents non-causally depend on their relations with other agents, for example in their in capacity to think, or to be rational. In defending this view, Pettit, like Winch, relies on the link between meaningful action and rule-following, and agrees with Winch that meaning, thus interpreted, is social (Pettit 2000).

With the help of these two claims, Pettit (2002, pp. 308ff.) develops the following answer to the question of what accounts for the resilience of social norms. Pettit identifies three requirements for a rule to be a social norm: (i) it is a regularity with which people generally conform; (ii) conformity with the regularity attracts widespread approval and deviation attracts widespread disapproval, and (iii) the fact of approval and disapproval helps ensure that people generally conform to the norm. Pettit emphasizes the importance of the third requirement, as without it there is nothing that ties the first two requirements together. In other words, to explain the power of social norms compared to other regularities, it has to be shown how conformity to the norm relates to the normativity of the norm.

Pettit tries to show this by advocating an extension of rational choice theory which starts, not from the question which course of action is rational, but from the question what attitudes rational individuals should adopt. Such an attitude-based derivation of social norms, Pettit claims, can "show that a certain sort of behaviour is bound to attract approval, its absence

disapproval, and that such sanctions ought to elicit the behaviour required, thus establishing norms" (Pettit 2002, p. 323). He gives the example of a norm against overgrazing in a "tragedy of the commons" situation. In this situation, standard rational choice theory predicts that free-riding will prevail. His attitude-based derivation of a norm against overgrazing relies on the possibility of people recognizing the costs and benefits of the different ways in which they interact with each other on the commons. Since avoiding overgrazing is better for all than an attitude which favours overgrazing, it is likely that there is approval of behaviour that avoids overgrazing and disapproval of other behaviour. The presence of disapproval, then, makes overgrazing more costly, and hence supports the norm. What Pettit argues, in sum, is that there is a social rationality which brings about attitudes that facilitate the emergence and persistence of this regulative rule. Note that this amounts to an exact reversal of Durkheim's claim, mentioned earlier, that it is the structural regularity or rule which produces the attitudes (sentiments) necessary to sustain it.

3. Rational Choice Theory

RCT comes in different versions. In its most stripped-down version it only assumes that agents act consistently according to Savage's axioms of rationality, without making any assumption about how agents form preferences. However, when RCT is applied in the social sciences, it is usually assumed that agents maximise their own welfare (Hechter & Kanazawa 1997), presupposing a specific motivation that determines preferences.[5] We call this the standard model of RCT. The idealized agent of these rational choice models is often called a "homo economicus" to underline the focus on personal welfare and on the maximization of payoffs. Over the years, the standard model has been changed or extended in many ways to incorporate social phenomena that are difficult to capture with the assumption of personal welfare maximization.

An important charge against the standard model of RCT is that it fails to model rules, norms, and commitments in an adequate way. We assess this charge from several

[5] More precisely, the claim is that in many situations agents act *as if* they were maximising their welfare. Therefore, RCT is not refuted by showing that agents do not always actually maximise welfare.

perspectives and look at proposals to extend RCT to address this problem. First, there are inherent problems in game theory to explain how agents coordinate actions. These problems are addressed by introducing the notion of conventions. Second, game theory predicts the breakdown of cooperation in one shot mixed motive games such as the prisoner's dilemma (PD). Since cooperation problems are often solved through norms of cooperation, we look into recent attempts to extend the game-theoretical analysis to model such norms. Third, we examine whether committed action is necessarily outside of the explanatory reach of RCT, or whether the standard model of RCT can be extended to incorporate it.

Rational Choice Theory and Sanctions

Standard rational choice explanations of norms focus on how it may be rational for a homo economicus to act in accordance with social norms. Such explanations must take into account that acting in accordance with norms often poses a dilemma of cooperation: it is attractive to free-ride while others comply. The dilemma of cooperation may take on the form of a public good problem: Mutual compliance is a public good, but it is difficult to provide the public good because self-interested agents find it more attractive not to comply and benefit from whatever level of compliance is reached by other agents.

Rational choice theorists argue that sanctions can change agents' payoffs such that acting in accordance with norms is in their rational self-interest (Hausman & McPherson 2006, pp. 72-76, 80-85). Sanctions decrease the utility attached to outcomes produced by undesired actions, and the subject of the sanction is supposed to change her preferences and behaviour in anticipation of these sanctions. One can distinguish between formal and informal sanctions. Threat of bodily harm or death, imprisonment, unpaid work, etc., are formal sanctions typically applied to enforce legal norms. Formal sanctions may also play a role for the enforcement of social norms in violent societies or subcultures (think of enforcement of social norms in the mafia or street gangs). However, most of the time social norms are not enforced with formal sanctions, but in more subtle, informal ways. Individuals do not only care for their bodily integrity, their freedom, and their money, but also for less measurable

goods such as social contacts, approval, recognition, and reputation. Informal sanctions are based on these desires.

Ostracism can be a powerful sanction to enforce social norms. In laboratory experiments where participants could choose their level of contribution in a public good game, contributions were significantly higher when participants were able to exclude low contributors for the next rounds (e.g. Cinyabuguma et al. 2005). The option to deprive agents of cooperation gains from future interactions is a strong incentive to cooperate or comply with norms.[6] Another incentive to comply with norms is the *social approval* agents receive for complying, and the disapproval for norm violations. Social approval can be intrinsically or instrumentally valuable. If it is intrinsically valuable, agents care for social approval as such. If agents care for social approval instrumentally, they consider their reputation and how a good or bad reputation will influence future interactions with other agents.

Sanctions are important to enforce social norms. However, a second-order question emerges: Since meting out sanctions is costly in terms of time and effort, how do groups solve the collective action problem of sanctioning? Some claim that humans are biologically disposed to punish those who do not cooperate. Others maintain that some of the most effective sanctions are costless. For instance, Brennan and Pettit (2004) argue that we reward and sanction people by holding them in esteem or disesteem, which is costless for the person supplying or withholding esteem. This is part of Pettit's proposal to extend standard RCT by incorporating attitudes, as mentioned above.

These considerations show that RCT in its standard form or with some extensions can offer explanations why it may be instrumentally rational for people to comply with norms. We now address problems arising within the game-theoretical foundations of RCT and will later return to the question as to whether RCT and its focus on sanctions gives an adequate motivational account as to why people comply with norms.

[6] See also Spiekermann (2007) for a formal model.

Conventions

We want to start with a discussion of conventions as solutions to coordination problems. Game theory struggles to explain seemingly innocuous coordination problems between two or more people. To understand the coordination problems at hand, consider an example. Ann and Bob drive towards each other on a road. Both drivers can either drive on the left or on the right. The two cars can pass each other if they drive on different lanes (i.e. both drivers drive on their left or right side), but they will crash if they both try to use the same lane. Table 1 shows the associated coordination game. The numbers in this table (and all tables below) are utility indices, and all agents maximize expected utility.

Table 1: Driving Game.

		Bob	
		left	right
Ann	left	1, 1	0, 0
	right	0, 0	1, 1

There are two pure strategy, strict Nash equilibria in this game: (left, left) and (right, right). In these two equilibria, Ann and Bob choose the best strategy available, conditional on the strategy of the other player: If Ann drives on the left, it is best for Bob to drive on the left, and vice versa. If Ann drives on the right, then it is best for Bob to drive on the right, and vice versa.[7]

A standard assumption in game theory is that players have common knowledge of the structure of the game and their respective perfect rationality. Ann's and Bob's rationality is common knowledge if Ann knows that Bob is rational, and Bob knows that Ann is rational. Also, Ann knows that Bob knows that she is rational, and Bob knows that Ann knows that he is rational, and so on. Rational players should end up in one of the two pure strategy

[7] There is also one mixed strategy Nash equilibrium available: both players randomize their choice with probability 0.5. It is a Nash equilibrium because both players have no better response, given the other player's strategy. But it is a Pareto-inferior, unstable equilibrium, and hardly ever found in reality, so we leave the mixed strategy equilibrium aside.[7]

equilibria, but the game-theoretic rationality does not tell us in which. Ann reasons that she should use the same strategy as Bob. She knows that Bob reasons that he should use the same strategy as Ann. This leads to an infinite regress: Ann knows that Bob knows that Ann knows…. No equilibrium can be selected on grounds of rationality.

One might think that the problem of equilibrium selection occurs only in symmetrical games, where the players are indifferent between both equilibria. But this is not the case. Consider the "Hi-Lo" game in table 2. Ann and Bob cannot communicate, but they have common knowledge of their rationality and the game. According to common sense, Ann should play top and Bob should play left, thereby realizing the Pareto-optimal outcome. But this does not follow from the assumptions of rationality and common knowledge of rationality. There are still two strict, pure strategy Nash equilibria: (top, left) and (bottom, right). The Pareto-inferior (bottom, right) is a Nash equilibrium because Ann and Bob each play their best strategy, given the opponent's strategy. If Ann expects Bob to play left, she should rationally play top, if she expects Bob to play right, she should rationally play bottom, and vice versa for Bob. The problem is that, within standard game theory, there is no reason to expect one equilibrium or the other. Both Nash equilibria are the result of rational play, and the result is underdetermined (Bacharach 2006, ch. 1).

Table 2: Hi-Lo Game.

		Bob	
		left	right
Ann	top	2,2	0,0
	bottom	0,0	1,1

David Lewis (1969) argues that coordination problems like the driving game and the Hi-Lo game are solved by *conventions* (see also Cubitt & Sugden 2003). His work is probably the first formal analysis of conventions. However, it has roots in Hume's (1978) notion of a convention as a rule that emerges in repeated interactions, and draws on Thomas Schelling's (1960) work on coordination games. Schelling finds that most individuals have no difficulty

to coordinate on one equilibrium in practice, despite the game-theoretical problems described. For instance, when asking persons what they would do to meet a person in New York if they had not agreed on a time and a place, most suggest "noon, central station". Schelling calls these intuitive equilibria "focal points".[8]

Game theory informs us about the Nash equilibria, but it does not tell us which equilibrium the agents should aim for, and therefore the agents may fail to coordinate. A convention creates expectations as to which equilibrium is preferred. Conventions can emerge spontaneously by relying on precedence (Lewis) or focal points (Schelling). The driving game is usually solved by precedence: Ann and Bob have seen many drivers (in the UK) driving on the left, and they have reasons to believe that their counterpart has seen them, too, and therefore knows of and follows the convention to drive on the left. The Hi-Lo game, by contrast, is more likely to be solved by identifying outcome (top , left) as the focal point. Both players realize that (top, left) is pareto-optimal. Even though they have never played this game before, they expect the opponent to play top (or left). Real players usually succeed to coordinate on (top, left) immediately.

According to Cristina Bicchieri, a convention is a behavioural rule for a coordination game.[9] The convention exists if (a) there is a sufficiently large number of agents in the population who know of the rule and know that it applies to coordination games in certain situations; (b) a sufficiently large number of agents prefers to conform to the rule in the coordination game if they expect sufficiently many other agents to conform with the rule in the coordination game; and (c) a sufficient number of agents believe that enough other agents will conform with the rule and therefore they prefer to conform with the rule (Bicchieri 2006, pp. 31-38, compare Lewis 1969, p. 78). Conventions are thus conditional rules: If we expect enough other people to aim for one equilibrium, then we also prefer to aim for this equilibrium. An important feature of this analysis of conventions is that it is arbitrary which

[8] Schelling's focal points are an informal explanation of how conventions come into being. In the driving case, there are also legal norms and enforcement mechanisms to ensure driving on the right side of the road. But the convention would work even without these, as long as all people have non-crazy preferences over the avoidance of car crashes.

[9] Bicchieri rules out games with nonstrict Nash equilibria because this would imply that one or more players would not prefer to coordinate on one equilibrium, but are indifferent between two or more actions.

equilibrium is primed by the convention, it only matters that the convention creates expectations to coordinate on one of them. This feature can also be used to distinguish conventions from social norms. Norms create not just empirical expectations but normative expectations as well.

Social Norms

Problems of cooperation, which must be distinguished from problems of coordination, may be solved by norms. Consider a two person PD as in table 3 (we focus on two person games for the sake of simplicity, but the argument can be extended to multi person games).

Table 3: Prisoner's Dilemma

		Bob	
		cooperate	defect
Ann	cooperate	2,2	0,3
	defect	3,0	1,1

The PD has only one strict Nash equilibrium, (defect, defect), because the strategy "cooperate" is dominated by "defect", i.e. no matter what the opponent plays, defection is always preferred over cooperation. The "dilemma" in the PD is that rational players are unable to achieve mutual cooperation, even though both prefer it over mutual defection. In experiments where subjects face payoffs in the structure of a PD, subjects cooperate much more frequently than the game theoretical analysis suggests. Either many agents are irrational, or they perceive the situation differently and do not maximise their payoff. Interestingly, communication before the game increases the level of cooperation, suggesting that subjects may be able to agree on or remind each other of norms of cooperation and commit themselves to cooperate, often successfully (for a meta-analysis see Sally 1995).

Bicchieri (2006, p. 3) argues that social norms transform mixed motive games such as the PD into coordination games. If both players endorse a social norm of cooperation, it changes the players' utility of "cooperate" such that they prefer to cooperate as long as their opponent

cooperates. It also reduces the utility of playing "defect". Thus, if both players expect to play against someone who also endorses the social norm, they play a coordination game in the form shown in table 4, as argued by Bicchieri (2006).[10]

Table 4: Coordination game.

		Bob	
		cooperate	defect
Ann	cooperate	3,3	0,1
	defect	1,0	2,2

Endorsing a social norm means that an agent knows that the norm applies to specific situations and that the agent has a conditional preference for cooperation: The agent prefers to comply, conditional on the agents' *empirical expectations* that enough others comply, and on the *normative expectation* that enough others expect the agent to comply, and may sanction non-compliance (Bicchieri 2006, p. 11). These expectations transform the cooperation problem into a coordination game.

Note that a coordination game does not lead to cooperation by default. The coordination game has two pure strategy Nash equilibria: (cooperate, cooperate) and (defect, defect). This also implies that the norm can exist without being followed: It might be the case that all relevant agents know the norm, and have conditional preferences for norm-following, but do not expect others to comply, and consequently do not comply themselves.

Bicchieri's analysis shows that game theoretical concepts can be used to analyse norm compliance in an extended rational choice framework. A key ingredient is the formation of expectations about other players in an environment of private information. However, it remains an open question what motivates the transformation of the mixed motive game into a coordination game. Sanctions are one possible cause, but Bicchieri thinks that normative

[10] Vanderschraaf (2006), in contrast, thinks that a norm can turn the game into an "assurance game" (Sen 1967), also called a "stag hunt" (Skyrms 2004, referring to Rousseau 1984).

expectations alone can also give reasons to comply. The issue of normative reasons will resurface in the analysis of commitment in the next section.

Commitment

Amartya Sen criticises standard RCT for advocating a view of human agents as "rational fools" (Sen 1977). As we have seen, the standard model of RCT assumes that agents always act to maximise their personal welfare. Sen thinks that this view is too simplistic. By introducing the notion of commitment and contrasting it with the notion of sympathy, he shows how personal choice and personal welfare come apart.

If an agent is motivated by *sympathy*, she cares for another agent because seeing the other agent suffer decreases her welfare. In the case of sympathy, welfare and personal choice are aligned. By contrast, if an agent is *committed* to help another agent, she provides help even though it does not increase her own welfare, and may well reduce it. Welfare and personal choice come apart. Sen asserts that committed agents are rational, and that rationality is therefore not equivalent to the maximisation of one's own welfare. More controversially, Sen also claims that a committed agent may not even act to pursue her own goals and still be rational. He offers an example (2007, p.348): You have the window seat on a plane. Your neighbour, playing a (in your opinion) silly video game, asks you to draw the window blind so that he can see his screen. You oblige, even though you would have preferred to see the sun, and even though you disapprove of your neighbour wasting his time with silly computer games. Sen claims that you show "socially normed behavior" (p. 349). The social norm prescribes not to frustrate the goals of other people unnecessarily. You, by shutting the window blind, do not maximise your own welfare, nor are you following your own goals. However, even though you helped your neighbour to pursue his goal, it would be wrong to say that his goal has become your goal: you would rather see him read the New York Times than play video games. But you are willing to restrain the pursuit of your own goals because you are committed to a social norm of tolerance and helpfulness.

Sen gives many examples of committed action, among them voting, contributing to public goods, activities to protect the environment, cooperating in a game with the payoffs of a PD, and many instances of moral action. If Sen is right, almost all norm-guided behaviour, and in particular moral behaviour, is motivated by commitment, not self-interest. This is why standard RCT fails where social or moral norms matter, i.e. in most areas of human interaction, with the possible exception of some economic interactions. The standard model of RCT, with its limitation to the self-interested homo economicus, fails to address important factors of human behaviour, in particular being motivated by social and moral norms. The failure is both descriptive and normative: On the descriptive side, RCT is unable to explain and predict committed behaviour. On the normative side, RCT recommends an impoverished notion of rationality.

It is important to distinguish Sen's notion of commitment from other uses of the term. Schelling (1960) and Elster (1979) talk about *causal commitment devices*. Elster uses Homer's famous example of causal commitment: Ulysses is sailing home to Ithaca. En route his ship will pass the Sirens' island. Ulysses knows that once he and his sailors hear the song of the Sirens, they will not want to stay their course, lured away by their voices. To prevent this, Ulysses stops the ears of his sailors with wax and has himself tied to the mast. This ensures that Ulysses's ship sails on to Ithaca. In this example, Ulysses physically restricts his set of options in the future, taking the option to change course off the table. Similar causal commitments can be achieved if the subject can change future preferences such that it leads to the preferred future action.

Schelling's and Elster's causal commitment is easy to model in a decision theoretic framework: this is a sequential decision where the agent first decides whether she chooses to use the commitment device, and the future decision nodes are changed accordingly, i.e. options are unavailable or payoffs differ (see Güth & Kliemt 2007). But this is not the commitment that Sen has in mind. For Sen, a commitment is to certain *normative reasons* other than maximizing one's personal welfare, and it may conflict with the goal of maximising one's welfare. This conflict cannot adequately be modelled in a single all-things-

considered preference ranking. While one could include the effects of commitment into the agent's preferences, such that the preferences reflect the choices after the consideration of commitments, this approach cannot account for the possible conflict between self-interest and commitment.

In his 1977 paper Sen proposes to add more structure to agents' preferences. Each agent should have several preference orderings, and these orderings should be ordered in a meta-ordering, according to Sen. For instance, an agent may have an ordering of alternatives according to his narrow self-interest, a second ordering based on sympathy for others, and a third ordering that respects relevant norms and commitments. In addition, the agent also ranks these different preference orderings according to which ordering is most preferable to act upon. While this added structure allows the modeller to capture commitment, it has an important drawback: it sacrifices the notion of unified, all-things-considered, action-guiding preferences (Hausman 2007). Sen's richer model leaves it open how agents derive choices from their richer preference structures, while the standard model has a clear answer to that question: Agents do what they prefer most, according to their preferences.

The issues raised by Sen have been taken up in debates on dynamic choice and the rationality of plans. Edward McClennen (1990), David Gauthier (1986, 1997), and Michael Bratman (1987) argue that acting according to plans is a core feature of human rationality, and they attempt to revise RCT to accommodate for plans. McClennen (1990, 2004) proposes a theory of "resolute choice" to accommodate planning and rationality. Resolute choice is a mode of deliberation that allows agents to make plans and stick to them, even if it requires rejecting alternatives that are preferred while the agent follows the plan. This means that agents can plan a certain course of actions over time and stick to this plan in future choice situations; even though they may prefer to abandon the plan once they face the choice. David Gauthier develops related ideas regarding interpersonal choices in his "Morals by Agreement" (1986) and discusses intrapersonal resolute choice in subsequent work (e.g. Gauthier 1997).

One interesting question is which psychological mechanisms could allow agents to stick to plans and commitments. First, there is the option to develop dispositions and to internalise norms. Second, agents may be boundedly rational and stick to plans simply because a constant recalculation of utilities is too demanding. Bratman (1987) discusses this mechanism among others. Bounded rationality is a good explanation for the rationality of planning in some cases, but fares less well to explain commitments that are usually honoured not because agents are cognitively limited, but because they feel *obliged* to do so. This leads to the third mechanisms: Agents may stick to plans, and in particular commitments, because they have normative reasons to do so. Bicchieri endorses normative reasons as a motivation to comply with norms when she writes that a "reason for compliance with a norm is that one accepts others' normative expectations as well founded" (Bicchieri 2006, p. 23). Such normative reasons go beyond the instrumental rationality of RCT, but appear indispensable for a complete picture of human rationality (Sugden 1991; Hollis & Sugden 1993; Hausman & McPherson 2006, p.85-95; Verbeek 2007, see: Rational Agency). Following Sen, the failure of RCT is to either ignore normative reasons, or to trivialise them by subsuming them under one single preference ranking for each agent. The characterisation of normative reasons for committed action leads to difficult psychological and philosophical questions. Gibbard (1990, p. 30) remarks that "the relevant psychology [of norms] is not sitting neatly arranged on library shelves". The question how norms motivate also leads to intricate problems with regard to practical reasoning and metaethics, which are beyond the scope of this review (see Wallace 2008 for a survey). The tension between the focus on the individual rational agent on the one hand, and the desire to incorporate normative reasoning to aim for a richer, social notion of human rationality on the other, is a contemporary version of the earlier debates in sociology regarding the relation between the individual and the social, as examined above. We return to this question in the section on collective intentions.

4. Evolution and Cooperation

We have seen that one explanation of why social norms are beneficial is that they transform mixed-motive games into more cooperative games. However, showing that a rule is beneficial is not sufficient to explain the existence of a rule. One also needs to show how the rule came into being and how it was able to persist. Answers to the questions as to how norms evolve and how they are maintained can be addressed with tools borrowed from theoretical biology, in particular evolutionary game theory. Models that were originally developed to analyse the biological evolution of organisms are applied to related questions in the social sciences, leading to evolutionary models of cooperation and culture. This transfer raises difficult questions, but it has also sparked off an interesting and productive research literature on the evolution of human cooperation and norms (see: Evolutionary Approaches). We focus on a few models that aim to explain the evolution of norms and give one example as to how theoretical and empirical research from biological evolution, in particular evolutionary psychology, can matter for these models.

Evolutionary Game Theory

Martin A. Nowak succinctly summarises the approach taken in evolutionary game theory:

> "Evolutionary game theory does not rely on rationality. Instead it considers a population of players interacting in a game. Individuals have fixed strategies. They interact randomly with other individuals. The payoffs of all these encounters are added up. Payoff is interpreted as fitness, and success in the game is translated into reproductive success. Strategies that do well reproduce faster. Strategies that do poorly are outcompeted." (Nowak 2006, p. 46)

This approach is applicable to both biological and social contexts. Following Dennett (2006, p. 341), evolution is "substrate neutral" and "will occur whenever and wherever three conditions are met:

1. replication
2. variation (mutation)

3. differential fitness (competition)".

In biological evolution, a gene replicates through the offspring of its organism, variation is provided by recombination and mutation of genes, and differential fitness is the relative success to replicate compared to other genes, which in turn depends on the success of the organism to survive and replicate. Roughly speaking, genes of more successful organisms (where success means replicating the gene as often as possible) are selected for. For the social sciences, by contrast, the most likely unit of selection are patterns of behaviour (called strategies) and evolution happens through learning (Young 1998). The behavioural pattern is replicated if individuals learn a behavioural pattern from other individuals; variation is caused by mistakes (or learners try something new), and differential fitness is given by the competition between different patterns of behaviour.

One important (but by far not the only) question that can be addressed with evolutionary game theory is whether and under what conditions cooperative strategies can evolve. We briefly discuss two attempts to answer this question: Axelrod's "Evolution of Cooperation"[11] and models of cooperation based on indirect reciprocity and assortation.

To explain the emergence of cooperation, it is useful to focus on repeated games. Robert Axelrod (1984) conducted computer tournaments of different strategies for iterated PDs. Axelrod invited researchers to submit small computer programs that had to play 200 PDs against each other in a round-robin tournament. The programs had to decide between "cooperate" or "defect" in every PD and could use the outcome of previous games as input to decide on the next move. In Axelrod's tournaments, "TIT-FOR-TAT" emerged as the most successful strategy. TIT-FOR-TAT cooperates in the first round, and copies the last move of the opponent in all subsequent rounds. If two agents play TIT-FOR-TAT against each other, they cooperate in all rounds. If a player with TIT-FOR-TAT strategy plays against an opponent who defects in all rounds, he only gets exploited in the first round, but not in any further round. However, TIT-FOR-TAT is not the only successful or even the best strategy in

[11] Strictly speaking, Axelrod' original computer tournament does not apply evolutionary game theory in Nowak's sense, but it is inspired by concepts derived from evolutionary game theory.

iterated PDs.[12] The important insight from Axelrod's work is not the focus on TIT-FOR-TAT, but the fact that cooperation can emerge in iterated settings, and that successful cooperative strategies should be cooperative in a conditional way, that is they should "reward" cooperation and "punish" defection. In Axelrod's view, TIT-FOR-TAT is a rudimentary norm of reciprocity. For instance, Axelrod argues that a TIT-FOR-TAT norm evolved between French and German troops in trench warfare. Axelrod's computational results tie in with the game-theoretical analysis of iterated games. The folk theorem (e.g. Osborne & Rubinstein 1994, pp. 143-149) implies that mutual cooperation is one of many Nash equilibria in the infinitely repeated two-person PD, if the discount rate for future payoffs is low.

Axelrod's analysis is restricted to the prolonged and potentially infinite interaction between two individuals. However, regarding norms it is more fitting to consider repeated interactions between different people. We consider two approaches to explain cooperation in these settings: indirect reciprocity and assortation. Firstly, *indirect reciprocity* works if there is a public track record of how individuals behaved in the past (see Nowak & Sigmund 2005 for a review). Given this track record, non-cooperative behaviour can be reciprocated (this could be interpreted as "retaliation"), even though victim and reciprocator do not have to be identical. Apart from reciprocating non-cooperative behaviour, it is also possible, secondly, to exclude defectors and work towards an *assortation* of cooperators and defectors. The internet auction platform EBay is a good example for indirect reciprocity and assortation. It invites its customers to rate the behaviour of their trading partners. Having a good track record of previous trades is essential for doing business on EBay, and this creates an incentive to comply with the relevant social and legal norms of trading. Even though it would be beneficial for a rogue trader to cheat if he considered only the current round, it will damage the reputation of the trader in all future rounds, and will result in fewer trades, or even in reciprocal, retaliatory cheating by future trading partners.

[12] There are strategies that systematically outperform TIT-FOR-TAT (Nowak & Sigmund 1993). Also, TIT-FOR-TAT is very sensitive to trembling and mistakes: A single mistake can lock two TIT-FOR-TAT players into a vicious circle of retaliation (Fudenberg & Maskin 1990).

Brian Skyrms's influential "Evolution of the Social Contract" (1996) has popularised evolutionary models of norm emergence. The idea is that norms can be universal replicators in Dennett's sense: norms replicate through learning, mutation happens by mistakes in the transmission (or attempts to try something new), and differential fitness is given by the relative success or failure of a norm to spread through a society. The idea is that some norms are more easily learned or transmitted than others. This is particularly plausible if some norms create more utility for agents than other norms, such that individuals learn to follow the "high-utility norm".

While these models are highly simplified, they can give indications as to how certain norms may have evolved and how stable they are. Models of cultural evolution are most easily applied to rules of prudence and technological know how (Sterelny 2006). For instance, Henrich and McElreath (2003) describe how Australian Aborigines have developed elaborate techniques to gather and process food to survive in a scarce environment. These rules of prudence ("this is how you hunt a fish", "this is how *nardoo* seeds are processed") are successfully passed on between members of these societies and from generation to generation because they are useful for the agent who knows these rules. The evolutionary perspective is apt because the rate of replication (i.e. transmission) for rules of prudence is likely to be positively correlated with how useful the rule is. In the case of social norms it is less clear how the content of a norm relates to the rate of its replication. Skyrms (1996), Dennett (2006), Binmore (2005) and Alexander (2007) develop explanatory evolutionary models of normative content with regard to norms of distributive justice, mutual aid, and even religion. Research in this area is still at an early stage and it is unlikely that evolutionary game theory can fully capture the rich processes involved in the emergence of norms. Alexander (2003, section 4.2) offers a sceptical outlook:

> "Although an evolutionary game theoretic model may exclude certain historical sequences as possible histories (since one may be able to show that the cultural evolutionary dynamics preclude one sequence from generating the phenomenon in

question), it seems unlikely that an evolutionary game theoretic model would indicate a unique historical sequence [that] suffices to bring about the phenomenon."

Apart from explanatory underdetermination, there are at least three further conceptual problems. Firstly, we currently have a very limited understanding as to what makes one norm "fitter" than another. Secondly, it is unclear whether the unit of selection should be norms, systems of norms, or perhaps even societies applying norms. Thirdly, if evolutionary models are applied to norms, it is difficult to disentangle genetic and cultural effects. Recent movements towards multilevel selection and gene-culture co-evolution try to address these difficulties (Richerson, Boyd, & Henrich 2006).

Evolutionary Models and Empirical Support

There are many links between evolutionary theory and human behaviour in general (Laland & Brown 2002). While evolutionary game theory is a primarily theoretical undertaking, other approaches take a more empirical route. One important field within biological evolutionary theory relevant for the analysis of cooperation and norms is *evolutionary psychology*. Briefly put, evolutionary psychology assumes that human minds have evolved in an environment of evolutionary adaptedness (primarily hunter-and-gatherer societies of the Pleistocene); that evolution has therefore created adapted brain "mechanisms" or "modules" to solve certain groups of problems (thereby rejecting the claim that the brain has evolved as a general all-purpose reasoning device); and that these modules still influence human cognition and behaviour today, such that testable predictions can be made (Barkow, Cosmides & Tooby 1992). One example for a potentially evolved mechanism is *cheater detection*. Since our ancestors in the environment of evolutionary adaptedness frequently encountered dilemmas of cooperation, and since these dilemmas can be solved more efficiently if it is possible to identify cheaters, one can expect such an evolved mechanism for cheater detection. There is some evidence for the existence of such a module: Experiments shows that people are much better in solving cognitive tasks when these tasks are framed in the form of a cheater detection problem, compared to a logically equivalent task framed in

different ways (Cosmides & Tooby 1992). In a related body of work, the economist Robert Frank (1988, ch. 3) argues that human *emotions* function as commitment devices to avoid cheating. He argues that cooperators and defectors send out different emotional signals. By picking up these emotional signals cooperators can recognise each other and cooperate with their own kind, while defecting against defectors. Theoretical and empirical studies support the claim that emotions facilitate cooperation in dilemma situations (Sally 2000) and that even short ex ante interactions allow agents to predict with better-than-random probability whether their opponents will cooperate or defect in a PD (Frank, Gilovich & Regan 1993). From an evolutionary point of view, being able to commit through emotions can increase fitness because cooperation gains can be accrued. However, Frank notes that "mimicry" results in even higher fitness: An agent who can pretend to be committed is able to defect and exploit the other committed agents, underlining the need for cheater detection.

The models and approaches discussed can only give a glimpse of manifold attempts to link evolutionary thinking with the social sciences. Such models can be of help to explain the emergence and stability of norms. Attempts to link theoretical models with empirical work (such as the psychological experiments to corroborate theories of evolutionary psychology) seem particularly promising because they provide empirical micro-foundations for an otherwise theoretical and frequently speculative literature.

5. Collective Intentions

The rational choice approaches discussed in section 3 shed only limited light on the social processes that lead to the emergence and sustenance of social rules, especially of social norms. Many have argued that this limitation is a result of restricting the analysis to the consequences of individual intentions. At the same time, many are reluctant to give up the ontological commitment to individualism and to embrace a Durkheimian collectivism. Some have started thus to explore the question whether there is a form of intentional analysis which could complement the analysis based on individual intentions. What these scholars suggest is

that social rules are, at least in part, the result of collective intentions (Gilbert 1989, 2007; Tuomela 1984, 2000; Hollis and Sugden 1993; Searle 1995; Bratman 1999). Their focus is on non-summative accounts of collective intentions, or "we-intentions". On the summative account, collective intentions are simply the sum of individual intentions, plus a common knowledge assumption. Such an account is both too weak and too strong. It is too strong, because it demands that all members of a collective have a particular intention for there to be a collective intention. But an utterance such as "Warwick University is committed to a high standard of excellence in research" may be meaningful without it being true that all of its members are individually so committed. And it is too weak, because it assumes that intentions are necessarily "I-intentions" and neglects the possibility of intentions at the collective level.

Collective Intentions and Social Rules

There are two main arguments about the relation between social rules and collective intentions. The first relates to the creation of social facts. The starting-point is the claim that social groups create social facts which are not explicable in terms of individual intentions. Consider the following simple example (Sugden 2000). You are a member of a group of friends, who originally organized regular trips to explore new pubs and sample little-known beers. While the main activities of the group may have shifted over time, it is quite possible that, when they are out as a group, beer remains the preferred drink, even if, individually, they all prefer wine over beer. John Searle defends the more general claim that the creation and sustenance of all those facts which are not brute facts, i.e. facts which do not exist independently of human institutions, depend on collective intentions. His argument links collective intentions to constitutive rules ("X counts as Y in context C"). In the first instance, collective intentions assign functions to things that have nothing to do with their physical properties. An example is to use the shadow of a tree as a classroom. Another of Searle's examples, as already mentioned, is a piece of paper that gets assigned the function of money. Beyond this initial assignment of function, collective intentions formulate and support

constitutive rules. As such, they ensure that these functions gain permanence and help create social institutions.[13]

The second argument about the relation between collective intentions and social rules relates to the creation of normativity, i.e. to the explanation of how certain social rules acquire binding force. The most prominent advocate of this argument is Margaret Gilbert (1989; 2006). According to Gilbert, collective intentions are the product of social groups. She calls such groups "plural subjects" (Gilbert 1989: 18). A group of individuals "constitute a plural subject … if and only if they are jointly committed to doing something as a body" (Gilbert 2006: 145). The idea of a "joint commitment" of all group members, which is necessary for collective intentions in her sense, also entails an account of how social rules acquire normative status and become action-guiding. As she puts it (Gilbert 1989: 411):

> "Being a group member takes work. … In order to enter a group … one must give over one's will to a sum or pool of wills which is itself dedicated to some cause… This entails taking on or accepting a new set of constraints on one's behaviour. (One also accepts certain new entitlements.)"

According to her, social rules are "of the fiat form"; they are "rules which we as group members prescribe for ourselves" (Gilbert 1989: 387). "The fiat forms is … the expected form for rules whose force is seen as deriving from judgment or will" (1989: 400). Her point is that this "fiat form" of rules created by collective intentions constitutes a source of obligations which is different from moral obligations and prudential recommendations. Because it includes an explanation for the special normative force of social norms, Gilbert's account differs not just from Searle's take on collective intentions and their relationship to social rules, but also from Lewis' account of conventions, or other accounts based on game theory. She objects to these accounts that they capture mere regularities. Her account, she argues, focuses on the binding rules that social groups impose on themselves and can thus explain how these rules become action-guiding.

[13] For a related argument, see Tuomela (1995).

The persuasiveness of arguments for the importance of collective intentions depends on how plausible the concept itself is. How should one make sense of the very concept? David Velleman (1997) has a particularly clear presentation of the problem. Take Michael Bratman's (1984) distinction between intentions as goals and intentions as plans as a starting-point. An agent may have two mutually exclusive goals and let the world decide between them. This is not possible in the narrower sense of intentions as plans. Interpreted as plans, intentions refer to things that are up to the agent. An agent cannot rationally plan to pursue two mutually exclusive outcomes. Intentions interpreted in this narrower sense raise the following challenge for theories of collective intentions:

> "how can I frame the intention that 'we' are going to act, if I simultaneously regard the matter as being partly up to you? And how can I continue to regard the matter as partly up to you, if I have already decided that we really are going to act? The model seems to require the exercise of more discretion than there is to go around" (Velleman 1997: 35).

If collective intentions are interpreted in the narrow sense of plans, this implies that there has to be one token intention.

This interpretation of collective intentions rules out summative accounts of collective intentions and points to the need of a non-summative account. But it also rules out some non-summative accounts, such as the one put forward by Searle (e.g. Searle 1995). Searle takes "we-intentions" to be a biologically primitive phenomenon, located in individual brains. Individuals are thus capable of forming two types of intentions: one type takes the form "I intend" and the other "we intend", and neither is reducible to the other. On a view of this sort, there is collective action based on collective intentions if each individual member of the collective forms the corresponding we-intention in his or her brain. That is to say, each individual holds the same token we-intention, but there is no single token intention at the collective level.

Gilbert's "plural subject theory", by contrast, is compatible with the interpretation of intentions as plans. Gilbert, as discussed, insists on the obligation-generating force of collective intentions. In Searle's account, the coordination and cooperation that is necessary

to create and sustain social rules happen as long as the we-intentions of different individuals happen to coincide. Such individuals can thus not "think as a team" (Schmid 2003). In Gilbert's account, once different groups members have expressed the joint commitment that constitutes a "plural subject", they are bound to perform their part of the collective action until they jointly rescind their commitments (Gilbert 2006, p. 141ff.).[14]

Collective Intentions and Commitment

An analysis based on collective intentions may help shed light on individual's motivations to follow norms and conventions. Such an analysis shifts the perspective from the question "what should *I* do?" to the question "what should *we* do?" As such, the analysis contrasts with rational choice approaches to conventions and norms, as discussed in section 3. Both the standard model of rational choice theory and extensions such as Bicchieri's only invoke I-intentions. The advantage of moving to the analysis of collective intentions is that the question "what should we do?" suggests natural solutions to coordination problems such as the Hi-Lo game or cooperation problems such as the PD that are not available to those who merely ask "what should I do?". More generally speaking, it suggests that committing to rules may be the rational thing to do in many social situations.

Does this shift of perspective necessarily imply that an analysis of social rules based on collective intentions is in conflict with rational choice theory – and perhaps provides an alternative to it – or can the two be integrated? The answer depends, again, on how collective intentions are interpreted. In the following we want to briefly discuss two opposing answers to this question. Elizabeth Anderson (2001) argues that Gilbert's account provides a strategy for overcoming the limitations of standard rational choice theory and points the way towards an alternative theory of rational action. Robert Sugden (2000), meanwhile, rejecting Gilbert's idea that collective intentions generate obligations, argues that collective intentions can be incorporated in rational choice theory.[15]

[14] There are important objections to Gilbert's theory, most importantly that it is circular, as the concept of quasi-readiness already invokes some form of we-intentions. See Tuomela (1984, 1995), Velleman 1997, Bratman 1999) for discussions and alternative accounts. We lack space to pursue this issue any further.

[15] For a comprehensive discussion which covers a range of approaches, see also Gold and Sugden (2007).

Anderson's starting-point is Sen's concept of commitment. She tries to show that the perspective of we-intentions allows for an account of what makes committed action rational: "committed action turns out to be action on principles (reasons) that it is rational for *us* to adopt, and thus that it is rational for any individual who identifies as a member of that group to act on" (Anderson 2001, p. 24). In a first step, she argues that the recommendations of standard rational choice theory, which focus on what is rational for an isolated individual, must be rejected when individuals identify as members of a group. Individuals who do indentify as members of a group should reason in terms of strategies which make sense for them as a group. Next, she combines this with Gilbert's interpretation of collective intentions as obligations-generating. This yields an explanation for why individuals who reason in this way end up being rationally committed to conform to a social norm such as "reciprocate favours". Finally, Anderson stresses that this model of practical reasoning is compatible with the Kantian idea that moral action is continuous with rational action. This is not to say that all principles that members of a group might regard as rational to adopt for them amount to moral norms. There is thus still a difference between social and moral norms. But those principles that are rationally adopted from a universal perspective are moral norms: "If it would be rational for a collective to encompassing all of humanity to adopt a certain principle of committed action, then action on that principle is morally right" (Anderson 2001: 24).

Let us grant that the "we-perspective" provides a helpful answer to the question as to what might motivate individuals to act according to principles in coordination or cooperation games which, if they look at the situation from an isolated perspective, they will not be inclined to adopt. But does that indeed necessitate a radical departure from rational choice theory, as Anderson suggests? Sugden (2000) argues that this is not the case. His theory of "team agency", he claims, is compatible with a generalized version of rational choice theory. That is to say, the received interpretation, which focuses on the reasoning of isolated individuals, is just the special case of a team that only includes one member.

Sugden rejects Gilbert's normative take on collective intentions. He argues, instead, that both the existence of a "team" and its objectives are empirical issues.[16] His theory of team reasoning starts with individuals who take themselves as members of a team – e.g. members of a football team, or members of a family. If all individuals have some confidence that the team actually exists, Sugden claims, they will be prepared to engage in "team-directed reasoning". In a coordination game, for example, such as the Hi-Lo game, team-directed reasoning assigns a single utility index to each outcome, as opposed to separate utility-indices for each individual. Individuals thus do not approach the game by asking what is rational for them, individually, to do, but what it is rational for them, as a team, to do. As such, team-directed reasoning escapes the infinite regress that individual-directed preferences may generate for coordination games.[17] Team reasoning and team agency, then, become possible if each member of a team engages in such team-directed reasoning, and each is confident that the team exists and that each member will do his or her part. Under these conditions, what individuals are motivated to do derives from team preferences and team reasoning. This approach, too, has thus an answer to the question why individuals might be ready to commit to conform to social norms. But it rejects the idea that such a commitment may be binding.

References

Alexander, J. M. (2003). Evolutionary Game Theory. In E. N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy* (Summer 2003 edition), from: http://plato.stanford.edu/archives/sum2003/entries/game-evolutionary/.

Alexander, J. M. (2007). *The Structural Evolution of Morality*. Cambridge: Cambridge University Press.

Anderson, E. (2000). Beyond Homo Economicus: New Developments in Theories of Social Norms. *Philosophy and Public Affairs*, 29, 170-200.

[16] Susan Hurley (1989), by contrast, argues that rationality decides what the appropriate unit of agency should be.

[17] See also Hollis and Sugden (1993) and Gold and Sugden (2007).

Anderson, E. (2001). Unstrapping the Straitjacket of 'Preference': on Amartya Sen's Contributions to Philosophy and Economics. *Economics and Philosophy*, 17, 21 – 38.

Axelrod, R. M. (1984). *The evolution of cooperation*. New York: Basic Books.

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

Barkow, J. H., Cosmides, L. & Tooby, J., Eds. (1992). *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. New York & Oxford: Oxford University Press.

Bicchieri, C. (2006). *The Grammar of Society. The Nature and Dynamics of Social Norms*. Cambridge: Cambridge University Press.

Binmore, K. (2005). *Natural Justice*. Oxford: Oxford University Press.

Bohman, J. (1993). *New Philosophy of Social Science*. Cambridge, MA: MIT Press.

Bratman, M. (1987). *Intentions, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press.

Bratman, M. (1993). Shared Intention. *Ethics*, 104, 97 – 113.

Bratman, M. (1999). *Faces of Intention*. Cambridge: Cambridge University Press.

Brennan, G. & Pettit, P. (2004). *The Economy of Esteem: An Essay on Civil and Political Society*. Oxford: Oxford University Press.

Cinyabuguma, M.; Page, T. & Putterman, L. (2005). Cooperation under the threat of expulsion in a public goods experiment. *Journal of Public Economics*, 89, 1421-1435.

Coleman, J. S. (1990). *Foundations of Social Theory*. Cambridge: Harvard University Press.

Cosmides, L. & Tooby, R. (1992). Cognitive Adaptations for Social Exchange. In J. H. Barkow, L. Cosmides & J. Tooby (Eds.), *The Adapted Mind: Evolutionary Psychology and the Generation of Culture* (pp. 163-228). New York & Oxford: Oxford University Press.

Cubitt, R. P. & Sugden, R. (2003). Common Knowlegde, Salience and Convention: A Reconstruction of David Lewis' Game Theory. *Economics and Philosophy,* 19, 175-210.

Dennett, D. (2006). *Breaking the Spell: Religion as a Natural Phenomenon*. London: Penguin.

Durkheim, E. (1938). *The Rules of Sociological Method* (8th translation). London: Macmillan.

Durkheim, E. (1951). *Suicide*. New York: The Free Press.

Elster, J. (1979). *Ulysses and the Sirens*. Cambridge: Cambridge University Press.

Elster, J. (1989). *The Cement of Society*. Cambridge: Cambridge University Press.

Elster, J. (2000).*Ulysses Unbound*. Cambridge: Cambridge University Press.Giddens,. (1984). *The Constitution of Society: Outline of a Theory of Structuration*. Cambridge: Polity Press.

Frank, R. H. (1988). *Passions within Reason: The Strategic Role of the Emotions*. New York & London: Norton.

Frank, R. H., Gilovich, T. & Regan, D. T. (1993). The Evolution of One-Shot Cooperation - an Experiment. *Ethology and Sociobiology*, 14, 247-256.

Fudenberg, D. & Maskin, E. (1990). Evolution and Cooperation in Noisy Repeated Games *The American Economic Review,* 80, 274-279.

Gauthier, D. (1986). *Morals by Agreement*. Oxford: Clarendon.

Gauthier, D. (1997). Resolute Choice and Rational Deliberation: A Critique and a Defense. *Noûs,* 31, 1-25

Gibbard, A. (1990). *Wise Choices, Apt Feelings: A Theory of Normative Judgment*. Oxford: Clarendon Press.

Giddens, Anthony. 1984. *The Constitution of Society. Outline of the Theory of Structuration*. Cambridge: Polity Press.

Gilbert, M. (1989). *On Social Facts*. Princeton: Princeton University Press.

Gilbert, M. (2006). *A Theory of Political Obligations*. Oxford: Oxford University Press.

Gold, N. and Sugden, R. (2007). Theories of Team Agency. In Peter, F. and Schmid, H. B. (Eds.) *Rationality and Commitment* (pp. 280 – 312). Oxford: Oxford University Press..

Güth, W. & Kliemt, H. (2007). The Rationality of Rational Fools: The Role of Commitments, Persons, and Agents in Rational Choice Modelling. In F. Peter & H. B. Schmid (Eds.), *Rationality and Commitment* (pp. 124-149). Oxford: Oxford University Press.

Habermas, J. (1984). *The Theory of Communicative Action. Volume 1*. T. McCarthy (transl.). Boston: Beacon Press.

Hart, H.L.A. (1961). *The Concept of Law*. Oxford: Oxford University Press.

Hausman, D. & McPherson, M. S. (2006). *Economic Analysis, Moral Philosophy, and Public Policy*. 2nd ed. Cambridge: Cambridge University Press.

Hausman, D. (2007). Sympathy, Commitment, and Preference. In F. Peter & H. B. Schmid (Eds.), *Rationality and Commitment* (pp. 49-69). Oxford: Oxford University Press.

Hechter, M. & Kanazawa, S. (1997). Sociological Rational Choice Theory. *Annual Review of Sociology*, 23, 191-214.

Henrich, J. & McElreath, R. (2003). The evolution of cultural evolution. *Evolutionary Anthropology: Issues, News, and Reviews*, 12(3), 123-135.

Hollis, M. & Sugden, R. (1993). Rationality in Action. *Mind*, 102 (405), 1-35.

Hume, D. (1978). *A Treatise of Human Nature*, Ed. L. A. Selby-Bigge, 2nd ed., Oxford: Clarendon.

Hurley, S. (1989). *Natural Reasons.* Cambridge: Harvard University Press.

Laland, K. N. & Brown, G. R. (2002). *Sense & Nonsense: Evolutionary Perspectives on Human Behaviour*. Oxford: Oxford University Press.

Lewis, D. (1969). *Convention*. *A Philosophical Study.* Cambridge, MA: Harvard University Press.

McClennen, E. F. (2004). The Rationality of Being Guided by Rules. In A. R. Mele & P. Rawling (Eds.), *The Oxford Handbook of Rationality* (pp. 222-239). Oxford: Oxford University Press.

Miller, S. (2001). *Social Action: A Teleological Account*. Cambridge: Cambridge University Press.

Nowak, M. A. (2006). *Evolutionary Dynamics: Exploring the Equations of Life*. Cambridge, MA: Belknap Press.

Nowak, M. & Sigmund, K. (1993). A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game. *Nature,* 364, 56-58.

Nowak, M. A. & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature,* 437 (7063), 1291-1298.

Osborne, M. J. & Rubinstein, A. (1994). *A Course in Game Theory*. Cambridge, MA & London: MIT Press.

Parsons, T. (1968). *The Structure of Social Action*. New York: Free Press.

Pettit, P. (1993). *The Common Mind*. Oxford: Oxford University Press.

Pettit, P. (2000). Winch's Double-Edged Idea of a Social Science. *History of the Human Sciences*, 13(1), 63 – 77.

Pettit, P. (2002). *Rules, Reasons, and Norms*. Oxford: Oxford University Press.

Rawls, J. (1955). Two Concepts of Rules. *Philosophical Review,* 64, 3 – 32.

Richerson, P. J., Boyd, R. T. & Henrich, J. (2006). Cultural Evolution of Human Cooperation. In P. Hammerstein (Ed.), *Genetic and Cultural Evolution of Cooperation* (pp. 357-388). Cambridge, MA: MIT Press.

Rousseau, J.-J. (1984). *A discourse on inequality*. Transl. M. Cranston. Harmondsworth: Penguin.

Sally, D. (1995). Conversation and Cooperation in Social Dilemmas: A Meta-Analysis of Experiments from 1958 to 1992. *Rationality and Society,* 7, 58-92.

Sally, D. (2000). A general theory of sympathy, mind-reading, and social interaction, with an application to the Prisoners' Dilemma. *Social Science Information,* 39, 567-634.

Savage, L. J. (1954). *The Foundations of Statistics*. New York: John Wiley.

Schelling, T. C. (1960). *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.

Schmid, H. B. (2003). Can Brains in Vats Think as a Team? *Philosophical Explorations*, 6(3), 201 – 217.

Searle, J. R. (1995). *The Construction of Social Reality*. New York: Free Press.

Sen, A. (1977). Rational Fools: A Critique of the Behavioral Foundations of Economic Theory. *Philosophy and Public Affairs,* 6, 317-344.

Sen, A. (2007). Rational Choice: Discipline, Brand Name, and Substance. In F. Peter & H. B. Schmid (Eds.), *Rationality and Commitment* (pp. 339-361). Oxford: Oxford University Press.

Skyrms, B. (1996). *Evolution of the Social Contract*. Cambridge: Cambridge University Press.

Skyrms, B. (2004). *The Stag Hunt and the Evolution of Social Structure*. Cambridge: Cambridge University Press.

Spiekermann, K. (2007). Integrity, Translucency and Information Pooling: Groups Solve Social Dilemmas. *Politics, Philosophy & Economics*, 6(3), 285-306.

Sterelny, K. (2006). Memes Revisited. *British Journal for the Philosophy of Science*, 57, 145-165.

Sugden, R. (1991). Rational Choice: A Survey of Contributions from Economics and Philosophy. *The Economic Journal*, 101, 751-785.

Sugden, R. (2000). Team Preferences. *Economics and Philosophy*, 16, 175 – 204.

Taylor, C. (1971). Interpretation and the Sciences of Man. *Review of Metaphysics* 25: 3 – 51.

Tuomela, R. (1984). *A Theory of Social Action*. Dordrecht: Kluwer.

Tuomela, R. (1995). *The Importance of Us: A Philosophical Study of Basic Social Notions*. Palo Alto: Stanford University Press.

Ullmann-Margalit, E. (1977). *The Emergence of Norms*. Oxford: Clarendon.

Vanderschraaf, P. (2006). War or Peace? A Dynamical Analysis of Anarchy. *Economics and Philosophy*, 22, 243-279.

Velleman, D. (1997). How to Share an Intention. *Philosophy and Phenomenological Research*, 57, 29 – 50.

Verbeek, B. (2007). Rational Self-Commitment. In F. Peter & H. B. Schmid (Eds.), *Rationality and Commitment* (pp. 150-174). Oxford: Oxford University Press.

Wallace, R. J. (2008). Practical Reason. In: E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2008 Edition), from: http://plato.stanford.edu/archives/fall2008/entries/practical-reason/.

Weber, M. (1947). *The Theory of Social and Economic Organization*. Edited by Talcott Parsons. New York: Free Press.

Winch, P. (1958). *The Idea of a Social Science and its Relation to Philosophy*. 2nd Edition. London: Routledge.

Young, H.P. (1998). *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions*. Princeton: Princeton University Press.