Viktor Vanberg/James M. Buchanan

Rational Choice and Moral Order

> "The problem which science has to solve here consists in the explanation of a social phenomenon, of a homogeneous way of acting on the part of the members of a community for which public motives are recognizable, but for which in the concrete case individual motives are hard to discern."
> (Carl Menger 1985, 152)

Abstract: The article discusses some of the fundamental conceptual and theoretical aspects of rational choice and moral order. A distinction is drawn between constitutional interests and compliance interests, and it is argued that a viable moral order requires that the two interests somehow be brought into congruence. It is shown that with regard to the prospects for a spontaneous emergence of such congruence, a distinction between two kinds of moral rules which we call trust-rules and solidarity-rules is of crucial importance.

I. Introduction

The crucial dependence of the character of a social and economic order on the framework of rules and institutions within which individuals act and interact has been a central theme of classical political economy, and it is a theme that has gained renewed attention in modern economics (Brennan/Buchanan 1985). One of the persistent issues in this context concerns the tension that is perceived to exist between rational, self-interested behavior - as postulated in economics - and the viability of a moral order. While the general benefits that a moral order generates are quite obvious, it is far less obvious how rational pursuit of self-interest should induce the kind of conduct that such an order requires.

Our purpose in this paper is to discuss some of the fundamental conceptual and theoretical aspects of the 'rational choice and moral order' issue, an issue that, at least since Thomas Hobbes, has plagued social theorists. In fact, it is often referred to as the 'Hobbesian problem of social order' or simply the 'Hobbesian problem'. In sociology, an influential theoretical program - associated with the names of Emile Durkheim and Talcott Parsons - is even based on the claim that the economic, individualistic-utilitarian tradition has not provided a satisfactory solution to the Hobbesian problem and, for intrinsic reasons, is unable to do so. Our purpose here is to

Analyse & Kritik 10 (1988), S. 138–160

argue the opposite; to show that, and how, an answer to the Hobbesian question can be provided from an individualistic, rational choice perspective.

The paper is organized as follows: Sections II and III analyze the 'rational choice and moral order' issue in terms of the contrast between what we call constitutional interests, on the one side, and action interests or compliance interests, on the other. We argue that the practical solution to the 'problem of social order' is that of bringing people's action interests into congruence with their constitutional interests. Sections IV and V discuss the issue of how such congruence may be spontaneously brought about as a by-product of incentives that are 'naturally' generated in the process of social interaction. Of particular interest in this context is the mechanism of reciprocity. Sections VI and VII are about a distinction between two kinds of rules that we call trust-rules and solidarity-rules, the difference between which is of crucial relevance to the issue of spontaneous emergence of a moral order. Section VIII elaborates on some implications of our analysis.

II. Moral Rules, Constitutional Interests, and Action Interests

Explanatory accounts of moral rules are not always sufficiently careful to avoid the 'functionalist fallacy'. This fallacy consists in assuming that by identifying the 'benefits' that a moral code provides to a group (community, social system, etc.), one has provided an answer to the question of why the respective code is honoured. The functionalist fallacy[1] is tempting because it seems quite natural to presume that the beneficial consequences of rules and institutions must have something to do with the fact that they exist and persist. The problem with the functionalist fallacy is not its focus on the beneficial consequences or, in its own terminology, on the functions that rules and institutions serve within a group. The problem rather is that the functionalist linkage provides no more than the illusion of an explanation and distracts attention from the genuine challenge which is that of identifying the actual processes or mechanisms that establish the critical linkage between beneficial consequences and effective causes for behaviorally generated rules and institutions.

In discussions on the rationale for and the effective causes of moral rules, the familiar contrasts between 'individual and group interests', 'private and common interest', or 'individual and collective interest' are potentially misleading because these tend to suggest that a conflict of interests is at issue, either of interests experienced by differents entities ('the individual' and 'the group') or interests that have to be traded off within the behavioral calculus of a single person (a 'narrow' self-interest and a

more 'noble' common interest). We propose an alternative interpretation that differentiates between different levels of choice and between different kinds of interests that are related to these levels of choice. In our framework, the issue is not one of conflicting preferences with regard to the same kind or level of choice; the issue is, by contrast, one of different kinds of choice, which, in turn, involve different interests. We separate, define, and contrast two kinds of individual interests: (1) constitutional or rule interests, and (2) operational or action interests.[2] An actor's constitutional interests are reflected in his preferences over potential alternative 'rules of the game' for the social community or group within which he operates. His constitutional interests inform his choices insofar as these choices pertain to the kind of institutional order or order of rules under which he is to live. Or, stated somewhat differently, they reflect preferences that would emerge if he were to participate in choosing the constitution, in the broadest sense, for his respective social community. By comparison, a person's operational or action interests are reflected in preferences over potential alternative courses of action under given situational constraints, including the constraints that pertain to the given structure of rules and institutions.

Constitutional interests and operational interests, as defined here, are experienced by the same person, and possibly simultaneously, yet there is no reason to expect that these interests will be either in 'natural harmony' or 'natural conflict'. As mentioned, trade-offs of the ordinary sort are not relevant here because these separate interests reflect an actor's preferences over different kinds of alternatives: Constitutional interests concern the imagined or real choice among alternative institutional characteristics of one's social environment. They reflect, in other terms, a person's preferences over alternative institutional environments. Operational or action interests concern alternative courses of action within a given environment. They reflect a person's behavioral preferences under given environmental constraints. Whether these two interests are in congruence or harmony in the sense that a person prefers to comply with a rule that he prefers constitutionally, is an empirical question the answer to which will depend on certain characteristics of the relevant social environment. More importantly, a person's constitutional interests do not automatically translate into corresponding action interests. My interest in living in a community where promises are kept, for instance, does not per se imply that I must have an interest in always keeping promises on my part. There is nothing 'inconsistent' in preferring a certain rule constitutionally and, at the same time, given the situational constraints as they are, violating the rule in pursuit of one's action interests.[3]

Presupposing that there are certain rules on which people's constitutional interests converge, the central issue that a 'theory of social order' has to deal with concerns the social forces or mechanisms that tend to bring

constitutional interests and action interests into correspondence. Such correspondence is needed in order for a generally preferred constitutional order to be operative. Accordingly, the analytical focus has to be on the reasons why and the conditions under which individuals can be expected to comply with rules that are in their constitutional interest.

III. Constitutional Interests and Moral Philosophy

The disjunction between people's constitutional interests and their operational or action interests, though not stated in these terms, has long been a puzzling issue to moral philosophers. Kant's reflections on the 'categorial imperative', for instance, lend themselves to be interpreted as reflections about constitutional interests, independent of whether this may pass as an 'authentic' interpretation or not. He examines arguments that might guide people's constitutional interests in terms of preference for rules which could qualify as general laws. He does not, however, examine the reasons that make people adopt as private maxims of behavior the general rules that they want to see practiced in the community. In our terms, Kant does not explain how people's action or compliance interests are supposed to come into congruence with their constitutional interests.[4]

More recently, David Gauthier (1986) has made a major effort to establish a rational-choice link between constitutional and action interests. Gauthier analyzes moral choice in terms of "a choice among dispositions" (1986, 183) rather than as a matter of possible morality-based responses to choice alternatives in specific situations.[5] Gauthier's central argument is that the choice of a general disposition to be moral can be rational even if this implies occasionally missed opportunities to earn larger pay-offs by non-moral behavior. According to Gauthier, a moral disposition can be rational because it allows a person to secure access to cooperative arrangements, to potential gains from cooperation, from which persons without such a disposition would be excluded.[6]

Though we shall develop a similar argument in a later section, there is a noteworthy difference between Gauthier's and our own account of the potential correspondence between constitutional and action interests. Gauthier's aim is to show that such a correspondence is implied by a proper conception of rationality. He seems to deny the possibility of a systematic gap "between rational compliance and rational agreement" (1987, 9) when he claims that "agreement on a set of principles carries with it, in some manner, adherence to those principles" (ibid., 13).[7] In our view, Gauthier's attempt to establish a direct link between the rationality of constitutional agreement and the rationality of compliance is not successful, and it cannot be succesful. Whether or not it is rational for persons to

comply with rules that they constitutionally may agree on is a matter of contingent, factual circumstances and not of rationality per se. It depends on whether or not the constraints that persons face after the agreement - i.e. post-constitutionally - make it rational for them to comply with previously agreed on rules. There is, to be sure, a rational link between constitutional agreement and compliance, but it is of an indirect rather than a direct nature as suggested by Gauthier. If it is rational for persons to agree on rules, it is rational for them to see to it that compliance is rational and, where necessary, to agree on some appropriate enforcement scheme, provided the costs of enforcement are warranted by the prospective cooperative gains.[8]

It has already been stated that our interest here is precisely in identifying the conditions under which compliance on agreeable rules can be rational, and in analyzing the social mechanisms and processes that tend to bring about those conditions. There are two basic ways in which a correspondence between constitutional interests and operational or action interests can be brought about, two ways which are complementary rather than mutually exclusive. The one has been stressed by Thomas Hobbes who, in his theory of social order, essentially argued that people who agree in their constitutional interests can rationally choose to modify the structure under which they act so as to bring about an explicit correspondence of constitutional and compliance interests (by deliberately changing the payoff structure of the generalized prisoners' dilemma matrix). In the Hobbesian conception the correspondence between the two kinds of interests is viewed as a product of individuals' rational capacity to implement their constitutional interests, to diagnose the problems they face and to change the choice-environment so as to make mutually preferred behavior individually rational.

The other solution to the 'correspondence problem' is most often associated with David Hume, Adam Ferguson and Adam Smith as well as other Scottish moral philosophers of the eighteenth century. These philosophers suggested that in some contexts of social interaction spontaneous forces may be present that will bring about a correspondence between constitutional and compliance interests, as if by an 'invisible hand'.[9] This conception focusses on the 'non-intentional' linkage between the two kinds of interests. It argues that the constraints that make it rational to comply with constitutionally preferred rules are, at least to some extent, an unintentional - but systematic - by-product of actions that persons take in pursuit of their immediate interests, without any explicit regard to their constitutional preferences. It is this interpretation that has been at the heart of the 'spontaneous order tradition', a prominent contemporary advocate of which is F.A. Hayek.

As indicated before, the two views on how a correspondence between constitutional and action interests may be generated are by no means mutually exclusive. The two principles, the invisible-hand and the constitutional-constructivist variant, may supplement each other and operate in combination. In the following sections we shall seek to explore the potential range over which 'spontaneous forces' may be expected to generate a tolerable correspondence between the two kinds of interests and to determine the critical limits beyond which deliberate concerted effort seems to be essential. In other words, we shall examine the possible forces that spontaneously, as a by-product of ordinary social interaction, tend to generate compliance with constitutionally preferred moral rules.[10]

IV. Coordination Rules and Prisoner's Dilemma Rules

The spontaneous order tradition contains a certain ambiguity in its analytical approach to the rules and institutions issue, an ambiguity that results from the failure sufficiently to distinguish between two different kinds of interaction problems, namely, in the terminology of modern game theory, coordination problems and prisoner's dilemma problems.[11] There is a tendency throughout this tradition - from David Hume over Carl Menger to F.A. Hayek - to argue as if the kind of explanation that applies to coordination-type rules can be generalized to other kinds of rules as well, including those of the prisoner's dilemma type.

David Hume, for instance, in the context of his discussion on a "theory concerning the origin of property, and consequently of justice" (1975, 307), refers to the example of two men pulling the oars of a boat (1975, 306; 1967, 490), as if the way oarsmen come to coordinate their respective activities could be considered to illustrate the general characteristics of the process by which people come to respect property and to follow the rules of justice.[12] In the same context Hume (ibid.) cites as further "examples" the ways in which "gold and silver are made the measures of exchange" or "speech and words and language are fixed". And it is in the same sense that Hume refers to the example of the rules of the road when he talks about "the necessity of rules, wherever men have any intercourse with each other" (1975, 210).[13] All these examples are, however, concerned with problems of the coordination type rather than with the kind of prisoner's dilemma problems which seem to be typically at the basis of what we use to classify as moral rules.[14] The 'perverse incentives' that characterise prisoner's dilemma problems are absent in coordination problems, allowing rules to emerge and to be maintained much more smoothly in the latter case than in the former.[15] For recurrent coordination problems individuals' constitutional interests and action interests are typically in harmony, at least in the sense that there is little or no incentive for 'defection', once a rule is established. As a coordination rule

emerges - whether it concerns rowing a boat, the use of a general medium of exchange, the use of language, or the rules of the road - there exists, under standard conditions, no 'temptation to defect'. While having a co-ordination rule established in a community can, in some sense, be considered a 'public good', there is definitely no 'free-rider-problem' in the sense that a person may hope to gain extra-benefits by unilaterally defecting. Coordination rules are, in other terms, largely self-enforcing.[16] Rules providing solutions to recurrent prisoner's dilemma type problems are, by contrast, typically not self-enforcing. There is no 'natural' harmony between constitutional and action interests, even if there is perfect agreement on the former among all members in a community. Rather, additional incentives, 'additional' to the payoffs embodied in the problem-defining payoff structure, have to be generated somehow in order to bring constitutional interests and action interests into harmony.[17]

The ambiguity in Hume's discussion of the 'emergence of rules' issue is paralleled in Carl Menger's discussion on the same issue. Menger's explanation on the 'origins of money', commonly cited as the paradigmatic example for an invisible-hand explanation of rules and institutions, is apparently concerned with a coordination type problem and has little direct implications for prisoner's dilemma type rules. In this sense it is misleading indiscriminately to list, as Menger does, phenomena like law, language, the state, money and markets as if all involve the same kind of explanatory problem.[18] The same criticism finally applies to F.A. Hayek (1964, 5) when he talks about the spontaneous emergence of "useful institutions ... such as language, morals, law, writing, or money", implicitly suggesting by such a list that the emergence of rules of morals can basically be explained along the same lines as the emergence of language or money.

In view of the failure of the spontaneous order tradition adequately to account for the fundamental difference between coordination rules, like rules of language or rules of the road, and prisoner's dilemma rules, like rules of morals, it cannot be strongly enough emphasized that an explanation of the first cannot be simply considered a model for an explanation of the second. This does not mean, however, that an invisible-hand explanation of the emergence of prisoner's dilemma type rules is inconceivable. It only means that such an explanation will have to be stated in somewhat different terms; in particular, it will have to specify the forces or mechanisms that curb the ever-present utility-maximizing temptation to defect. Important suggestions for such an explanation have, in fact, been made within the spontaneous order tradition, suggestions that center around the notion of reciprocity as a fundamental principle in human interaction.

V. Reciprocity and Cooperation

The prisoner's dilemma notion and the public goods notion are equivalent conceptual tools to characterize the incentive structure underlying the moral order problem. The public goods interpretation draws attention to the question of what kinds of incentives may induce an individual to contribute to the production of the public good 'moral order', where 'to contribute' is typically seen as a matter of an individual's own behavioral compliance with the relevant rules. This interpretation is not complete, however, because there are two ways in which individuals can contribute to the production of this 'good': By their own rule-compliance and by providing incentives for others to comply. An invisible-hand theory of how prisoner's dilemma type, moral rules come to be effective would have to show how, in the process of social interaction, selective incentives are spontaneously created which induce people to contribute, in the two ways mentioned, to the production of a moral order. The notion of reciprocity is a central one in this context, i.e., the notion that in social settings where individuals repeatedly interact they are in a position mutually to reinforce each other's behavior, to reward 'desirable' and to punish 'undesirable' behavior. That reciprocity works as a spontaneous enforcement mechanism which encourages cooperative behavior has been stressed again and again throughout the history of social theory and across the various social sciences. It was central to the social theory of David Hume and other eighteenth century Scottish moral philosophers, and it is central to the so-called 'exchange theory' in modern sociology.[19]

More recently, in his book The Evolution of Cooperation, R. Axelrod (1984) has added some interesting new aspects to the study of reciprocity. By way of computer experiments Axelrod simulated competition among potential alternative behavioral strategies that actors may adopt in recurrent prisoner's dilemma type interaction situations. The principal result that Axelrod found is that the simple strategy of TIT FOR TAT (the strategy of cooperating in the first move and then doing whatever the opponent did in the previous move) performed better than any of the other strategies that were included in the experiment. The essential reason for TIT FOR TAT's success is its combination of readiness to cooperate on the one side, and preparedness to 'punish' defection on the other. The willingness to cooperate (i.e. to comply with 'moral' rules) allows an actor to realize gains from cooperation in interactions with others who are equally disposed. Being prepared to punish defection protects an actor against continuous exploitation.

TIT FOR TAT obviously reflects the basic pattern of the type of behavior that the concept of reciprocity describes. Though human reciprocating behavior is likely to be much more complex than TIT FOR TAT, Axelrod's results are of obvious relevance for the study of reciprocity as a spontan-

eous enforcement mechanism in everyday social life. Reciprocating behavior has been universally observed across cultures and through time, and Axelrod's study illuminates an obvious reason why reciprocity can be expected to be a universal feature of human social conduct: It is likely to be adopted simply because it tends to be more successful than alternative behavioral strategies.[20] There are basically two forms that such 'adoption' may take, or, stated differently, two mechanisms by which 'success' can be expected to result in the behavioral pattern's diffusion, namely (1) genetic evolution and (2) individual learning. The observed patterns of human reciprocity can probably be best understood as the combined outcome of both mechanisms.

In his 1971 article The Evolution of Reciprocal Altruism, the biologist R.L. Trivers sought to explain reciprocating behavior in evolutionary terms. Trivers' analytical interest is in those kinds of behavioral patterns that produce some apparent benefit to another organism while involving some cost to the organism performing it.[21] As Trivers points out, beyond the relatively narrow limits of close kinship, where "kin selection" may allow for the evolution of genuinely self-sacrificing behavior, natural selection can be expected to favor helping patterns of behavior only where "in the long run they benefit the organism performing them" (1971, 35). This is, however, as Trivers argues, typically the case where such behavior is reciprocated. To the extent that reciprocity allows for mutual net-benefits, natural selection will tend to favor reciprocating behavior. It allows for realization of benefits from 'mutual helping' or cooperation without being vulnerable to systematic exploitation by 'cheating', i.e. non-reciprocating, individuals.

For reasons like those studied by Trivers, the disposition to reciprocate is likely, to some extent, to be genetically 'hard-wired' into human nature.[22] Learning certainly supports the same behavioral tendencies and accounts for some of the extraordinary complexities that characterize human reciprocity.[23] The interaction of genetically inherited and learned traits in human interaction appears to be exemplified by what Trivers (1971, 49) calls "moralistic expression". Since rewarding as well as punishing others are costly to the actor, learning will support such activities only to the extent that they are apt to generate beneficial consequences to the actor himself. On these grounds moralistic aggression may be learned as 'successful' behavior in settings where the initiator and the addressee of the aggression are likely to meet again and where, therefore, the 'shadow of the future' provides a rationale for incurring the costs of the aggression. There are, however, apparent instances of moralistically aggressive behavior that do not seem to fit such a description because the aggressor cannot reasonably expect that the effects on the addressee's future behavior will generate benefits to him that will outweigh the costs of his punishing act. Out of emotions like anger people sometimes tend to

reciprocate or, more descriptively, to retaliate against defectors even in situations where the potential future payoffs from such behavior seem to be in obvious disproportion to the costs incurred.[24]

The seemingly 'irrational' readiness to punish others that sometimes appears to be caused by emotions like anger seems to be difficult to account for in terms of individual learning. An evolutionary explanation might, however, be constructed in terms of potential advantages that such behaviour may generate in the 'very long run'. To be disposed to punish defectors even in cases where rational calculation would suggest not incurring the costs of doing so, may well be beneficial in the longer run by providing better protection from other actor's exploitative inclinations. To be perceived as somebody who is willing to hurt himself only to get the satisfaction from taking revenge may be a most effective deterrent.[25]

VI. Trust Rules and Solidarity Rules

To the extent that the production of 'moral order' involves the same problems as public goods production, in general, rational self-seeking actors cannot be expected to contribute, except if there are selective incentives, i.e., benefits that are contingent on their own contributions. The principle of reciprocity can, in the sense described before, be expected to generate such selective incentives, at least to some extent. In terms of our earlier analysis: Reciprocity can be expected to bring persons' operational or action interests into accordance with their constitutional interests in recurrent prisoner's dilemma type interaction situations. The potential role that reciprocity may play in this respect requires, however, some qualification.

Reciprocity seems likely to emerge and to be effective as a behavioral pattern only in critically small-number settings, where individuals both identify others in the social interaction and expect to experience further dealings within the same group. The question for us becomes one of identifying conditions under which persons are likely to form small-number groups or 'cooperative clusters' that internally secure rule-following through reciprocity. In this regard it is useful to distinguish between two types of rules which we shall call trust rules and solidarity rules.

Trust-rules are rules such as "keeping promises", "telling the truth" or "respecting others' property". Trust-rules have their significance typically in dealings among particular persons. By his compliance with or transgression of trust-rules, a person selectively affects specific other persons. Because compliance with and non-compliance with trust-rules is, in this sense, 'targeted', the possibility of forming cooperative clusters exists: Any subset of actors, down to any two individuals, can realize cooperative

gains by following those rules in their dealings which each other. Adoption of and compliance with trust-rules offers differential benefits to any group or cluster, independently of the behavior of other persons in the more inclusive community or population. Even in an otherwise totally dishonest world any two individuals who start to deal with each other honestly - by keeping promises, respecting property etc. - would fare better than their fellow-men because of the gains from cooperation that they would be able to realize. To be sure, they would be even better off if all their fellowmen could be trusted to act honestly. But, what is crucial in the present context, there are gains from rule-compliance that can be realized within any subset, however small, without any need to achieve inclusive compliance within some predefined group. It is precisely the possibility of forming such cooperative clusters, i.e., the possibility of gains from cooperation to be realized by any subset of actors, that allows the mechanism of reciprocity to be effective in enforcing trust-rules.

Solidarity-rules are rules such as "not littering in public places", "respecting waiting lines", "not driving recklessly", "paying one's fair contribution to joint endeavours", "not shirking one's duties in a team", etc.. In contrast to trust-rules, compliance with or violation of solidarity-rules cannot be selectively targeted at particular other persons, at least not within some 'technically' - i.e., by the nature of the case - defined group. There is always a predefined group, all members of which are affected by their respective rule-related behavior. Whether the relevant group is a work team (as in case of the shirking problem) or the world population (as in case of certain pollution problems), a person cannot avoid by his compliance or non-compliance with the applicable solidarity-rule indiscriminately affecting all members of the predifined group. For solidarity rules it is not true, as it is for trust-rules, that any two individuals can start to form a 'cooperative cluster' that would allow them to realize differential gains from which their unconstrained fellow-men are excluded. Solidarity-rules require adherence by some inclusively defined persons before providing differential mutual benefits to those who adopt compliance behavior.

The very fact that, in the case of solidarity-rules, clustering is not possible, or possible only in a much more restricted sense,[26] makes reciprocity a much less effective mechanism of spontaneous enforcement for those kinds of rules. The crucial difference that separates the two kinds of rules in this regard is reflected in the differences between Axelrod's study, The Evolution of Cooperation (1984) and his more recent study, The Evolution of Norms (1986b). Though this is not an explicit part of Axelrod's own interpretation, the 1984 study can be said to be about the spontaneous emergence of trust-rules, while his 1986 study is an attempt to explain the spontaneous emergence of solidarity-rules (the example that Axelrod uses in this study is the norm "not cheating on exams").

Characteristically, the notion of clustering is central to the first study but plays no role whatsoever in the latter. Instead, the crucial explanatory role is played here by assumptions about 'vengefulness' as an inherent emotional energy that makes people willing to incur some cost in order not only to punish others whom they observe cheating, but also - and this turns out to be the central part in Axelrod's account of the 'evolution of norms' - to punish others for failing to punish observed defections.

VII. Clustering and Compliance

Where individuals repeatedly interact with each other, there are direct personal gains to be made by obeying rules like "keeping promises" and by punishing others for defecting. It is the 'shadow of the future', the expected effects of one's own current behavior on the opponent's future behavior, that is crucial for one's current behavioral choices.[27] In dealing with reciprocating opponents one can not expect to be able to get away with 'cheating', and the only way to secure their ongoing cooperation is by playing by the rules. It is in an individual's direct interest to behave in such a way that he is perceived by others as someone who can be trusted as an honest person. Being trustworthy makes one an attractive partner for cooperation and, thus, increases one's prospects of realizing cooperative gains.[28] On the other hand, the interest in protecting oneself against exploitation provides an immediate incentive for punishing cheaters. The most obvious and least costly form of punishment is simply to exclude a cheater from cooperation until he makes up for his dishonest behavior and proves himself to be a trustworthy person. But an individual may very well have an incentive to take stronger punitive measures, even though they are more costly to himself. Such behavior sends a message to one's direct opponent as well as to third parties. It indicates that one is prepared strictly to retaliate whenever one is being cheated on. In addition, by signaling to other members of the group that the opponent is a cheat, one is able to inflict - at little cost to oneself - an even more effective punishment as others will be more reluctant to deal with the defector in the future.

With trust-rules, the mechanism of reciprocity is capable of activating private interests in following such rules and in punishing others for rule violations. In this sense, 'moral order' can be expected to be generated, at least to some extent, spontaneously, through reciprocity, an observation that might be taken as an example for how a public good may be produced as a by-product of individuals' separate pursuits of purely private interests. It should be noted, though, that, as far as trust-rules are concerned, 'moral order' can be considered a public good in a limited sense only. To be sure, there are certain benefits from living in a community of

honest people that are genuinely public. Consequently, there are apparent opportunities for defectors to 'free-ride' on these benefits in the sense of taking advantage of an environment where people are generally honest and expect others to be generally honest. But the mechanism of reciprocity does not allow for someone systematically to 'free-ride' on other persons' compliance with trust-rules. Defectors will be inevitably excluded from those benefits that can only be realized in ongoing cooperative relations.

For solidarity-rules it is obviously true - as it is for prisoner's dilemma type rules in general - that a rational actor's constitutional interests in such rules do not, per se, generate compliance. Separate, selective incentives are required for bringing operational or action interests into accordance with constitutional interests. It is with regard to the way in which such selective incentives can be expected to be generated that a crucial difference exists between solidarity-rules and trust-rules, a difference that has to do with the extent to which the benefits that result from obeying and enforcing these rules are genuinely public goods, the benefits of which 'spill over' among a large number of non-excludable recipients. The mix of benefits a person generates by obeying and enforcing solidarity-rules, systematically tends to include (other things being equal) more public and less private elements, as compared to trust-rules. In fact, rather than being viewed as a dichotomy, the distinction between trust-rules and solidarity-rules may be more appropriately interpreted in terms of a continuum along which rules may be located according to the degree of 'publicness' of the benefits from rule-obedience.

The incentives for complying with trust-rules and for punishing others who defect derive from the expected effects of one's own actions on other actors' future behavior. These effects include, in the first place, the effects on one's direct opponents' future behavior: The gains one can expect from making them more inclined to cooperate and less inclined to defect in future interactions. In addition, expected indirect effects on third parties may also provide incentives for an individual to comply with rules and to punish cheaters. To be perceived as a trustworthy but also vengeful person increases one's prospects of realizing gains from cooperation and, at the same time, makes one an unsuitable target for exploitation.[29] In any event, it is the expected effects on other persons' future behavior towards the actor himself that provide private inventives for complying with trust-rules and for punishing defectors. The same cannot be said for solidarity-rules. By complying with rules like "not shirking one's duties in a team", a person generates benefits that are public to the relevant group. These benefits cannot be selectively allocated in order to affect the behavior of particular members within the group. On the other hand, by punishing others who defect, a person may make their future compliance more likely. Yet he will share the benefits from such 'improved behavior' with all members of the relevant group, without selective rewards

to himself. In other words, by complying with solidarity-rules and by punishing others for not complying, a person is producing a genuine public good, i.e., benefits that are shared by all members of some predefined group and that, as such, do not qualify as selective incentives.

The above arguments do not imply that there exist no selective incentives at all for individuals to comply with solidarity-rules and to punish defectors. Such incentives may exist, for instance, where a person's behavior toward solidarity-rules affects his reputation. Such behavior may be perceived by others as a signal about what type of person he is. And this again may affect their future behavior toward the person.[30] A person's revealed willingness to comply with solidarity-rules may be interpreted by (direct or indirect) observers as indicative of his general trustworthiness. And a parallel argument may apply to a person's revealed willingness to contribute to the enforcement of solidarity-rules. In addition to such kinds of selective incentives, the emotional factors that have been discussed above under the label 'moralistic aggression' can also be expected to contribute, to some extent, to a spontaneous enforcement of solidarity-rules. In fact, as mentioned before, it is these emotional factors that play a crucial role in Axelrod's model of the 'evolution of norms' (1986).[31]

VIII. From Hobbesian Anarchy to Moral Order

A major implication of our analysis for the 'Hobbesian problem of social order' is that the clustering option that exists for trust-rules makes the leap out of the Hobbesian anarchy somewhat less difficult than the common public goods interpretation of moral order suggests. As far as trust-rules are concerned, individuals do have means, even in large number settings, 'privately' to orchestrate the transition from anarchy to moral order. For the first step towards a normative order to be taken, no more is required than that just two, any two, 'inventive' individuals realize that they can fare better by dealing 'honestly' with each other, by following in their dealings with each other certain rules. Such a two-person cooperative cluster can get the order-creating process started because the differential success of the initial cooperators can be expected to provide incentives for others, either to join the existing cluster or to copy the successful cooperative arrangement. Reciprocity will, at least to some extent, protect existing cooperative clusters against invasion by defectors: Reciprocating actors will allow only those actors to be included in their cooperative network who are willing to submit to the rules.

Since the possibility of discriminating between 'cooperators' and 'defectors' is critical for the stability of cooperative clusters, there are apparent limits to the group-size up to which the principle of reciprocity may serve as a workable mechanism of spontaneous rule enforcement. But a plausible

explanatory account seems possible of a gradual process by which a more extended, 'segmented' moral order may emerge, a moral order that extends beyond the limits of single cooperative clusters that exist as scattered 'islands' within a Hobbesian world. Such a more extended, but still largely 'spontaneous' moral order might be achieved through a kind of 'second order' clustering process. Just as, on the individual level, any two actors can profit from forming a cooperative cluster, on the group level any two groups can realize additional cooperative gains by entering some kind of mutual collective surety arrangement. By collectively accepting responsibility for each group member's rule compliance in his dealings with members of the other group, the intra-group enforcement potential is made effective for creating a normative order in between-groups dealings, thus allowing for mutually profitable transactions to be carried out beyond the limits of the original cooperative clusters. The requirements for such 'second order' cooperative clusters to emerge are equally parsimonious as the requirements for the initial emergence of cooperative clusters: It takes no more than just any two groups that are 'inventive' enough to realize the gains that can be made by such a surety arrangement, and, once a 'model' exists, other groups have an incentive to participate in or to imitate such arrangements.[32]

A more general conclusion concerning the relation between group size and the prospects for a spontaneously created moral order is implied in our analysis. In discussions on this issue a common supposition is that there exists an inverse relation between group-size and the likelihood of a moral order spontaneously to emerge and to be sustained. It is typically argued that - in the absence of deliberately organized enforcement - persons' willingness to contribute to the production of 'moral order' will decrease as group size increases, for the same reasons that are familiar from the general discussion on the significance of group size for the production of public goods:[33] First, the individual will have less and less reason to expect that his own contribution (his own compliance and his punishing of defectors) will be decisive for the persistence of moral order. And, second, the informal, spontaneous mechanisms of enforcement will be less effective in larger and more anonymous groups.

Our analysis of the differences between trust-rules and solidarity-rules suggests that the standard diagnosis concerning the relevance of group size for the moral order issue needs to be qualified. It should be apparent that solidarity-rules and trust-rules are not affected in exactly the same way by growing group size, in particular, that the latter are much less vulnerable to increasing numbers than the former. The formation of co-operative clusters which is possible with regard to trust-rules, makes rules like "keeping promises" much more robust and resistant against the detrimental effect of increasing numbers. So far as trust-rules are concerned, individuals do have means, even in large number settings, to

start building a 'moral order', means that are not available to them in the same way where solidarity-rules are concerned.

To the extent that different kinds of social settings or environments in which persons interact can be meaningfully arrayed along the trust-rules/solidarity-rules distinction, the arguments that we have elaborated here have implications for our understanding of the working principles of these different settings. The fundamental Hayekian distinction between two kinds of social order - between 'spontaneous order' and 'directed social order', or, more specifically, between market order and organization - directly comes to mind in this context.[34] The 'rules of the game' that characterize or define a market-type order are apparently more of the trust-rules than of the solidarity-rules variety, while the opposite is true for organization-type orders - a fact that should have relevant implications for the relative robustness of the respective kinds of order. Markets possess the great advantage, over other types of social arrangements, that they are based on two-party transactions when finally reduced to their basic elements. It is this feature that gives reciprocity its effectiveness as a compliance-inducing device. It is not at all surprising that the 18th century discoverers of the self-enforcing characteristics of market order were exited. The Hobbesian problem of order had been, in large part, resolved. Recognition of the same reciprocity characteristic of market interaction has led David Gauthier (1986, 83ff.) to call the idealized market a "moral-free zone".

All this is not to suggest, of course, that the self-enforcing capacity inherent in markets would make the explicitly constructed arms and agencies of the law dispensable. It seems nonetheless clear, however, that markets remain particularly robust social arrangements for the reasons noted here.

Notes

1 Instances for the 'functionalist fallacy' can notably be found in sociology and anthropology, in particular in the functionalist schools within these disciplines. Economics, because of its dominantly individualistic orientation, has been less susceptible in this respect though it has not been perfectly immune from this type of fallacy, e.g. in some of the analyses concerning the emergence of 'efficient' institutions.

2 A similar distinction is made by Heckathorn 1987 who uses the terms "inclinations" and "regulatory interests" in order to distinguish between the interests ("inclinations") that make rational actors in prisoner's dilemma situations choose the mutually destructive strategy and their ("regulatory") interests in having the choice situation regulated in a way that would allow them to realize the mutually advantageous cooperative outcome.

3 In this sense, many of the common uses of the 'practice what you preach' argument are inappropriate if the inference of inconsistency is made. It is not necessarily inconsistent to advocate a social rule while at the same time behaving differently from the way that might be dictated by generalized adherence to the social rule being advocated.

4 Although, in his writings on the philosophy of law (Metaphysik der Sitten), Kant is well aware of the difference between the two kinds of interests. Cf. Kant 1887, 91ff., 155ff., 163ff. (We owe these references to Hartmut Kliemt).

5 For a discussion of the 'rationality of morality' issue in terms of a choice of dispositions rather than case-by-case choices see also Vanberg 1988.

6 Gauthier 1986, 183: "The essential point in our argument is that one's disposition to choose affects the situations in which one may expect to find oneself." - Harman (1986, 6) identifies the same kind of argument in David Hume's explanation of morals out of self-interest: "Self-interest is involved because, if you cannot be trusted to tell the truth, keep your promises or avoid injuring your associates, people will not join up with you in common enterprises and you will lose out in comparison with other people who do tell the truth, keep their promises, and avoid injury to associates."

7 Gauthier's 1987, 8 sees a shortcoming of John Rawls' contractarian conception in the fact that Rawls shows why it is rational for persons to agree on certain principles, but "does not show, or attempt to show, the rationality of their compliance with the agreed principles".

8 In fact, Gauthier's argument is not always perfectly unambiguous in this respect since, at some places, he apparently presupposes that conditions are de facto given under which compliance can be expected to be in a person's interests. Cf. e.g. Gauthier 1987, 14: "So what we suppose is that I find reason to comply with constraining principles in the benefits that accrue to me through the response of my fellows..."

9 The paradigmatic notion is, of course, that of the spontaneous order of markets. It should be kept in mind, though, that the spontaneous coordination within markets and the enforcement of the legal-institutional framework of markets are different issues. The notion of spontaneous market coordination can very well be combined with a more 'constructivist' view on the institutional framework.

10 See Buchanan 1975, 1977, 1988 for a discussion that puts more emphasis on the constitutional-constructivist perspective.

11 In terms of the typical pay-off structure in a 2x2 matrix the two kinds of interaction problems can be characterised as follows:

Coordination Problem:

		B: b-1	B: b-2
A	a-1	R,R	P,P
	a-2	P,P	R,R

(with R>P)

Prisoner's Dilemma Problem:

		B: b-1	B: b-2
A	a-1	R,R	S,T
	a-2	T,S	P,P

(with T>R>P>S)

12 In the setting that Hume obviously has in mind - namely - the two men pulling one oar at different sides of the boat - the oarsmen are clearly facing a pure coordination problem. There exists, in such a setting, simply no opportunity for 'cheating'. The situation is different, of course, if the oarsmen are pulling two oars each, one sitting behind the other.

13 In the context from which the quotation is taken, Hume draws a comparison between rules for the conduct of ordinary games and the "rules of justice, fidelity, and loyalty" (1975, 210) upon which a society is based. After emphasizing that the comparison is in several ways "very imperfect", Hume states: "We may only learn from it the necessity of rules, wherever men have any intercourse with each other. They cannot even pass each other on the road without rules. Waggoners, coachmen, and postillions have principles, by which they give the way". (ibid.)

14 The relation between moral rules and prisoner's dilemma problems is stressed in Gauthier 1986. On Gauthier's argument cf. Buchanan 1987, 8f.; cf. also Vanberg 1988, 3f.

15 The two kinds of problem situations and the respective kinds of problem-solving rules are discussed in more detail in Vanberg 1986.

16 Their self-enforcing character implies on the other hand that, while coordination rules may be spontaneously established, a spontaneous transition from some established coordination rule (e.g. "driving on the right side of the road") to a different one may be unlikely or even impossible, even though the other rule may be preferable in terms of people's constitutional interests. In this sense, and only in this sense, people's constitutional interests and their action interests may be in 'disharmony' even for coordination rules. But such potential disharmony is, of course, a totally different issue than the typical disjunction between constitutional and compliance interests in case of prisoner's dilemma type moral rules, the issue that we are interested in here.

17 See the different ordinal rankings of pay-offs in footnote 11.

18 Menger 1985, 147: "Law, language, the state, money, markets, all these social structures in their various empirical forms and in their constant change are to no small extent the unintended result of social development. The prices of goods, interest rates, ground rents, wages, and a thousand other phenomena of social life in general and of economy in particular exhibit the same peculiarity. Also, understanding of them ... must be analogous to the understanding of unintentionally created social institutions. The solution of the most important problems of the theoretical social sciences in general and of theoretical economics in particular is thus closely connected with the question of theoretically understanding the origin and change of 'organically' created social structures."

19 The role of the reciprocity notion in social theory, in particular with regard to the Scottish moral philosophy, to anthropology and to exchange-sociology, is discussed in Vanberg 1975, 15ff., 55ff., and Vanberg 1982, 129ff.

20 It has been occasionally argued (cf. e.g. Gouldner 1960) that people's disposition to reciprocate reflects a 'norm of reciprocity' which requires such behavior as 'proper' conduct. Though it is certainly true that normative expectations are often attached to reciprocating behavior (concerning 'gratitude' as well as 'revenge'), the universality of such behavior strongly indicates that those normative expectations are a secondary, not primary, phenomenon, that they are a consequence rather than the cause of the general behavioral tendency to reciprocate.

21 Trivers 1971, 35 labels such behavior "altruistic": "Altruistic behavior can be defined as behavior that benefits another organism, not closely related, while being apparently detrimental to the organism performing the behavior, benefit and detriment being defined in terms of inclusive fitness." - The terms "altruistic" and "altruism" are probably not the best to describe the behavior under investigation, since these terms tend to presuppose certain assumptions about the 'underlying motivation'. It would seem to be preferable to use a term that is purely descriptive of the behavior that is to be explained, and that is neutral about how it is to be explained. Hirshleifer 1978, 240 suggests to use the term "helping behavior", a term that does not presuppose what the "determinants of helping behavior" are. As Hirshleifer (ibid.) points out: "The patterns of helping are grouped by biologists into three categories: those associated with kinship; those merely incidental to selfish behavior; and those involved in reciprocal interaction."

22 Rawls 1971, 494f. refers to such a "hard-wired" tendency to reciprocate as a crucial ingredient to the "capacity for a sense of justice": "The basic idea is one of reciprocity, a tendency to answer in kind. Now this tendency is a deep psychological fact. Without it our nature would be very different and fruitful social cooperation fragile if not impossible. ... Beings with a different psychology either have never existed or must soon have disappeared in the course of evolution. A capacity for a sense of justice built up by response in kind would appear to be a condition of human sociability."

23 Trivers 1971, 46 refers to some of those complexities when he argues: "because human altruism may span huge periods of time, a lifetime even, and because thousands of exchanges may take place, involving many different 'goods' and with many different cost benefit ratios, the problem of computing the relevant totals, detecting imbalances, and deciding whether they are due to chance or to small-scale cheating is an extremely difficult one." - For an interpretation of reciprocity as learned behavior cf. e.g. Homans 1974, 51ff.

24 An analogous argument applies to 'positive' emotions like gratitude which sometimes seem to make people reciprocate the helping behavior of others in situations where there is little prospect of future benefits which outweigh the cost of the 'act of gratitude'. - Cf. in this context Witt 1986.

25 On the role of emotional behavior for the enforcement of rules in social communities cf. Mackie 1985.

26 It is certainly possible for subgroups within a given community to realize differential gains from internally practicing solidarity rules. For instance, a work team whose members refrain from shirking will be more productive than teams whose members are 'morally unconstrained'. The crucial point, however, is that with solidarity-rules not any subset can realize differential gains from rule compliance. There always exists a technically (by the nature of the collective good that rule-compliance produces) defined group for which some inclusive rule-compliance has to be secured.

27 Axelrod 1984, 12: "What makes it possible for cooperation to emerge is the fact that the players might meet again. This possibility means that the choices made today not only determine the outcome of this move, but can also influence the later choice of the players. The future can therefore cast a shadow back upon the present and thereby affect the current strategic situation."

28 The point that is of interest here has been nicely articulated by Sumner 1918, 95: "Some say that a man cannot afford to be honest unless everybody is honest. The truth is that, if there was one honest man among a lot of cheats, his character and reputation would reach their maximum value. ... If a man ... does right, the rewards of doing right are obtained. They are not as great as could be obtained if all did right, but they are greater than those enjoy who still do wrong."

29 It should be added, though, that being perceived as a vengeful person may involve a certain trade-off: While providing protection from exploitation, it may also decrease, to some extent, one's attractiveness as a potential partner for cooperation because others may worry about the risk of being vengefully prosecuted if they should ever inadvertently defect or if they should be mistaken for defectors.

30 Axelrod 1986a, 1107: "An important, and often dominant, reason to respect a norm is that violating it would provide a signal about the type of person you are ... This is an example of the signaling principle: a violation of a norm is not only a bit of behavior that has a payoff for the defector and for others, it is also a signal that contains

information about the future behavior of the defector in a wide variety of situations."

31 As indicated above, Trivers 1971 arguments on "moralistic aggression" imply that a certain disposition to retaliate is likely to be selected for, a disposition that is relatively (though, of course, not totally) independent of the prospective costs and benefits involved in particular retaliatory acts.

32 Evidence for 'real world' examples of such collective responsibility or surety arrangements are in fact provided by legal historians and anthropologists. On the Anglo-Saxon frankpledge system cf. Morris 1910 and the references in Liggion 1977, 273f. - Anthropological evidence is reported in Moore 1978, chpt. 3: "Legal liability and evolutionary interpretation: some aspects of strict liability, self-help, and collective responsibility." - Heckathorn's 1987 discussion on the role of "collective sanctions" is also of interest in this context.

33 The classical contribution on the general significance of group size for the provision of public goods is, of course, Olson 1965. In Buchanan 1965 the issue is discussed with specific regard to the problem of moral order. For a more recent discussion cf. e.g. Taylor 1982 and Raub 1986.

34 See Hayek 1964 and 1973, 35. See also Vanberg 1982, esp. 88ff.

Bibliography

Axelrod, R. (1984), The Evolution of Cooperation, New York

- (1986a), An Evolutionary Approach to Norms, in: American Political Science Review, 80, 1095-1111

- (1986b), The Evolution of Norms, New York

Brennan, G./J.M. Buchanan (1985), The Reason of Rules - Constitutional Political Economy, Cambridge

Buchanan, J.M. (1965), Ethical Rules, Expected Values, and Large Numbers, in: Ethics, 76, 1-13 (reprinted in J.M. Buchanan 1977, 151-168)

- (1975), The Limits of Liberty: Between Anarchy and Leviathan, Chicago

- (1977), Freedom in Constitutional Contract, College Station/Texas-London

- (1987), The Gauthier Enterprise, mimeographed

- (1988), Economics - Between Predictive Science and Moral Philosophy, College Station/Texas

Gauthier, D. (1986), Morals by Agreement, Oxford

Gauthier, D. (1987), Morality, Rational Choice, and Semantic Representation: A Reply to my Critics, mimeographed, University of Pittsburgh

Gouldner, A.W. (1960), The Norm of Reciprocity, in: American Sociological Review 25, 161-178

Harman, G. (1986), Moral Agent and Impartial Spectator, The Lindley Lecture, University of Kansas, Dept. of Philosophy

Hayek, F.A. (1964), Kinds of Order in Society, in: New Individualist Review 3.2, 3-12

- (1973), Law, Legislation and Liberty, Vol. 1. Rules and Order, London

Heckathorn, D. (1987), Collective Incentives and the Creation of Prisoner's Dilemma Norms, mimeographed, University of Missouri at K.C.

Hirshleifer, J. (1978), Competition, Cooperation, and Conflict, in: Economics and Proceedings 68, 238-243

Homans, G.C. (1974), Social Behavior - Its Elementary Forms, New York

Hume, D. (1967), A Treatise of Human Nature, Oxford

- (1975), Enquiries Concerning Human Understanding and Concerning the Principles of Morals, Oxford

Kant, I. (1887), The Philosophy of Law, Edinburgh

Liggion, L.P. (1977), The Transportation of Criminals: A Brief Political-Economic History, in: R.E. Barnett/J. Hagel III (eds.), Assessing the Criminal - Restitution, Retribution, and the Legal Process, Cambridge/Mass.

Mackie, J.L. (1985), Morality and the Retributive Emotions, in: J.L. Mackie, Persons and Values, Oxford, 206-219

Menger, C. (1985), Investigations into the Method of the Social Sciences with Special Reference to Economics, New York-London

Moore, S.F. (1978), Law as a Process. An Anthropological Approach, London

Morris, W.A. (1910), The Frankpledge System, New York

Olson, M. (1965), The Logic of Collective Action, Cambridge

Raub, W. (1986), Problematic Social Situations and the "Large-Number Dilemma" - A Game-Theoretical Analysis, mimeographed, University of Nuernberg (to appear in Journal of Mathematical Sociology)

Rawls, J. (1971), A Theory of Justice, Cambridge/Mass.

Sumner, W.G. (1918), The Forgotten Men and Other Essays, ed. by A.G. Keller, New Haven

Taylor, M. (1982), Community, Anarchy and Liberty, Cambridge

Trivers, R.L. (1971), The Evolution of Reciprocal Altruism, in: The Quarterly Review of Biology 46, 35-57

Vanberg, V. (1975), Die zwei Soziologien - Individualismus und Kollektivismus in der Sozialtheorie, Tübingen

- (1982), Markt und Organisation, Tübingen

- (1986), Spontaneous Market Order and Social Rules - A Critical Examination of F.A. Hayek's Theory of Cultural Evolution, in: Economics and Philosophy 2, 75-100

- (1988), Morality and Economics - De Moribus Est Disputandum, Original Papers Series, Social Philosophy and Policy Center, New Brunswick

Witt, U. (1986), Evolution and Stability of Cooperation without Enforceable Contracts, in: Kyklos 39, 245-266