Rationality, Rule-Following and Emotions:
On the Economics of Moral Preferences

Abstract: The long-standing critique of the 'economic model of man' has gained new impetus not least due to the broadening research in behavioral and experimental economics. Many of the critics have focused on the apparent difficulty of traditional rational choice theory to account for the role of moral or ethical concerns in human conduct, and a number of authors have suggested modifications in the standard model in response to such critique. This paper takes issue with a quite commonly adopted 'revisionist' strategy, namely seeking to account for moral concerns by including them as additional preferences in an agent's utility function. It is argued that this strategy ignores the critical difference between *preferences over outcomes* and *preferences over actions*, and that it fails to recognize that 'moral preferences' belong into the second category. Preferences over actions, however, cannot be consistently accounted for within a theoretical framework that focuses on the rationality of single actions. They require a shift of perspective, from a theory of rational choice to a theory of rule-following behavior.

1. The Economic Model of Man: Rationality and Self-Interest

The model of man that has dominated the neoclassical tradition in economics comprises two separable core assumptions, rationality and self-interest. Agents are assumed to act rationally in pursuit of what they wish to achieve. And what they wish to achieve is assumed to be defined in terms of their own well-being.[1] Technically this notion of rational, self-interested behavior has been specified as the assumption that agents maximize a utility function subject to the constraints they face. In this construction the rationality component of the economic model of man is specified in terms of the maximization assumption, and the self-interest component is specified in terms of the entries that are included in the utility function. In fact, the agents that populate the standard economic models are essentially 'reduced' to utility functions (Witt 2005: 4ff.).

[1] Sen (2002a: 22f.): "It is the self-interest view of rationality that has been effectively dominant in contemporary economics…(T)he narrow view of rationality simply as intelligent pursuit of self-interest, and the corresponding characterization of the so-called 'economic man,' have been very influential in shaping a dominant school of thought in modern economics…Not only is this assumption widely used in economics, but many of the central theorems of modern economics (e.g., the Arrow-Debreu theorem…) significantly depend on it."

Once the agents' utility functions are specified, the analyzing economist need not know more about them in order to predict what they will choose given the choice options and the constraints they face.[2]

Both components of the economists' standard model, rationality and self-interest, have long since been the target of criticism, from non-mainstream approaches within the field, and even more so from other social sciences (Vanberg 2004). In recent times such critique has gained new impetus, not least due to research findings in behavioral and experimental economics,[3] and there is a broadening discussion on whether and, if so, how the economic model of man might be modified in order to account for behavioral observations that appear to conflict with its traditional interpretation. What is quite obvious from this discussion is that economists are much more conciliatory with regard to the self-interest component of their traditional model than with the rationality component. As far as revisions of the model are suggested, they are typically about modifying the *content* of the utility function, while maintaining the assumption that agents maximize their utility function, whatever its content may be.[4] This 'revisionist' strategy is programmatically stated e.g. when Gary S. Becker (1996:4) notes about the purpose of his *Accounting for Tastes*: "This book retains the assumption that individuals behave so as to maximize utility while extending the definition of individual preferences to include … love and sympathy, and other neglected behavior." The same spirit is reflected in "a remarkably large literature on skillfully 'elongating' the self-interest model" (Sen 2002a: 24) that seeks to deal with "the dissonance between the theory and

[2] Walras (1054: 256): "In our theory each trader may be assumed to determine his own utility or want curves as he pleases. Once these curves have been determined, we show how prices result from them under a hypothetical régime of absolutely free competition." – In reference to V. Pareto's article "Mathematical Economics" (*International Economic Papers*, Nr. 5, 1955:. 61) Georgescu-Roegen (1971: 343) notes: "As Pareto overtly claimed, once we have determined the means at the disposal of the individual and obtained a 'photograph of his tastes … the individual may disappear." As Georgescu-Roegen (ibid.) comments: "The individual is thus reduced to a mere subscript of the ophelimity function Φ_i (x)."

[3] For references to research findings that "contradict the neoclassical model of rational choice" see e.g. McFadden (2005: 12ff.).

[4] Sen (2002a: 24): "A definition of rational choice theory that reflects the 'revisionist spirit' is given by C.Biechieri (2004: 183): 'The theory of rational choice's central assumption is that a decision maker chooses the best action available according to her *preferences.* The content of preferences is unrestricted. Agent's preferences may be selfish or altruistic, self-defeating or even masochistic. Preferences mirror values and dispositions that are beyond the pale of rationality. What is required is that preferences are well behaved in the sense of fulfilling certain formal conditions… If preferences are well behaved, they can be represented by utility functions, and rationality consists in maximizing one's utility function, or finding the maximum value of one's utility function.' "

the actuality of behavior" (ibid.) that has been observed in numerous experiments and real world settings.[5]

The focus of the present paper is on attempts to account in such manner for the role of moral or ethical concerns in human conduct. In the literature on empirical evidence for 'behavioral anomalies' references are quite often made, for instance, to the fact that "standards of fairness" (Kahnemann, Knetch and Thaler 1987: 114) appear to influence agents' behavior, and that including "a preference for fairness in the objective function" (ibid.: 115) can help to resolve the recognized anomalies. Such references to the role of concerns for fairness, equity or justice are particularly prominent in the literature on ultimatum game experiments, presumably the most widely applied and discussed experiments in behavioral economics.[6]

The fact that in the ultimatum game experiments "proposers offer an average of 40 percent of the money (many offer half) and responders reject small offers of 20 percent or so half the time… (falsifies V.V.) the assumption that players maximize their own payoffs as clearly as experiment data can," (Camerer 2003: 43). Ernst Fehr and coauthors, in particular, have argued in a number of articles (Fehr and Schmidt 1999; 2003; Fehr and Falk 2003; Fehr and Fischbacher 2000) that the empirical and experimental evidence for deviations from the predictions of rational choice theory can be accounted for if one relaxes the assumption of self-interest, allowing for other-regarding concerns to be included in individuals' utility functions, while maintaining the assumption that agents are fully rational maximizers given their utility functions. Observations such as proposers' willingness to share in ultimatum games and responders' unwillingness to accept small offers can, so Fehr and coauthors argue, be explained by assuming that a person's utility function may include "social preferences, in particular, preferences for reciprocal fairness" that make her care "not only about the material

[5] Kliemt (2005: 207): "To accommodate the findings they argued that the utility function would not be dominated by material, in particular monetary payoffs but rather by more complex motivations. All that matters is that behavior can be *described as if* individuals would *maximize* some utility function or other representing their given preferences *whatever the latter may be.*"

[6] For the original experiment see W. Güth, R. Schmittberger and B. Schwarze 1982. The experiment involves two subjects one of which, the 'proposer' is provided by the experimenter with a sum of money that he can divide between himself and the second subject, the 'responder.' If the responder accepts the share assigned to him by the proposer both get the respective amounts. If the responder rejects, none of them gets anything.

resources allocated to her but also … about the material resources allocated to relevant reference agents" (Fehr and Fischbacher 2000: C1f.). According to Fehr and coauthors the explanatory power of the rational-choice paradigm can be restored in the face of observed "deviations from purely self-interested behavior" (Fehr and Falk 2003: 40) if one allows agents to be concerned not only for their own well-being but also to be moved by an aversion against inequality and concerns for "reciprocal fairness" (ibid.).

The claim made by Fehr and coauthors regarding the role of inequality and fairness concerns have been subject to critical scrutiny. It has been questioned, for instance, whether it is in fact fairness concerns that motivate proposers' 'generosity' in ultimatum game experiments, or if it is not, instead, their anticipation of the responders' rejection of small offers that motivates their behavior.[7] And alternative explanations have been suggested that operate on more parsimonious assumptions, such as that agents are not concerned with equity or fairness per se but with their own relative standing.[8]

However experimental economists may settle this dispute among themselves is of secondary importance to the issue that is of principal interest in the present paper, namely whether moral preferences can be consistently accounted for as entries in individuals' utility functions, if and to the extent that they are acknowledged to play a role in human conduct. There is, to be sure, no reason why one should not speak in a general sense of a preference for fairness, equity, justice, and so forth, if this is meant to imply that people

[7] Elster (1998: 68f.): "In early studies of the Ultimatum Game it was often argued that the players deviate from self-interest because they are motivated by fairness or a sense of justice. Later experiments have largely ruled out this explanation. In the Dictator Game, where the second player has no choice at all, the first player is usually less generous (Roth 1995: 270). Rather what explains the generosity of the first player is his anticipation that the second player will prefer to take nothing rather than a small amount. – As Bolten and Ockenfels' reference to "responder concerns for equity" (Bolten and Ockenfels [2000: 169]: "Proposers may care about equity (they *do* give money in the dictator game), but it appears that it is responder concern for equity that drives the ultimatum game.") indicates, the fact that proposers act in anticipation of responders' unwillingness to accept small offers implies, of course, that proposers do not expect their counterparts to behave as rational choice theory would predict them to behave, namely to prefer a positive payout over a zero payout. That is to say, the subjects in ultimatum game experiments clearly do not act on the theory that people behave as the standard economic model of man presumes they do. And it is prudent for them not to take their lead from that model, because if they did it would work out to their disadvantage. As Hartmut Kliemt (2005: 211) notes: "The proposer in the ultimatum game who assumes full rationality will normally pay a high price by not earning any money."

[8] Bolten and Ockenfels assume agents to maximize a "motivation function" that includes their own payoff and their "relative share of the payoff" (2000: 171). As they reason (ibid.: 189): "[S]everal studies find that people are willing to sacrifice little to defend egalitarianism. The same experiments cast doubt on the notion that people care about payoff distribution in a way we would expect a purely unselfish altruist to care. People appear self-centered, albeit in a way that differs from received theory."

do not only care about the payoffs that they reap from their own choices and the choices of others with whom they interact, but also care about whether or not their own behavior and that of others is in accordance with generally accepted standards of 'fair,' 'just,' or 'ethical' conduct. The issue that is of interest in the present context is whether such 'moral' preferences can be treated en par with 'ordinary' preferences for pecuniary payoffs, consumable goods, and other objects of desire. The critical line that, I submit, must be drawn here is hinted at in K. Arrow's (1996: xiii) succinct statement: "Choice is over sets of actions, but preference orderings are over consequences."

Rational choice theory looks at actions in a strictly instrumental fashion. Actions are seen as the means or instruments by which agents seek to bring about desired consequences or outcomes. Accordingly, an agent's choice among actions is explained in terms of his preferences over the outcomes that he predicts to result from the alternatives considered. If action *A* is expected to result in outcomes that are more desirable for the agent than the expected consequences of potential alternative actions, rational choice theory predicts the agent to choose *A*. There is no place in this theoretical framework for preferences over actions *per se* in addition to and separate from preferences over the outcomes that they are expected to produce. It is this very fact that, as I shall argue, renders inconsistent attempts to account, within a rational choice framework, for moral preferences simply by modifying the content of the utility functions.

Moral principles, standards of fairness, justice, etc. are typically about *actions*, not about *outcomes.*[9] They are *codes of conduct* that require persons to act in fair, just, or ethical ways. They tell them not to steal, not to lie, to keep promises, etc. They are typically concerned not so much with *what* a person wants to achieve but with *how* she seeks to achieve what she wants. If the notion of a 'moral preference' is to make any sense, this is my principal claim, it can only mean a *preference for acting morally*, i.e. in accordance with moral rules of conduct. In other words, moral preferences are, if anything, *preferences over actions as such,* not preferences over outcomes.

The fact that in addition to, and different from, preferences for outcomes preferences for actions as such may play a role in human decision making has, of course,

[9] This is not to say that moral principles are not concerned with outcomes. Yet, as they are typically stated they are not about outcomes *per se* but about the ways in which one is supposed to go about achieving them.

not been entirely ignored by economists. Bruno S. Frey, for instance, has in a number of contributions drawn attention to this issue by emphasizing the role of "intrinsic motivation"[10] and "procedural utility."[11] What to my knowledge has found little attention, however, is the fact that accounting for preferences over actions requires a shift of focus from a theory of rational choice to a theory of rule-following behavior.

2. Preferences Over Actions and Rule-following Behavior

The reason why experimental economists who invoke 'moral preferences' in their behavioral explanations fail to recognize the implicit shift of perspective from rational choice to rule-following can be found, as I suppose, in their tendency to gloss over the difference between *other-regarding preferences* and *moral preferences.* There is, however, a significant difference between, on the one hand, claiming that agents evaluate outcomes not only in terms of their own narrowly defined interests but also in terms of how they affect the wellbeing of other persons and, on the other hand, claiming that agents are motivated to act in accordance with ethical rules or principles of fairness. It is one thing to claim that individuals' "subjective evaluations of payoffs differ from economic payoffs" (Fehr and Fischbacher 2003: 788) and that, therefore, "non-selfish motives" should be accounted for in specifying their utility functions. It is something quite different to claim that agents have a "predisposition to reward others for cooperative, norm-abiding behaviors, and …a propensity to impose sanctions on others for norm violations" (ibid.: 785).[12] And one must surely distinguish between the claim "that we can treat altruistic preferences in a manner perfectly parallel to the way we treat

[10] B.S. Frey and F. Oberholzer-Gee (1997: 746): "Human behavior is influenced by both extrinsic and intrinsic motivation. The former is activated from the outside. In particular, individuals follow the generalized law of demand. Intrinsic motivations, on the other hand, relate to activities one simply undertakes because one likes to do them or because the individual derives some satisfaction from doing his or her duty."

[11] B.S. Frey, Benz and Stutzer (2004: 377): "The economic concept of utility as generally applied today is outcome-oriented…" (Ibid.: 379): "Procedural utility, in contrast, means that there is something beyond instrumental outputs as they are captured in a traditional economic utility function. People may have preferences about *how* instrumental outcomes are generated. These preferences about processes generate procedural utility."

[12] Also Fehr and Fischbacher (2003: 786).

money and private goods in individual preference functions" (Gintis and Khurana 2006:11), and the claim that "character virtues" such as honesty and fairness can be included as "argument(s) in one's preference function, to be traded off against other valuable objects of desire and personal goals" (ibid.: 18).[13]

A rational choice approach that represents human agents by utility functions, and that seeks to explain human behavior as the maximization of such utility functions, has its inherent focus on single acts of choice and accounts for these acts of choice exclusively in terms of the consequences that potential alternative courses of actions are predicted to bring about in the particular instance.[14] Such a rational choice approach looks at every single act of choice separately and interprets each action in a purely instrumental fashion, as a means to bring about desired consequences. In each instance, a rational agent is predicted to choose from among the choice options available the action that, in the particular instance, is predicted to result in the most preferred consequences. To be sure, a rational choice theory, so defined, may allow for 'altruistic' or 'other-regarding' preferences, as long as these preferences are interpreted as preferences over outcomes. The basic logic of a rational maximization account is in no way compromised if agents are assumed to judge the 'utility' of the predicted consequences of actions in terms of how they affect not only their own immediate wellbeing but also the wellbeing of others. Whether this is the case or not is an empirical matter. In its purely instrumental outlook at actions a rational maximization account can, however, not allow for actions to be chosen

[13] Gintis and Khurana (2006: 17) recognize that one needs to make a distinction here when they note: "*Character virtues* are ethically desirable behavioral regularities that individuals value for their own sake, while having the property of facilitating cooperation and enhancing social efficiency. The character virtues include *honesty, trustworthiness, promise-keeping,* and *fairness.* Unlike such other-regarding preferences as strong reciprocity and empathy, these character virtues operate without concern for the individuals with whom one interacts. An individual is honest in his transactions because this is a desired state of being, not because he has any particular regard for those with whom he transacts." Yet, they nevertheless insist (ibid.: 78): "One might be tempted to model honesty and other character virtues as self-constituted constraints on one's set of available actions in a game, but a more fruitful approach is to include the state of being virtuous in a certain way as an argument in one's preference function, to be traded off against other valuable objects of desire and personal goals. In this respect, the character virtues are in the same category as ethical and religious preferences, and are often considered subcategories of the latter."

[14] As H. Kliemt (2005: 205) has noted about the 'rational agent' in standard economic theory: He "is acting opportunistically rational in view of the future causal consequences of his choice making for the pursuit of his own advantage. Regularities in his behavior emerge if and only if he faces the same incentives that appeal to his self-interest repeatedly in the same way. But he is always taking each situation separately on its own merits…Once the future looks different from what it looked in the past [he] instantaneously shifts his behavioral gears if this is to his advantage."

in terms of criteria that are different from, and independent of, the agent's preferences over outcomes, such as preferences over actions per se. Such criteria or preferences over actions are, however, inevitably – if only implicitly – invoked when "character virtues" or "predispositions" are argued to guide human behavior, since the very point of "character virtues" and "predispositions" is that agents do not act in response to the payoffs that alternative courses of action are predicted to produce in the particular situation but according to *preconceived* notions or criteria of what kinds of actions are ethically required or appropriate in the kinds of situations they are facing.[15] To act on such preconceived criteria is equivalent, though, to rule-following behavior, since behavioral rules may be stated as "if-then" instructions, where the "if"-component identifies types of situations and the "then"-component specifies the kinds of actions that are called for (Vanberg 2002a: 16). Accordingly, as the terms are used here, to say that a person's behavior is guided by her preferences over actions per se is equivalent to saying that she acts in a rule-following manner.

As agents adopt dispositions to follow rules of action they will presumably experience *emotional consequences* from complying with or going against their behavioral inclinations.[16] They may, for instance, feel uneasy if they 'deviate' from rules they are disposed to act on. Since these emotional consequences may appear to be like other consequences agents consider in their choice of actions, one might be inclined to conclude that behavioral dispositions can, after all, be accounted for by rational choice analysis, as components in agents' utility functions. Such conclusion would disregard, however, the essential fact that the very point of being disposed to follow rules is to act in certain ways in certain types of situation *without* considering the expected consequences in each instance. To be sure, agents may on occasion deliberately act against their rule-following inclinations, giving less weight to the 'bad conscience' from rule-violation than to the benefits it promises. And there are surely cases of calculated rule-compliance where agents consider the benefits to be had from rule-violation insufficient to compensate for the uneasiness felt from acting against their dispositions. Yet these cases

[15] The criteria implied in behavioral dispositions are 'preconceived' or 'categorical' relative to the particular choice situation that the agent faces. In terms of the agent's overall learning history they are a product of previously experienced consequences of alternative ways of acting.
[16] To the role of emotions in rule-following behavior I shall return below (section 5).

are the very instances in which agents shift from a rule-following mode to situational, case-by-case choice, even if their situational calculus includes the emotional implications of their behavioral dispositions. They do definitely not represent the 'standard' cases of rule-following, i.e. the cases in which behavioral dispositions induce agents to act on preconceived notions of appropriate behavior without calculating the expected payoffs from potential alternative courses of action. It is these cases, however, that do not fit the rational choice model.

3. A. Sen on 'Sympathy' and 'Commitment'

In a number of contributions A. Sen has addressed the very issue that is at stake here, and it is instructive to take a closer look at his arguments. In reference to suggestions for how the rational choice model may be revised in order to account for observed behavior that appears to contradict the assumption of rational self-interest Sen argues that a distinction must be drawn between accounting for *sympathy* and accounting for *commitment*. According to Sen, sympathy can without difficulty be accounted for within a rational choice framework, simply by broadening the concept of self-interest. "Indeed," he argues, "being self-interested does not require one to be self-centered in any way, since one can get joys and pains from sympathy to others, and these joys and pains are quintessential one's own" (Sen 2002a: 31). Not only can concern for others be easily accommodated "within the utility function of the persons involved" (ibid.), concerns for any kind of 'goal' or 'value' that a person may be supposed to pursue can, as Sen argues, be accounted for in a rational choice framework, if 'rational choice' is defined in the minimal sense of maximizing an identifiable maximand.

This is categorically different, though, so Sen insists, with commitment.[17] While our everyday experience as well as many empirical studies "indicate that committed behavior has its actual domain" (Sen 2002a:9), it cannot be accounted for by standard

[17] Sen (2002c: 214): "Sympathy – including antipathy when it is negative – refers to one person's welfare being affected by the position of others…, whereas 'commitment' is concerned with breaking the tight link between individual welfare (with or without sympathy) and the choice of action."

rational choice theory, even in its minimal version.[18] Sen's own suggestion for how committed behavior can be accounted for is that we must even relax the assumption of "self-goal choice", i.e. the assumption that a person's choices reflect her own goals, and allow for the pursuit of private goals to "be compromised by the consideration of the goals of others" (Sen 2002c: 215). Commentators like Philip Pettit have criticized Sen's suggestion as highly implausible."[19] And, indeed, it is difficult to see in what sense human choice can be anything other than – in Sen's terminology –"self-goal choice."[20] Yet, the difficulties inherent in Sen's concept of choices other than "self-goal choice" can be easily avoided if one restates his argument on the nature of "committed behavior" in terms of the theoretical perspective that I seek to advance in this paper, i.e. in terms of the distinction between preferences over outcomes and preferences over actions as such, and that draws attention to the intimate link between preferences over actions and rule-following behavior. Such 'restatement' is in fact invited by Sen (2002c: 214) himself when he notes that "the violation of self-goal choice" involved in commitment may "arise from self-imposed restrictions on the pursuit of one's own goals (in favor of, say, following particular rules of conduct)."[21] Apparently it is, in particular, commitment to rules of behavior that, in Sen's view, poses a "more fundamental" problem to standard rational choice accounts than accommodating other-regarding preferences or non-self welfare goals or values (Sen 1973: 249ff.). Accepting "certain rules of conduct as part of obligatory behavior" is, so Sen (2002c: 216f.) argues, "not a matter of asking each time,

[18] Sen (2005a: 8): "A reason for the importance of taking note of commitment is that it can help to explain many patterns of behavior that we actually observe which are hard to fit into the narrow format of contemporary rational choice theory."

[19] As Pettit (2005: 19) charges, Sen's claim that "people may become the executors of a goal-system that outruns the private goals that they endorse in their own name… is highly implausible, at least on the face of it."

[20] Pettit (2005: 19): "According to the minimal version of rational choice theory, people can be represented in action as maximizing an identifiable maximand, or as acting on their own goals: satisfying the assumption, as Sen calls it, of 'self-goal choice.' … Rational choice theory in the minimal sense is close to common sense. … The claim that we can be executors of a goal system that outruns our own goals is bound to raise a question."

[21] See also Sen (2002a: 7): "[A] person's choice behavior may be constrained or influenced by the goals of others, or by rules of conduct…, thereby violating the self-goal choice." – Sen (2002c: 219f.): "[A] rejection of self-goal choice reflects a type of commitment that is not able to be captured by the broadening of the goals to be pursued. It calls for behavior norms that depart from the pursuit of goals in certain systematic ways … and it has close links with the case for rule-based conduct, discussed by Adam Smith."

What do I get out of it? How are my own goals furthered in this way?, but of taking for granted the case for certain patterns of behavior towards others."[22]

As Sen conjectures, it is the very fact that real world human beings act as rule-followers and not as goal-maximizers of standard rational choice theory that allows them to realize many of the mutual gains from cooperation that appear unobtainable for strategically acting rational maximizers. Situations of strategic interdependence as paradigmatically described in the prisoners' dilemma game are, so Sen (1973: 250) argues, "precisely the type of situation in which moral rules of behavior have traditionally played an important part. Situations of the type of the prisoners' dilemma occur in many ways in our lives and some of the traditional rules of good behavior take the form of demanding suspension of calculations geared to individual rationality."[23] That rule-following behavior has to do with *preferences over actions as such* as opposed to 'ordinary' preferences over outcomes Sen (2002b: 191f.) explicitly recognizes when he notes: "This issue is close to Adam Smith's general point that many behavioral regularities can be explained better by understanding people's attitude to *actions*, rather than there valuation of final outcomes. Similarly Immanuel Kant gave a central position in social ethics to…the 'categorical imperative' … While the focus of Smith's and Kant's reasoning is normative rather than descriptive the two are closely linked in their analysis, since both understood actual behavior to be partly based on norms. There behavioral analysis included seeing the process of actual choice through $K(S)$, and not just through an 'everything considered' grand preference ranking."[24]

---

[22] Sen 2002b: 178: "However, in following rules… the motivating factor need not be any concern about the well-being of others…, but simply following an established rule."

[23] Sen (1973: 251): "Suppose each prisoner in the dilemma acts not on the basis of the rational calculations outlined earlier but proceeds to follow the dictum of not letting the other person down irrespective of the consequences for himself… [T]he choice of non-confession follows *not* from calculations based on this welfare function, but from following a moral code of behavior suspending the rational calculus."

[24] On the notation "$K(S)$" Sen (2002b: 189f.) comments: "The practice of enjoining rules of conduct that go beyond the pursuit of specified goals has a long tradition. As Adam Smith had noted, our behavioral choices often reflect 'general rules' that 'actions' of a particular sort 'are to be avoided'. To represent this formally we can consider a different structure from choosing a maximal element, according to a comprehensive preference ranking … from the given feasible set S (allowed by externally given constraints). Instead, the person may first restrict the choice options further by taking a 'permissible' subset $K(S)$, reflecting *self-imposed* constraints, and then seek the maximal elements $M(K(S),R)$ in $K(S)$. The 'permissibility function' K identifies the permissible subset $K(S)$ of each option set (or menu) S."

4. The Reason of Rules and the 'Rationality' of Moral Preferences

F.A. Hayek has made it a central theme of his work that the limits of our knowledge and reason require us to follow rules rather than deciding each case in a discretionary manner on its own merits. The "whole rationale of the phenomenon of rule-guided action," he submits, is to be found in our "inescapable ignorance of most of the particular circumstances which determine the effects of our actions" (Hayek 1976: 20).[25] In the same spirit R. Heiner has worked out a careful argument for why 'imperfect' agents, i.e. agents who are not endowed with the full knowledge and perfect power of reason ascribed to neoclassical rational man, may profit from following rules instead of attempting to maximize on a case-by-case basis.[26] For perfect agents, i.e. agents who are able to determine with perfect reliability what, in particular situations, is the maximizing choice, case-by-case maximization would clearly be the best policy. An imperfect agent, by contrast, may fare better overall by following rules, even though rule-following will inevitably result occasionally in less than optimal outcomes, i.e. in outcomes that are less advantageous than what a perfect agent would choose in the situation. Apparently, the relevant comparison on which the 'rationality' of rule-following for imperfect agents hinges is between, on the one side, the likelihood of 'mistakes' – and the damage resulting from such mistakes – he is bound to make in attempts to maximize case-by-case and, on the other side, the likelihood of – and the damage resulting from – missing out on 'preferred exceptions' when following a rule. The first risk is a function of an agent's competence, i.e. of the quality of the conjectures or theories on which he relies in predicting the consequences which alternative actions will produce, the completeness of his account of relevant situational circumstances, and the reliability with which he can carry out the necessary predictions and compute the associated payoffs.[27] The second risk

[25] The rationalist claim that "man is capable of coordinating his activities successfully through a full explicit evaluation of the consequences of all possible alternatives of action, and in full knowledge of all possible circumstances," represents as Hayek (1967: 90) argues "not only a colossal presumption concerning our intellectual powers, but also a complete misconception of the kind of world in which we live."

[26] The original contribution is Heiner 1983. For a more detailed discussion on the argument developed by Heiner in this and later papers see Vanberg 1983, sect.3.

[27] The perfect rationality assumption ascribes to economic agents unlimited competence. Accordingly for them the ‚error-rate' in case-by-case maximization would be zero, eliminating the reason for rule-following.

is a function of the quality of the rules that guide an agent's actions. It depends on how well these rules are adapted to relevant contingencies that pertain in the environment in which the agent operates, and on how well they focus the agent's attention on easily detectable clues that tell him when in ever new choice situations it is advisable to apply particular rules.[28]

Hayek's and Heiner's arguments draw attention to the role played by factual and conjectural knowledge in human decision making, a role that is essentially ignored by rational choice theories that model human beings as utility functions and claim that how an agent will act in any particular choice situation can be predicted from his utility function. Such theories, it appears, must either presume that all agents, including the analyzing economist, possess the same (perfect) factual and conjectural knowledge, or they must sacrifice the deceptive simplicity of the maximization paradigm by allowing for differences in agents' knowledge and their "mental models", thereby inviting all the explanatory complexities that arise as soon as one recognizes that how people act does not only depend on what they wish to achieve but also on what they know and what they believe.

To be sure, all behavior, rational choice as well as rule-following, must be guided by 'knowledge' about the environment in which agents operate. The demands on the agent's explicit knowledge are, however, critically different in the case of rational choices than in the case of rule-following behavior. Rational choice is about responding to particular, unique situations, considering all that is potentially relevant for the choice among available options. It requires an agent to be able to predict the specific consequences of all the choice options and to calculate the associated payoffs. In complex environments this can obviously be a quite demanding task, in many situations overtaxing the capacities of ordinary humans. Rule-following, by contrast, is about responding to *certain types of situations* by *certain kinds of behavior*. It requires an agent be to able to *classify* the particular situations he confronts as belonging to certain *types* and to identify the *kind of behavior* that according to the adopted rule is appropriate for the given type of situation. The rule itself embodies 'knowledge' of relevant

[28] Rules relieve agents from the burden of having to consider the 'inexhaustible complexity of everything' by singling out selected aspects of the choice situations they face as the only ones to be considered in choosing how to act (Vanberg 1993: 181f.).

contingencies in the agent's environment, 'knowledge' that the agent does not need to actively possess in order to benefit from it. Because of the very fact that agents can benefit from the 'wisdom' implicit in suitable rules, rule-following significantly reduces the demands on their explicit knowledge and cognitive powers compared to rational case-by-case choice.

Recognizing that the 'imperfect' agents that populate the real world may fare better by following rules than by discretionary rational choice requires one to adopt a broader understanding of human rationality than is implied in traditional rational choice theories. By focusing on single acts of choice such theories can only consider actions as 'rational' that in terms of an agent's goal-function and the contingencies of the particular choice situations, are the best means for achieving what the agent seeks to achieve. An action that in this sense, i.e. in terms of a situational account, is not the "best means" simply cannot qualify as a 'rational' action. By shifting the analytical focus from single actions to the level of rules, Hayek and Heiner draw attention to the fact that rule-following is 'rational' – in the sense of serving an agent's interests well – if it results in *patterns of outcomes* preferable to what error-prone discretionary case-by-case choice would produce, even though on many occasions it may call for actions that are not 'rational' in the sense of standard rational choice theory.[29]

As noted, whether or not rule-following promises, in fact, to bring about 'better' patterns of outcomes than discretionary case-by-case choice depends, of course, on the 'quality' of the rules that guide an agent's behavior. This leads one to the question of how agents come to adopt rules and how they come to adopt 'good' rules, i.e. rules that help them to live successful lives. It is obvious, and has often been noted, that agents cannot choose to adopt rules in the same sense in which they can choose among actions. Agents can however acquire *dispositions* to follow rules through processes of behavioral learning, and they may even deliberately take measures in order to enhance the likelihood of learning to acquire dispositions that they whish to possess. Such dispositions can be regarded as *preferences over actions as such* in the sense that they make an agent

[29] An action that is not "rational" in terms of the agent's utility function or his preferences over outcomes may well be induced by "rational" dispositions – or preferences over actions – that enable him to cope more successfully with recurrent problems he is likely to face in the kind of environment in which he operates.

inclined to act in particular ways in certain types of situations, more or less without calculating the costs and benefits that potential alternative courses of action may be predicted to generate in the particular instance.[30]

In particular, whether it is 'rational' for an agent to be guided by 'moral preferences' in the sense of being disposed to follow moral rules of conduct is dependent on the nature of the rules in question and the nature of the environment in which they are applied. Historical records as well as every-day experience provide ample evidence, though, that in human social life conditions are quite commonly established under which it is 'rational' for most – even if not for all – participants to adopt moral preferences.[31] As with any kind of rules, following 'moral rules' is 'rational' in the sense of furthering the agent's interests if by doing so better patterns of outcomes are produced than by following other kinds of rules or by rational case-by-case choice, even if there will be inevitably situations in which the morally disposed person misses out on opportunities that a morally unconstrained rational maximizer might capture. Yet, lacking the ability to reliably identify such 'preferred exceptions' imperfect agents are better off following the rule.[32]

Not different from rational choice theory, a theory of 'rational' rule-following takes a 'consequentialist' outlook at human action in the sense that it, too, assumes human behavior to be governed ultimately by its payoffs. The difference is that rational choice theory assumes *single actions* to be chosen in light of the expected consequences of the particular actions while a theory of rational rule-following assumes that the *dispositions* that guide an agent's conduct are shaped by the payoff-experiences that the

[30] The disposition to follow rules need not necessarily imply the total absence of situational calculation but may well coexist with a conscious deliberation of behavioral alternatives (more on this in section 4 below).
[31] Based on a simple computer simulation tournament Vanberg and Congleton (1992) have shown that in environments in which agents can easily avoid 'unwanted' counterparts a 'moral program,' defined as one that always cooperates when it interacts with others but walks away from defectors, performs best in competition with a set of plausible alternative programs that meet each other in PD-type encounters.
[32] This fact is described quite succinctly in a remark that A.P. Lerner has made in reference to free trade "as a general rule" (quoted here from Hayek 1960: 428): "As with all general rules, there are particular cases where, if one knew all the attendant circumstances and the full effects in all their ramifications, it would be better for the rule not to be applied. But that does not make the rule a bad rule or give reason for not applying the rule where, as is normally the case, one does not know all the ramifications that would make the case a desirable exception." – The tempting option, to take advantage of the benefits from rule-following, but also to take advantage of 'preferred exceptions' does not exist. Adopting this 'strategy' means nothing other than going back to discretionary case-by-case choice, judging each case by its own merits.

agent has made in the past with different kinds of behavioral practices. Both theories assume that, in a sense, human behavior is based on a "calculus of advantage." The difference is that rational choice theory locates the calculus of advantage exclusively at the level of single actions, while a theory of rule-following behavior insists that, though humans are surely making situational choices, their behavior is also guided by dispositions that are adopted based on a "calculus of advantage" at the level of *rules of behavior*.[33]

To say that behavioral dispositions are based on a 'calculus of advantage' is, of course, not meant to imply that they are the product of deliberate calculation. It is meant to say that the process in which dispositions are formed must include some 'method of accounting' that keeps track of the comparative performance of different behavioral practices in different types of situations, i.e. of how well they work in helping agents to cope with recurrent problems of the kind they are likely to encounter in the type of environment in which they operate. Such 'methods of accounting' can be presumed to exist at – and, accordingly need to be theoretically specified for – three levels, the level of biological evolution, the level of cultural evolution, and the level of individual learning. At each of these three levels processes of learning or 'accumulation of knowledge' take place that operate on the same general principle of trial and error elimination or variation and selective retention, even though the specific modes of their operation are quite different (Campbell 1987). In the process of biological evolution the genetically encoded dispositions or behavioral programs have been shaped that we refer to when we speak of 'human nature.'[34] The process of cultural evolution shapes the socially shared traditions

[33] The contrast, discussed here, between a rational choice approach that explains single actions in terms of their expected payoffs and a theory of rule-following behavior that explains the rules that guide actions in terms of the payoff-*patterns* that they produce obviously parallels the contrast between act-utilitarianism and rule-utilitarianism.

[34] Cosmides and Tooby (2000: 98ff.): "The ancestral world posed recurrent information-processing problems, such as, What substances are best to eat? or, What is the relationship between others' facial expression and their mental states? Information-processing programs – food preferences and aversions, or rules for inferring emotions from facial expressions – acquired one set of design features rather than many others because the retained features better computed solutions to these information-processing problems…Natural selection can extract statistical relationships that would be undetectable to any organism. It does this by testing randomly generated alternative designs, each of which embodies different assumptions about the structure of the world, and retaining the ones that succeed most effectively … Designs whose features exploited these real but ontogenetically unobservable relationships outperformed those that depended on different relationships, or that only responded to conditions an individual could observe during his or her lifetime."

and rules of conduct that we refer to when we speak of 'human cultures' (Vanberg 1994a). The learning process that a person undergoes over her lifetime, and through which her personal repertoire of behavioral dispositions is formed, operates on the basis of the genetic heritage with which she is born and it takes place within a social environment that is characterized by a particular cultural heritage (Witt 1987).

My focus in the remainder of this paper is on the 'method of accounting' that works at the level of individual behavioral learning and in particular on the role that emotions play in this context.

5. Rule-Following, Moral Preferences and Emotions

The relation between rationality and emotions is an issue that has in recent times found growing attention in psychology and neuroscience (Damasio 1994) as well as in economics (Frank 1988; Elster 1996, 1998; Loewenstein 2000; van Winden 2001; Bosman, Sutter and van Winden 2005). A variety of conjectures have been advanced about how emotions may interfere with sober and prudent conduct or, by contrast, may induce agents to act more successfully than they would if guided only by pure rational calculation.[35] Economists who seek to incorporate emotions into a rational choice framework most commonly view emotions "as psychic costs or benefits that enter into the utility function on a par with satisfaction derived from material rewards" (Elster 1998: 64).[36] Whether modeling emotions as elements in the utility function, to be traded off with other 'ordinary' preferences over outcomes, can be "the full story" (Bosman, Sutter and van Winden 2005: 408) has been questioned, though, by several authors who point to what J. Elster (1998: 73) calls the "*dual role of emotions*." "The role of emotions," so Elster (ibid.) argues, "cannot be reduced to that of shaping the reward parameters for rational choice." What must also be accounted for is their role in "shaping choices as well as rewards" (ibid.), i.e. their role as "action tendencies" rather than as

[35] See e.g. Elster 1998: 59ff.; Cosmides and Tooby 2000: 93ff.; Loewenstein 2000; van Winden 2001: 491f.
[36] Bosman, Sutter and von Winden (2005: 208): "The role of emotions in decision making attracts growing attention in economics … The standard modeling approach is to include emotions as additional (psychological) costs or benefits in the utility function. Emotions simply enter on a par with material rewards in the decision calculus."

costs or benefits.[37] Similarly Bosman, Sutter and van Winden (2005: 412) argue that "emotions can be defined in terms of an action tendency … or pattern of readiness, which is the urge to execute a particular form of action or to abstain from a particular action."[38] And Loewenstein (2000: 428) notes that emotions are often experienced as "feelings that one should or should not take certain actions," feelings that may be in conflict with what "an analysis of the expected consequences of one's actions" invites one to do.

The role of emotions as 'action tendencies' to which the above quoted comments refer can, I submit, be systematically captured, and be clearly separated from their role as costs and benefits in utility functions, if they are interpreted as *preferences over actions*, i.e. as a motivational force that induces agents to take or to abstain from particular courses of actions without considering the costs or benefits that may be expected to result from them in particular instances. Interpreted in this sense, emotions account for the 'strength' of dispositions to follow particular rules of conduct in certain kinds of situations. They are, figuratively speaking, the 'currency' in terms of which the 'calculus of advantage' operates at the level of behavioral rules. They reflect how strongly committed agents are to follow rules rather than choosing actions opportunistically in light of their situational payoffs.

Even though he does not discuss the matter in terms of emotions, what John H. Holland has to say in his theory of rule-based adaptive agents[39] about the 'strength of rules,' i.e. their force in guiding behavior, can illuminate the role that, as I wish to argue, emotions play in inducing agents to follow rules even in the presence of distracting situational incentives. At the heart of Holland's theory of adaptive rule-following as well as of the computer simulations that he has designed to model the process of rule-based learning is the notion that *adaptive agents*, i.e. agents who are able to use experience to modify their behavior in beneficial ways, are equipped with a repertoire of rules that they adapt to the contingencies of their environment "as experience accumulates" (Holland 1995: 10). The adaptation results from a process of variation and selection by

[37] Elster (1998: 99) notes that an emotion such as "envy is more plausibly interpreted as an action tendency than as a cost." – Elster (1996: 1388): "(M)ost emotions are associated with a characteristic action tendency."
[38] Bosman, Sutter and van Winden (2005: 412) quote in this context N.H. Frijda's (1986: 78) statement: "Evidently, then, action tendencies are programs that have a place of precedence in control of action and of information processing."
[39] Holland 1995, 1998. – For a more detailed discussion of Holland's approach see Vanberg (2004: 12ff.).

consequences. In order for selection to systematically favor 'beneficial,' and to work against 'inferior' rules a feedback or accounting mechanism must be in place that assigns 'credit' to behavioral practices according to the contribution they make to an agent's ability to operate successfully in the environment that he faces.

In order to serve its function the method by which credit is assigned must, in particular, be able to give proper credit to behavioral practices or rules that are not themselves followed by immediate rewards, but rather serve in a stage-setting role in the sense of being part of extended chains of actions only the last links of which are directly 'rewarded.'[40] This is, quite obviously, of special relevance in the case of moral practices that typically do not generate immediate reward. In Holland's computer simulation credit assignment is modeled as a mechanism, called "bucket brigade algorithm" (Holland 1995: 56; 1992: 176ff.), that works somewhat analogous to the ways in which credits or rewards are assigned in markets in which only the final sellers receive the 'ultimate reward,' namely the price paid by consumers, but in which the revenue earned in the consumer market is transferred back along the production chain to the producers of the final product, the producers of inputs for the final product, the producers of inputs for the inputs, and so forth.

It is my conjecture that the processes in which emotions and behavioral dispositions are formed must operate in terms of a mechanism for 'credit assignment' that works in ways analogous to Holland's "bucket brigade algorithm," where the strength of emotions associated with particular kinds of behavioral practices – or the strength of dispositions to act in certain ways – reflects the 'credits' assigned to the respective practices during the agent's past behavioral history. How such 'credit assignment' actually works in practice is an issue that may be left aside in the present context.[41] It suffices here to point out that the perspective outlined above provides an explanatory

[40] Holland et al. (1986: 16): "Credit assignment is not particularly difficult when the system receives payoffs from the environment for a particular action – the system simply strengthens all the rules active at that time (a kind of conditioning). Credit assignment becomes difficult when credit must be assigned to early-acting rules that set the stage for a sequence of actions leading to payoff."

[41] As Smith (2003: 469) puts it, "the brain is capable of off-line subconscious learning." - The theories of operant and classical conditioning provide at least a partial answer to the question of how such 'subconscious learning' works by explaining how secondary reinforcers can be 'learned' by virtue of their association with primary reinforcers (Witt 1987: 112ff.; 2005f.), and how actions may be indirectly reinforced by being associated with actions that are followed by rewards.

account of the role of 'ethical' concerns that is in contrast to – and, as I submit, more consistent than – including them, as Fehr and others suggest, as ordinary preferences in an agent's utility function.

Like all preferences over actions or behavioral dispositions, moral preferences are the product of learning processes – including the processes of biological and cultural evolution[42] – in which experiences with the capacity of alternative behavioral practices to further the agent's wellbeing have been accumulated and have been 'condensed' in the agent's emotional attachment to the respective practices, i.e. the strength of his disposition to act in certain ways in certain types of situations. The strength of a person's moral preferences or dispositions, so understood, is a function of her learning history, i.e. of what she has learned – through direct as well as through indirect experience – about the reward-generating potential of moral practices within the socio-cultural environment in which she operates, given the constraints and capabilities her genetic heritage has her endowed with. Accordingly, differences in persons' moral preferences or their readiness to adhere to ethical rules need not reflect at all differences in their preferences over outcomes – as Fehr and others suppose – but may, instead, reflect differences in their 'implicit' theories of what serves their interests, theories that they may not be conscious of but that are incorporated in the behavioral dispositions that they have adopted as a result of past learning experiences.

That moral preferences as preferences over actions rather than an outcome-related 'inequality aversion' or 'concern for fairness' may explain the 'behavioral anomalies' observed in ultimatum experiments is suggested when one compares the behavior that subjects display in the original version of the experiment with how they behave in the modified version in which their performance in some symbolically assigned task decides who is to act as proposer and who as responder (Smith 1998: 12ff.). The fact that proposers in the modified version – in which they have (supposedly) 'earned' their role by better performance in the assigned task – divide the 'pie' much less generously than in

[42] Sen (2002a: 25): "There is also a challenging issue involved in the possibility that the broader values themselves are the result of evolutionary survival rather than reasoned pre-selection." – Sen (2002c: 217): "Insofar as following such 'habitual' rules, as opposed to relentless maximization according to one's goals, produces better results (even in terms of those very goals …), there will also be a 'natural selection' argument in favor of such behavior modes, leading to their survival and stability … This is an 'evolutionary' influence that works in a direction quite different from the survival of profit-maximizers as seen by Friedman."

the original version of the ultimatum experiment, and that responders are willing to accept much smaller amounts, is difficult to explain in terms of a general aversion against inequality of outcomes. It can be more readily accounted for if one assumes agents to be guided by behavioral dispositions that make them classify the two experimental setting as different types of problem-situations in which different kinds of behavior are appropriate. The original experiment they may classify as one in which the distribution of a benefit is at stake that was bestowed on the proposer as a matter of pure luck, and for such cases they may have learned a more generous distribution rule to be appropriate, while the modified version they may classify as one which is about distributing a benefit that has been produced by effort and for such cases they may have learned to consider a distribution rule appropriate that favors the person who contributed most in producing it.

That behavioral dispositions of the kind described may have already been encoded into our genetic heritage has, in fact, been argued by evolutionary psychologists like L. Cosmides and J. Tooby who point out that, because food sharing has been one of the standard problems our ancestors had to deal with throughout the evolutionary history of our species, we should expect the human mind to include a specialized module for sharing problems that triggers different behavioral dispositions depending on whether the benefits to be shared are mainly a fruit of luck or of effort.[43] Whether or not biological evolution has, indeed, produced the domain-specific modules that Cosmides and Tooby postulate, what is highly plausible is that learning processes may produce behavioral dispositions of exactly the same kind because they provide in general useful guidance for how to deal with the two kinds of sharing problems. Evidence for such dispositions, whether genetically encoded or learned, is provided, for instance, by anthropological

[43] Cosmides and Tooby (1994b: 108f.): "(D)ifferent kinds of sharing rules benefit individuals in different situations. For example, when the variance in foraging success of an individual is greater than the variance for the band as a whole, bandwide food sharing buffers the variance. In essence, the individual stores food in the form of social obligation." – Food sharing as a form of social insurance is, as Cosmides and Tooby argue, typically functional for food items like large game, where luck plays a significant role in getting them and where the item exceeds the momentary consumption needs of the individual, while it may be dysfunctional for food items like herbs or fruits the collecting of which is mainly a matter of effort and where sharing might invite free-riding. Translating this into a general rule Cosmides and Tooby (1994a: 331) note: "These mechanisms should make sharing rules appealing in conditions of high variance, and unappealing when resource acquisition is a matter of effort rather than luck." And they conclude (1994b: 110): "Because foraging and sharing are complex adaptive problems with a long evolutionary history it is difficult to see how humans could have escaped evolving highly structured domain-specific psychological mechanisms for solving them."

reports on food-sharing (Ridley 1996: 89ff.) and by laboratory experiments which show that, quite generally, "agents behave differently if their own earnings are at stake ('effort') or a budget allocated to them by the experimenter ('no effort')" (Bosman, Sutter and von Winden 2005: 407), and that they display different attitudes towards others depending on whether they 'earned' their resources or got them by pure chance.

For agents to adopt moral preferences or dispositions to follow moral rules does not mean, of course, that they become entirely oblivious to the overall incentive structure of the choice situations they are facing, responding only to the 'clues' that let them classify a given situation as one to which a particular rule applies. Even though human behavior, including moral conduct, is surely 'routinized' to a large extent in the sense that much of our everyday conduct is carried out semi-automatically without any involvement of conscious deliberation, we cannot simply 'switch off' our capacity for rational calculation, and anything unusual in the choice situations we encounter may activate this capacity.[44] The 'function' of emotions, i.e. their 'evolutionary rationale,' lies, I submit, exactly in the role they play in 'stabilizing' man's dispositions to follow rules in the presence of opposing situational incentives.[45] The conflict that persons experience in such situations is not about a trade-off between different elements in the utility function as is suggested by authors like Fehr who treat concerns for equity, fairness and the like as preferences over outcomes.[46] Instead , it is a conflict between a person's preference for acting according to a rule that the above discussed 'accounting mechanism' has identified

[44] As Vernon Smith (2003: 468f.) notes: "Since our theories and thought processes about social systems involve the conscious and deliberate use of reason, it is necessary to constantly remind ourselves that human activity is diffused and dominated by unconscious, automatic, neuropsychological systems that enable people to function effectively without always calling upon the brain's scarcest resource – attentional and reasoning circuity. This is an important economizing property of how the brain works. … The challenge of any unfamiliar action or problem appears first to trigger a search by the brain to bring to the conscious mind what one knows that is related to the decision context. Context triggers autobiographic experiental memory."

[45] Elster (1996: 1389): "(S)ocial norms as injunctions to behavior … are non-outcome-oriented … and are sustained by internalized emotions … (T)ey differ from the outcome-oriented injunctions of instrumental rationality in that the targeted action is to be performed because it is intrinsically appropriate, not because it is a means toward a desired goal."

[46] In discussions on how to account for subjects' behavior in experiments like the ultimatum game emotions are occasionally invoked as explanatory variables, typically though without explicit recognition of their role as preferences over actions rather than over outcomes. Camerer (2003: 44), for instance, speaks of the "emotional reaction to unfairness which is highlighted by the ultimatum game" and notes that it "is useful to distinguish between the emotions or reasons that cause responders to reject (call it 'anger') from the emotion A might feel when B does something unfair to a third party C (call it 'indignation')."

as one that works well in situations of the type currently encountered and her preference for outcomes that her situational calculation tells her she may achieve by deviating from the moral rule. The intensity of this conflict will depend on the strength of her 'moral emotions' on the one side and the attractiveness of the outcome that can be achieved by rule-violation on the other.[47]

What I have described above as the function of emotions in 'stabilizing' rule-following behavior has been extensively discussed by Robert Frank (1988) under the rubric of "emotions as commitment devices," emphasizing the fact that as such devices emotions can help persons to act in ways more conducive to their long term wellbeing than opportunistic rational calculation would have them.[48] As Frank (1988: 7) puts it, "emotions often predispose us to behave in ways that are contrary to our narrow interests, and being thus disposed can be an advantage." Summarizing the thrust of the "commitment model" that he proposes, Frank (ibid.: 258f.) states: "The commitment model is a tentative first step in the construction of a theory of un-opportunistic behavior. It challenges the self-interest model's portrayal of human nature in its own terms by accepting the fundamental premise that material incentives ultimately govern behavior. … The emotions that lead people to behave in seemingly irrational ways can thus indirectly lead to greater material wellbeing. Viewed in these terms, the commitment model is less a disavowal of the self-interest model than a friendly amendment to it. Without abandoning the basic materialist framework, it suggests how the nobler strands of human nature might have emerged and prospered."[49]

6. Conclusion

[47] Conflicts with moral preferences may not only arise from 'situational temptations' in the sense of preferences for outcomes that may be obtained by deviating from moral rules. Conflicts may also arise where different moral principles call for different actions in the given situation and where agent's moral preferences for the respective rules collide. As an example consider the famous stylized case of someone who is hiding an innocent fugitive and is asked whether the fugitive is in his home. The conflict is here between following the rule to speak truthfully or following the rule to help persons in danger.

[48] For a similar argument see Hirshleifer 1987.

[49] Ridley (1996: 132ff.) reports on Frank's and other contributions on emotions as "mental devices for guaranteeing commitment."

In response to behavioral evidence that calls the general validity of the standard model of self-interested, rational economic man into question a number of economists have started seriously to consider how the model may be modified in order to accommodate observed 'anomalies.' Their efforts are typically focused, though, on only one of the two components of the rational choice model, namely the self-interest assumption, while the possibility of modifying the rationality assumption is rarely considered. Quite apparently, economists find it much easier to give up the self-interest assumption and to be more flexible in how they specify the content of the utility function than to give up the notion that individuals rationally maximize their utility, whatever it may be that gives them utility. The reason for this characteristic asymmetry presumably is that merely redefining the content of the utility function allows them to continue applying the standard modeling techniques of their profession that they are accustomed to, techniques that since Walras and Jevons are inherently tied to the maximization logic.

In this paper I have developed an argument for why focusing on the content of the utility function means attacking the problem at the wrong end. What may be in need of revision is, I submit, not so much the self-interest assumption understood as the notion that individuals seek to further their own wellbeing, but the notion that in pursuing their own interests individuals always act as rational case-by-case maximizers whose choices are solely determined by the expected consequences of their actions. I have argued that many aspects of human conduct can be more consistently and more plausibly accounted for if one recognizes that human behavior may be guided by preferences over actions as such in addition to preferences over outcomes. And I have discussed how the notion of preferences over actions is related to the phenomenon of rule-following behavior, with a particular focus on moral preferences and on the role of emotions in moral conduct.

I have not addressed the issue of what distinguishes moral preferences from other behavioral dispositions, and which among the rules that the forces of biological evolution, of cultural evolution and of individual learning tend to favor qualify as 'moral rules.' Nor did I address the question under what conditions individuals are more or less likely to acquire moral preferences, and under what conditions moral rules are likely to gain effective recognition in social groups. Discussing these issues would require more

than one separate paper.[50] What can be said in general, though, is that the rules that are commonly classified as 'moral rules,' both in everyday life as in scholarly discourse, tend to be rules that help groups of persons to solve prisoners' dilemma type problems, i.e. problems where the direct pursuit of individual interest produces collective outcomes that are disadvantageous to all persons in the group compared to what they could realize if they were to follow suitable rules of conduct. Elsewhere (Vanberg 2002b) I have suggested that such rules can be said to be in the *common constitutional interest* of the persons involved, rules that they could agree upon as means for securing mutual advantage.[51] Moral rules can, in this sense, be viewed as rules that work to the common benefit of all parties but for which situational incentives to deviate exist.

The capacity of groups and societies to realize mutual gains from cooperation will depend on their ability to adopt rules of conduct that are in their members' common constitutional interests and on their ability to create conditions under which individuals are likely to acquire moral preferences or dispositions for following such rules.

[50] Some aspects of these issues are addressed in Vanberg 1987, 2002b; Vanberg and Buchanan 1988.
[51] In a similar spirit Sen (1973: 250) notes in reference to moral rules of behavior: "In different periods of history in different social situations in response to different types of problems particular rules of behavior have been proposed which have in common the analytical property of trying to generate the results of a social contract without there being such formal contract."

References

Arrow, Kenneth J. 1996: Preface, in: K.J. Arrow, E. Colombatto, M. Perlman and C. Schmidt (eds.), *The Rational Foundations of Economic Behavior*, Proceedings of the IEA Conference held in Turin, Italy, New York: St. Martin's Press, xiii-xvii.

Becker, Gary S. 1996: *Accounting for Tastes*, Cambridge, Mass., and London: Harvard University Press.

Bicchieri, Cristina 2004: "Rationality and Game Theory," in: A.R. Mele, P. Rawling (eds.), *The Oxford Dictionary of Rationality*," Oxford: Oxford University Press, 182-205.

Bolton, Gary E. and Axel Ockenfels 2000: "ERC: A Theory of Equity, Reciprocity, and Competition," *The American Economic Review* 90, 166-193.

Bosman, Ronald, Matthias Sutter and Franz van Winden 2005: "The impact of real effort and emotions in the power-to-take game," *Journal of Economic Psychology* 26, 407-429.

Camerer, Colin F. 2003: *Behavioral Game Theory. Experiments in Strategic Interaction*, New York: Russel Sage Foundation.

Campbell, Donald T. 1987: "Blind Variation and Selective Retention in Creative Thought as in Other Knowledge Processes," in: Radnitzky, Gerard and W. W. Bartley, III (eds.), *Evolutionary Epistemology, Rationality, and the Sociology of Knowledge*, La Salle, Illinois: Open Court, 91-114.

Cosmides, Leda and John Tooby, 1994a: "Better than Rational: Evolutionary Psychology and the Invisible Hand," *American Economic Review, Papers and Proceedings*, 84, 327-332.

Cosmides, Leda and John Tooby, 1994b: "Origins of domain specificity: The evolution of functional organization," in: Hirschfeld, Lawrence A. and Susan A. Gelman (eds.), *Mapping the mind – Domain specificity in cognition and culture*, Cambridge/New York: Cambridge University Press, 85-116.

Cosmides, Leda and John Tooby 2000: "Evolutionary Psychology and the Emotions," in: M. Lewis and J.M. Haviland-Jones (eds.), *Handbook of Emotions*, 2nd ed., New York, London: The Guildford Press, 91-115.

Damasio, Antonio R. 1994: *Descartes' error. Emotion, reason and the human brain*, New York: Putnam.

Elster, Jon 1996: "Rationality and the Emotions," *The Economic Journal* 106, 1386-1397.

Elster, Jon 1998: "Emotions and Economic Theory," *Journal of Economic Literature* XXXVI, 47-74.

Fehr, Ernst and Klaus M. Schmidt 1999: „A Theory of Fairness, Competition and Cooperation," *Quaterly Journal of Economics* 114, 817-868.

Fehr, Ernst and Armin Falk 2003: "Reciprocal Fairness, Cooperation and Limits to Competition," in: E. Fullbrook (ed.), *Intersubjectivity in Economics, Agents and Structure*, London and New York: Routledge, 28-42.

Fehr, Ernst and Urs Fischbacher 2002: "Why Social Preferences Matter – The Impact of Non-Selfish Motives on Competition, Cooperation and Incentives," *The Economic Journal* 112, C1-C33.

Fehr, Ernst and Urs Fischbacher 2003: "The nature of human altruism," *Nature* 425, 785-791.

Fehr, Ernst and Klaus M. Schmidt 2003: „Theories of Fairness and Reciprocity: Evidence and Economic Applications," in: M. Dewatripont, L.P. Hansen, S. Turnovski (eds.), *Advances in Economic Theory, Eighth World Congress of the Econometric Society*, Vol. 1, Cambridge: Cambridge University Press, 208-257.

Frank, Robert H. 1988: *Passions Within Reason. The Strategic Role of Emotions*, New York and London: W.W. Norton & Company.

Frey, Bruno S. and Felix Oberholzer-Gee 1997: "The Cost of Price Incentives: An Empirical Analysis of Motivation Crowding-Out," *The American Economic Review* 87, 746-755.

Frey, Bruno S., Matthias Benz and Alois Stutzer 2004: "Introducing Procedural Utility: Not Only What, but Also How Matters," *Journal of Institutional and Theoretical Economics* 160, 377-401.

Frijda, Nico H. 1986: *The Emotions*, Cambridge: Cambridge University Press.

Georgescu-Roegen, Nicholas 1971: *The Entropy Law and the Economic Process*, Cambridge, Mass., und London, England: Harvard University Press.

Gintis, Herbert and Rakesh Khurana 2006: "Corporate Honesty and Business Education: A Behavioral Model," Paper presented at IEA Workshop on Corporate Social Responsibility (CSR) and Corporate Governance, Trento, Italy, 11-13 July, 2006.

Gueth, Werner, Rolf Schmittberger and Bernd Schwarze 1982: "An Experimental Analysis of Ultimatum Bargaining," *Journal of Economic Behavior and Organization* 3, 367-88.

Hayek, Friedrich A. 1960: *The Constitution of Liberty*, Chicago: The University of Chicago Press.

Hayek, Friedrich A. 1967: *Studies in Philosophy, Politics and Economics*. Chicago: The University of Chicago Press.

Hayek, Friedrich A. 1976: *The Mirage of Social Justice*, Vol. 2 of *Law, Legislation and Liberty*, London: Routledge & Kegan Paul.

Heiner, Ronald A. 1993: "The Origin of Predictable Behavior," *The American Economic Review* 73, 1983, 560-595.

Hirshleifer, Jack 1987: "On the emotions as guarantors of threats and promises," in: J. Dupre (ed.), *The latest on the best: Essays on evolution and optimality*, Cambridge, MA: MIT Press, 307-326.

Holland, John H. 1992: *Adaptation in Natural and Artificial Systems*, Cambridge, Mass., and London: The MIT Press.

Holland, John H. 1995: *Hidden Order: How Adaptation Builds Complexity, Reading*, Massachusetts: Helix Books.

Holland, John H. 1998: *Emergence: From Chaos to Order*, Reading, Massachusetts: Perseus Books.

Holland, John H., H.J. Holyoak, R.E. Nisbett and P.R. Thagard 1986: *Induction*, Cambridge, Mass.: MIT Press.

Kahneman, Daniel, Jack L. Knetsch and Richard H. Thaler 1987: "Fairness and the Assumptions of Economics," in: R.M. Hogarth and M.W. Reder (eds.), *Rational Choice – The Contrast between Economics and Psychology*, Chicago and London: The University of Chicago Press, 101-116.

Kliemt, Hartmut 2005: "Public choice and political philosophy: Reflections on the works of Gordon Spinoza and David Immanuel Buchanan," *Public Choice* 125, 203-213.

Loewenstein, George 2000: "Emotions in Economic Theory and Economic Behavior," *The American Economic Review* 90, Papers and Proceedings, 426-432.

McFadden, Daniel 2005: "The New Science of Pleasure – Consumer Behavior and the Measurement of Well-Being," Frisch Lecture, Econometric Society World Congress, London, August 20, 2005.

Pettit, Philip 2005: "Construing Sen on Commitment," *Economics and Philosophy* 21, 15-32.

Ridley, Matt 1996: *The Origins of Virtue – Human Instincts and the Evolution of Cooperation*, New York: Viking.

Sen, Amartya 1973: "Behaviour and the Concept of Preference," *Economica*, *New Series*, 40, 241-259.

Sen, Amartya 2002: *Rationality and Freedom*, Cambridge, Mass., and London: The Belknap Press of Harvard University Press.

Sen, Amartya 2002a: "Introduction: Rationality and Freedom," in: A. Sen 2002, 3-64.

Sen, Amartya 2002b: "Maximization and the Act of Choice," in: A. Sen 2002, 158-205.

Sen, Amartya 2002c: "Goals, Commitments, and Identity," in: A. Sen 2002. 206-224.

Sen, Amartya 2005: "Why Exactly is Commitment Important for Rationality?" *Economics and Philosophy* 21, 5-14.

Smith, Vernon L. 1998: "The Two Faces of Adam Smith," *Southern Economic Journal* 65, 1-19.

Smith, Vernon L. 2003: "Constructivist and Ecological Rationality in Economics," *The American Economic Review* 93, 465-508.

Vanberg, Viktor J. 1993: "Rational Choice, Rule-Following and Institutions: An Evolutionary Perspective," in: B. Gustafson, Ch. Knudsen, U. Mäki (eds.), *Rationality, Institutions and Economic Methodology*, London and New York: Routledge, 171-200.

Vanberg, Viktor J. 1994a: "Cultural Evolution, Collective Learning and Constitutional Design," in: D. Reisman (ed.), *Economic Thought and Political Theory*, Boston, Dordrecht, London: Kluwer, 171-204.

Vanberg, Viktor J. 1994b: *Rules and Choice in Economics*, London and New York: Routledge.

Vanberg, Viktor J. 1987: *Morality and Economics: De Moribus Est Disputandum*, (Social Philosophy & Policy Center, Original Papers No. 7), New Brunswick: Transaction Book (reprinted in Vanberg 1994b)

Vanberg, Viktor J. 2002a: "Rational Choice vs. Program-based Behavior: Alternative Theoretical Approaches and their Relevance for the Study of Institutions," *Rationality and Society*, Vol. 14, 2002, 7-53.

Vanberg, Viktor J. 2002b: "Constitutional Economics and Ethics – On the Relation Between Self-Interest and Morality," in Brennan, Geoffrey, Hartmut Kliemt, Robert D. Tollison (eds.), *Methods and Morals in Constitutional Economics – Essays in Honor of James M. Buchanan*, Springer-Verlag, Berlin, Heidelberg, 485-503.

Vanberg, Viktor J. 2004: "The Rationality Postulate in Economics: Its Ambiguity, its Deficiency and its Evolutionary Alternative," *Journal of Economic Methodology* 11, 1-29.

Vanberg, Viktor J. and James M. Buchanan 1988: "Rational Choice and Moral Order," *Analyse und Kritik*, Vol. 10, 1988, 138-160 (reprinted in Vanberg 1994b).

Vanberg, Viktor J. and Roger D. Congleton 1992: "Rationality, Morality and Exit," *American Poltical Science Review* 86, 418-431.

Walras, Léon 1954 [orig. 1874]: *Elements of Pure Economics – Or the Theory of Social Wealth*, Homewood, Ill.: Richard D. Irwin Inc. (Reprinted 1984 by Orion Editions, Philadelphia, PA)

Winden, Franz van 2001: "Emotional Hazard, exemplified by taxation-induced anger," *Kyklos* 54, 491-506.

Witt, Ulrich 1987: *Individualistische Grundlagen der evolutorischen Ökonomik*, Tübingen: J.C.B. Mohr (Paul Siebeck).

Witt, Ulrich 2005: "From Sensory to Positivist Utilitarianism and Back – The Rehabilitation of Naturalistic Conjectures in the Theory of Demand," Papers on Economics & Evolution #0507, Max Planck Institute of Economics, Jena, Germany.