Koboldt, Christian

Working Paper

Rational Samaritans, Strategic Moves, and Rule-Governed Behavior: Some Remarks on James Buchanan's "Samaritan's Dilemma"

CSLE Discussion Paper, No. 95-02

Provided in cooperation with:
Universität des Saarlandes (UdS)

Suggested citation: Koboldt, Christian (1995) : Rational Samaritans, Strategic Moves, and Rule-Governed Behavior: Some Remarks on James Buchanan's "Samaritan's Dilemma", CSLE Discussion Paper, No. 95-02, http://hdl.handle.net/10419/23049

Rational Samaritans, Strategic Moves, and Rule–Governed Behavior: Some Remarks on James Buchanan's "Samaritan's Dilemma"

Christian Koboldt *

April 1995

> "Would you tell me, please, which way I ought to go from here?" "That depends a good deal on where you want to get to," said the Cat. "I don't much care where —" said Alice. "Then it doesn't matter which way you go," said the Cat. "— so long as I get *somewhere*," Alice added as an explanation. "Oh, you're sure to do that," said the Cat, "if you only walk long enough."
> Lewis Carroll: Alice in Wonderland, Wordsworth's Classics, p. 55

1 What to Do?

In his paper on the "Samaritan's Dilemma", first published in 1975[1], James Buchanan claimed to present an "essay in prescriptive diagnosis" (p. 169). "Many different social problems ... " are "separate symptoms of the same disease ...": "We may simply be too compassionate for our own well beeing or for that of an orderly and productive free society." (ibid.) or — put alternatively — "[m]odern man has become incapable of making the choices that are required to prevent his exploitation by predators of his own species." (p. 173) The treatment of this disease — once recognized as a disease — requires man to "accept the possible necessity of acting *strategically* instead of pragmatically." (ibid.)

As is well known from the game–theoretic analysis of interactive decisions[2]

*Center for the Study of Law and Economics, Department of Economics, Universitt des Saarlandes, Germany. I am indebted to Matthias Leder and Dieter Schmidtchen for helpful disussions and critical comments. A first version of this paper has been presented at the Conference "Freedom in Constitutional Contract — James Buchanan meets his german friends and critics", Bielefeld, April 19–21 1993.

[1]The article originally was published in the Volume "Altruism, Morality and Economic Theory", edited by E.S.Phelps. All further citations of page numbers without an author's name refer to the reprint of the paper in the collection "Freedom in Constitutional Contract" (Buchanan [1977]).

[2]Binmore [1992] provides an excellent introduction to game theory. Dixit/Nalebuff [1991] illustrate game–theoretic reasoning in a non–technical manner and with reference to situations from everyday life. The book on game theory by Rasmusen [1989] aims at the straightforward application of the theoretical apparatus to 'economic problems'.

the main problem with strategic moves is their credibility. For the prescribed treatment to be effective credible threats or promises must be possible. But, as Buchanan notes, for credibility to be established, the samaritan "must accept the prospect of personal injury." (p. 175) Even if it is correct that "[t]o the extent that the parasite believes that the samaritan has, in fact, adopted a strategic behavioral plan and that he will, in fact, abide by this plan once adopted, there need be *no* loss to the samaritan at all", the problem lies in the premises.

This paper focusses on the problem of credibility, i.e. on the question whether the samaritan can *in fact* claim to behave in accordance with some announced behavioral pattern. Therefore I would like to sidestep the issue of diagnosis: it will not be asked whether the situation presented by Buchanan can really be qualified as a disease or, if so, requires treatment. Rather it will be accepted as a matter of fact. The aim will be to look at the possibility of an effective cure, the preconditions for the therapy to be successful and the long–run effects if the preconditions are fulfilled.

The analysis will proceed as follows: In the next section the problem is presented again in the game–theoretic formulation employed by Buchanan. It turns out to be necessary for the applicability of strategic moves that the game is repeated. The object of further analysis therefore has to be the repeated (sequential) game (or supergame). Section three deals with the possibility of strategic behavior in repeated games with or without a finite time horizon. In this section the central problems of credibility will be tackled and the preconditions for effective strategic moves will be worked out. The final section looks for possible long run effects by pointing to extensions of the analysis to issues of evolutionary stability. This extensions are supposed to complement the remarks Buchanan makes on the evolutionary origin of the disease, i.e. that "modern man has 'gone soft' " (p. 173).

2 The Samaritan's Dilemma

The Samaritan's Dilemma is the result of an interaction between two players. Player A is a 'potential samaritan' (p. 174 ff.), the other player (B) is a 'potential parasite'. A has to choose between two courses of action, labelled 'bc' for 'behave charitably' and 'bn' for 'behave noncharitably'. Likewise B has two alternative options, namely 'w' for 'work' and 'rw' for 'refuse work'.

The pay–offs for A and B are determined by the combination of the options chosen by them. They can be represented by the matrix of figure 1 where the numbers in the lower left corner of the cells represent the pay–offs to A, the number in the upper right corners those to B[3].

[3]The terms 'active' and 'passive' samaritan's dilemma are due to Buchanan. The reason for this labelling should become clear when both situations are interpreted as sequential games (see the discussion below).

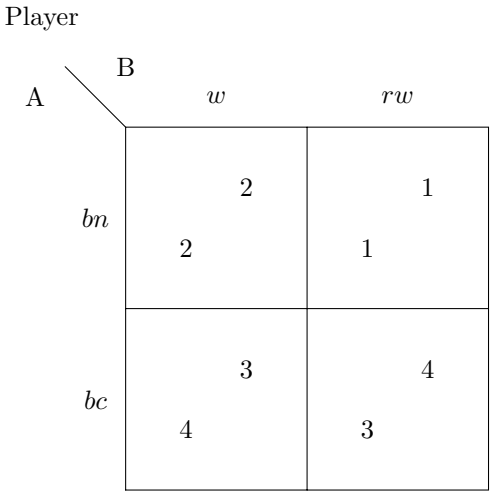

Figure 1: Active samaritan's dilemma: Pay–offs for the single encounter

If A and B meet once (and only once) then the resulting play will be (bc,rw) with pay–offs of 3 to A and 4 to B. The result holds for the simultaneous game as well as for the sequential game, where one player moves first and the move is observable for the other player[4]. Furthermore, the result is independent of the order of moves because bc strongly dominates bn. So A, if rational, always will choose bc regardless of B's choice or A's expectations about B's choice.

The important point is that there is no room for strategic manoeuvres. Any preplay announcement by A about his intention to choose bn — either simultaneously or after having observed B's move — is not credible, because this choice would be dominated by bc. Even if A is the first mover in a sequential game, he cannot gain by choosing bn, even if this has an effect on B's move. Choosing bn pays at best 2 instead of 3, the worst possible result from choosing bc.

Choosing bn instead of bc may make sense, but only if A expects his relationship with B to continue. Then the utility loss from choosing bn once may be compensated by the utility gain from reaching the combination (bc,w) in later encounters. Whether this actually will work and, if ever, under which

[4]From the discussion in Buchanan's paper it is not clear whether the single encounter has to be interpreted as a simultaneous or as a sequential game. Therefore, both interpretations should be considered. Note, however, that the matrix of figure 1 does not adequately represent the strategic form of the sequential game. Since a strategy in a sequential game requires every player to specify his decision at every decision node and the player moving second has two decision nodes in the game tree, he has 2×2 possible strategies, e.g. for A labelled (bn,bn), (bn,bc), (bc,bn) and (bc,bc). The strategic form of the sequential game is represented by a 2×4 matrix with 4 columns/rows for the player moving second.

conditions, will be the topic of the next section. But before turning to this question let us look at the second version of the samaritan's dilemma which can be represented ceteris paribus by the matrix of figure 2.

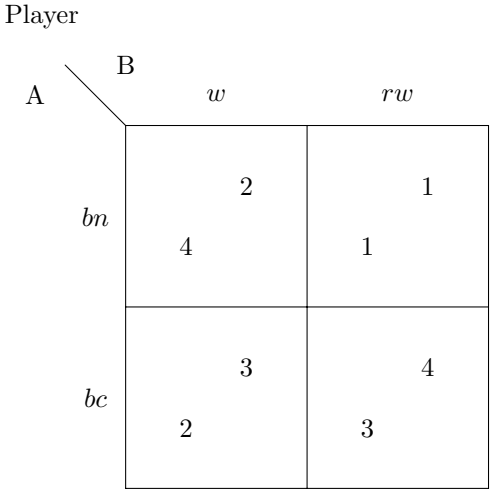

Figure 2: Passive samaritan's dilemma: Pay–offs for the single encounter

If interpreted as a simultaneous game, this version has two possible Nash–equilibria characterized by the plays (bn,w) and (bc,rw). The first one is preferred by the samaritan, the second one by the parasite. Which one of the two equilibria is to be selected in the simultaneous game cannot be answered by exlusively considering rationality on the side of the players. Both equilibria are equally likely, and even 'prominence' of one course of action will not help if both players are rational and know about their rationality[5].

If the game is played sequentially[6] there exists a so called 'first–mover–advantage': the player moving first can determine the resulting equilibrium by choosing his best option in expectation of the best answer given by the other player. Again, there is no room for strategic manoeuvres: if A moves first, then no problem arises — by not supporting the parasite the latter can be forced to work. But if B moves first, A cannot credibly threaten not to support the starving parasite, because this reaction to B's choice of rw would hurt A himself. And again, to hurt oneself once can make sense only if one expects this to have influence on the future behavior of the opponent in an ongoing relationship.

Although it is not intended to discuss questions of diagnosis here, a short

[5]This holds even if one equilibrium Pareto–dominates the other. To the problem of equilibrium selection in coordination games see Gilbert [1989] or Sugden [1991].

[6]As for the representation in strategic form the disclaimer of the previous footnotes applies.

remark seems to be in order: the dilemma does not result from the fact that the equilibrium of the first version or one of the equilibria of the second version are Pareto–dominated by another possible outcome, as it is the case in the prisoner's dilemma. Rather, the active samaritan's dilemma "that player A is worse off than he would be in cell III [the combination of (bc,w) in figure 1, C.K.]" (p. 170) seems to result from an exclusive concern for the well–being of the samaritan[7]. Buchanan's diagnosis of a disease refers to thism welfare loss to the samaritan, resulting from his concern for the parasite's well–being.

Let us accept the diagnosis and look for a possible therapy. The discussion will concentrate on the active samaritan's dilemma, because the problem of credibility is much more severe in this case. Additionally, the sequence of moves in one single encounter does not matter, so that we can analyze the situation as a supergame with the matrix of figure 1 as the (simultaneous) stage game.

3 Any Hope for a Therapy?

Buchanan claims that the dilemma can be resolved by "acting strategically instead of pragmatically" (p. 173). This is to say that player A should try to influence player B's choices by changing B's expectations about A's future choices. Thus, a strategic move is not meant to *force* one's opponent into a certain course of action (by only leaving open some options to him), rather it is aimed at changing his mind about one's own future behavior. It is exactly this feature of strategic moves that gives rise to problems of credibility.

As we have seen in the last section, any attempt on A's side to threaten B with the action bn is only "cheap talk" in the stage game. If A is rational in the sense of expected utility maximization and B knows about A's rationality, he can predict A's choice of bc with certainty.

But what happens when there are repetitions, i.e. if A and B will play the stage game n times ($n > 1$)? Then, perhaps, choosing bn at an early stage may pay on the whole, if (and only if) the expectations of player B thereby are changed in a way which makes him prefer the choice of w — even if A returns to choosing bc afterwards. Will this condition hold for rational players and will it depend on n?

[7]Suppose that a social welfare function exists which allows the ordering of all possible outcomes. This requires perhaps the assumption of interpersonal comparability of utility and therefore a cardinal utility conception. Then a dilemma requires the result of individual actions to be ranked inferior (with respect to the social welfare function) to some other outcome. For this to be the case, the parasite's pay–off has to be weighted less than the samaritan's to generate a dilemma in the first case, and it has to be weighted less than 6/7 of the samaritan's to generate a dilemma in the second case if one refers to the sum of the pay–offs as a measure for the social welfare. Only then will the combination of (bn,w) be better than the equilibrium (bc,rw).

3.1 Strategic Behavior in Finite Horizon Games

Let us assume that both players are rational and that the time horizon is finite (this means that n is a positive integer). The structure of the game (i.e. the pay–offs of the stage game and the actions available to the players at each stage) as represented by the matrix of figure 1 as well as n are common knowledge. The players try to maximize the (discounted)[8] sum of the pay–offs over the n stages of the game.

Thus, to choose bn instead of bc at stage i will pay only if the opponent thereby can be expected to choose w at future stages and the advantages from these choices at later stages outweigh the actual utility loss. This means that for A (with δ as discount factor)

$$1 + \sum_{t=i+1}^{n} \delta^{(t-i)} \cdot 4 > \sum_{t=i}^{n} \delta^{(t-i)} \cdot 3$$

Let us assume that this inequality holds. This will only be a necessary, not a sufficient condition for the choice of bn. In this inequality it is supposed that by choosing bn once the samaritan can induce B to choose w for the rest of the game, even if A turns back to bc after i.

This in turn requires B to expect A to choose bn at future stages whenever he chooses rw at some stage $j > i$ *and* that B does not gain by receiving a pay–off of 4 once (at stage j) and 2 afterwards. For A to react as specified, the inequality must hold for period j *and* A must expect B to choose w for the rest of the game in response to the choice of bn in j and so on[9]. The problem of credibility is one of iterated expectations about the opponents future behavior, which depends on the opponents expectations of one's own future behavior which ....

But as n is assumed to be finite there is a last stage where expectations do not matter. The choice at the last stage can be analyzed without any reference to expectations in the same way as the choice for the stage game. A will, if

[8] This assumption can easily be relaxed. As we shall see, discounting is not the problem in this case, so even if future pay–offs are not to be discounted, the solution Buchanan proposes will not work.

[9] Otherwise the aim of A, i.e. to arrive at a combination of (bc,w), can be reached only if B continuously and mistakenly expects A to play bn even if this expectation cannot be fulfilled whenever A's preferred outcome is realized. Because of the dominance of bc A would never try to induce B's choice of w if this confronts him with the cost of really choosing bn at each stage. Thus, in the active samaritan's dilemma with finite time horizon the solution in Buchanan's sense, i.e. where the Samaritan can credibly threaten to behave noncharitably by actually acting in this way for a limited time and then switching back to bc, seems to require that B chooses continuously on wrong expectations. This hardly can be classified as an equilibrium in any sense. The prescription would be more realistic for the passive samaritan's dilemma, where B can expect A to choose bn as punishment for deviation to rw if therby B is forced back to w. In this case for A to choose bn at stage i, the inequality $1 + \sum_{t=i+1}^{n} \delta^{(t-i)} \cdot 4 > \sum_{t=i}^{n} \delta^{(t-i)} \cdot 3$ must hold as well. But the chain of expectations is somewhat simplified. For B to play w at all future stages, he must expect A to play bn at all future stages. But A will play bn at stage $i+1$ only if the inequality holds for $i+1$ and so on.

rational, choose *bc* because nothing is to be gained by choosing *bn* — even if thereby B's expectations are to be changed, there is no future to enjoy the benefits. Thus, B will choose *rw* in the last period.

If A is rational and B knows A to be rational, he can predict A's choice at the last stage with certainty at the last but one stage. If B is rational, and if A knows that B is rational and if A knows that B knows that A is rational then A can predict B's choice at the last stage as well. But then the condition imposed by the inequality can never hold at stage $n-1$, because A can predict B's choice of *rw* at n and therefore can gain nothing by choosing *bn*. A should choose *bc* at stage $n-1$.

This argument can be extended backwards to the first period, showing that the only *subgame perfect equilibrium*[10] is the choice of (*bc*,*rw*) in each period. This troublesome result of requiring equilibria to be subgame perfect is well known and gives rise to an "inconsistency between game theoretical reasoning and plausible human behavior" (Selten [1978:127]). It is especially vexing in the case of dilemmas, where the prescription of "acting strategically" obviously fails to get rational players out of the equilibrium in which they are trapped: "Suppose that you and I face and know that we face a sequence of prisoner's dilemmas of known finite length (...) There is a well known argument — the backward induction argument — to the effect that, in such a sequence, agents who are rational and who share the belief that they are rational will defect in every round. (...) And yet, it appears that I might do better (...) to cooperate provided you reciprocate. This is the backward induction paradox." (Pettit and Sugden [1989:169]) To the extent that the samaritan's dilemma is a dilemma, the paradoxical effect of the game–theoretically correct reasoning is the same.

But there is one condition for the appropriateness of the backward induction argument: it is the iterated sequence of 'A knows that B knows that A knows ...' one can find in the reasoning backwards from the last stage, as sketched above. This iterated sequence becomes more and more complicated the greater is the distance to the last stage. An ingenious escape from this iterated sequence commonly employed by game theorists is the assumption of 'Common Knowledge of Rationality', or CKR. Additionally to the game structure and the number n, every player knows that every player is rational and that every player knows that every player is rational and so on ...

Given CKR, there is, at first sight, no problem in analyzing a sequential game with finite length (a repeated game with known end period) by starting at the last node and reasoning backwards. When taking a second look, however, one finds a possible inconsistency in the reasoning: it is necessary for the predictions of the choices at later stages to make sense that the premises employed in the derivation of these predictions are valid (cf. Bonanno [1989]). Thus, the predicted choice of *rw* by B at the last stage is valid only if CKR can be

[10] For the notion of subgame perfection and its extension to perfection (or "trembling–hand–perfection", as it is called by some authors) see Selten [1975].

maintained at the last stage.

More formally: Given CKR, one can derive for every stage t of our game a proposition P_t which says: 'If stage t is reached, rational play requires the choice of actions (bc,rw)'. All propositions are known to the players and together they define the subgame perfect equilibrium.

The important fact is that in the derivation of proposition P_t the proposition P_{t+1} has been assumed to be valid. This seems to be no problem, given CKR — but proposition P_{t+1} has to be derived even if at stage $t+1$ deviations from the rational play at earlier stages have been observed, which contradicts CKR. The derivation of the equilibrium path necessarily involves considerations about what would be off the equilibrium path — but the latter must be void if CKR is valid, because no one ever deviates from optimal play. Thus "[c]onventional arguments on subgame–perfect equilibria require counterfactuals of the form: 'if a rational player made the following sequence of irrational moves, then ..." (Binmore [1987:198]) The players must be able to deal with deviations from rational play while they *know* at the same time that their opponent is rational and knows that they are rational themselves and so on.

The only possibility for sticking to CKR despite the fact that moves have been observed which should not have been observable given CKR is to suppose — at every stage — that history doesn't matter. Only then deviations from optimal play give no rise to doubts about one's opponent's rationality. Exactly this is ensured by the fiction of a 'trembling hand' (Selten [1975]). Deviations from the equilibrium path are due to uncorrelated random errors which are the reason for not *implementing* the optimal choice. The observation of bn has to be interpreted as a tremble: A would have played bc, if not an exogenous random error had pushed his hand towards playing bn. And as the random errors are uncorrelated, the probability of playing bn again does not depend on this observation.

But then the paradoxical result comes as no surprise: Strategic moves are impossible by definition, as long as every player has to interpret observable deviations as trembles and not as signals. A cannot — and he cannot *by definition* — have any influence on B's expectations by playing bn at early stages. If one applies the backward induction argument with reference to CKR, then B has to interpret this choice as a tremble without any informational value. Hence, we have an impossibility of strategic moves in games where CKR is assumed because the trembling hand–interpretation of deviations is necessary to employ CKR consistently[11].

One way to escape the trembling hand–assumption is to replace CKR with a system of iterated beliefs in the opponent's rationality which have to be updated after every observed move. Then B can, for example, predict A's choice of bn at stage n even if he has observed deviations at earlier stages, because for the

[11] More detailed elaborations of this argument with reference to the problems of counterfactual reasoning can be found in Binmore [1987,1988], Bicchieri [1988,1989], Bonanno [1989], Pettit and Sugden [1989], Sugden [1991,1992] or Koboldt [1993].

prediction he needs only to believe that A is rational. This belief can be held even if A plays bn at stage $t-1$, because for bc to be optimal at stage $t-1$ A has to be rational and to believe that B is rational and that B believes that A is rational. The choice of bn at stage $t-1$ can be explained by the fact that at least one of these beliefs does not exist, i.e. that A does not believe that B is rational or that A does not believe that B believes that A is rational or both.

It is important to note, however, that it is impossible to single out one and only one explanation for bn played by A at stage $t-1$. It might also be that A is not rational at all. Which explanation is employed by B is rather arbitrary. There exists one system of changing beliefs which generates the subgame perfect equlibrium sketched above — but nothing enables us to say that this one is the only system of updating beliefs that is compatible with or even required by rationality itself. B cannot discriminate rationally between the different explanations for the failure of A to choose bc at stage $t-1$. B knows that at least one of the preconditions for the choice of bn is not fulfilled, but he cannot say which one.

The extrapolation of future behavior from observed past behavior is a problem that cannot be solved by rational considerations — at least not within finite time, because any attempt to do so ends in infinite regression. Suppose that B observes the choice of bn at one stage. This could be due to the fact that A is not rational. If this is the case, B should expect A playing bn again and therefore play w. But perhaps that is exactly what A wanted him to do, so that the choice of bn is no signal of irrationality. The problem is one of self–reference, generating undecidable situations, i.e. paradoxa like the well–known Epimenides–paradoxon.

The situation becomes worse without dominant strategies, as one can see by analyzing the passive samaritan's dilemma in the way sketched above. In general, to squeeze all information out of the observation of past play one must speculate on the opponent's speculations on the effects of past play on future play. These effects depend on the opponent's speculations on one's own speculations on the opponent's speculations and so on ad infinitum.

Thus, rationality has necessarily to be incomplete if a decision has to be reached in finite time. The incompleteness indicates an exit criterion which allows a break from the infinite chain of reasoning. This exit criterion can be modelled as a so–called 'guessing algorithm' (cf. Binmore [1988]). The results produced by this algorithm may sometimes be wrong and decisions therefore be suboptimal. Nevertheless, the algorithm is necessary to produce a result at all. To calculate his first move, each player must have an a–priori probability distribution about possible types of opponents (characterized by their guessing algorithms). Then he can determine by a finite chain of backward reasoning the expected pay–offs for every possible move. During play the probability distribution has to be updated according to the information from observed moves (Binmore [1988], see also Kreps and Wilson [1982], Milgrom and Roberts [1982] and Kreps, Milgrom, Roberts and Wilson [1982]).

To be sure: this procedure can be avoided if CKR is assumed. But the price to be paid is the necessity of a trembling hand–explanation for deviations, which precludes any influence of past play on future play: history and experience don't matter. In exchange one gets a criterion for the selection of equilibria that promises uniqueness. The process described by Binmore [1988], on the other hand, can generate almost any sequence of moves as the equilibrium of a game.

Hence, the prescription given by Buchanan, i.e to act strategically instead of pragmatically, may work, but not necessarily. It won't work, however, if all agents are rational and know about their rationality. To the extent that this is the setting that seems to be appropriate for analysis, the social scientist can only move from diagnosis to despair: if a situation is recognized as a dilemma no possibility for self–medication exists. Becoming aware of the situation is useless as long as strategic moves are precluded by the assumption that deviations of observed behavior from the rational course of action have to be interpreted as trembles, not as signals.

3.2 The Promising Infinite Horizon

The devastating consequences the requirement of subgame perfection has for the samaritan seem to vanish if one removes the assumption of a known end period. Then there is no last stage at which expectations do not matter and from which the whole analysis can unfold. Additionally, the problem of influences of history on the expectations diminishes, because "*any* history of play will necessarily be *short* compared with the possible length of the game. (...) The assumption that a player will behave rationally in the future, in spite of having behaved irrationally in the past, therefore ceases to be quite so implausible." (Binmore [1992:356])

An infinite horizon does not necessarily require the interaction to go on forever. It is sufficient that the probability of one more stage is high enough at every stage. The easiest way to model this is to assume a constant probability p (i.e. independent of the number of stages yet played) for the game to continue. The probability that the game will last n periods is p^{n-1} and the probability that it will last forever is zero. Nevertheless, the time horizon for any player is infinite, as the number of stages yet played has no influence on the probability that the game will go on[12]. The probability of continuation can be interpreted as a discount factor[13] for future pay–offs which are discounted because they are uncertain. Too high a discount rate (or too low a probability) means that future benefits may not be high enough to compensate for a present utility loss. Then a rational samaritan can never "convince the potential parasite of his willingness to suffer utility loss in order to insure that the expected value of this

[12] This is to say that the probability for the game to last another n periods after having reached the nth stage equals the probability for the game to reach stage n.

[13] ... equivalent to a discount factor δ which can be derived from a discount rate r as $\delta \equiv 1/(1+r)$.

loss is effectively minimized." (p. 176) So let us assume a discount factor (or a continuation probability) high enough for a 'future' to exist.

As a strategy for a repeated game requires the player to specify an action for the first choice and a function which maps any possible history of the game onto possible action(s) for every stage of the game (cf. Binmore [1992:349 ff.], the set of possible strategies for infinite horizon games is infinite. Attention will be restricted to pure strategies which could be represented by finite automata[14].

A finite automaton can be described as a computing device which accepts an input and produces an output. We will consider so called Moore-machines[15]. These machines can be described by a finite set of internal states S, a finite set of possible inputs I, a finite set of possible outputs O and two functions, a transition function $f_T : I \times S \to S$ and an output function $f_O : S \to O$.

For our purpose the set of inputs contains the actions of one's opponent at one stage, the set of outputs contains the actions of oneself at one stage, and the set of internal states builds up the memory for a finite piece of history. The machine accepts the action of the opponent at stage $n-1$ as input and produces an action for the player at stage t as output.

It is convenient to represent different finite automata graphically with circles describing the internal states where letters within circles label the outputs corresponding to the state, and with labelled arrows which show the transition from one state to another induced by the respective inputs. An unlabelled arrow points to the initial state of the machine.

The choice of a finite automaton to play the sequence of stage games equals the choice of a strategy for the supergame. For our samaritan's dilemma stage game let us consider the following machines, playing the samaritan's (figure 3) and the parasite's (figure 4) part respectively.

The complexity of a machine can be measured by the number of its possible states. The machines pictured here have a maximum complexity of two. You may imagine or invent machines of greater complexity[16], but for simplicity let us concentrate on these eight machines.

Since the complexity of each two game–playing machines is finite the number of possible combinations of states the two machines might occupy is also finite. This fact, together with the determination of state transitions by the two functions f_T and f_O ensures that any two finite automata playing a repeated game will settle down to a cycle of finite lenght, in which an identical sequence of plays of the stage game is realized. One can show (cf. Binmore [1992:365]) that maximization of the sum of expected (discounted) pay–offs is equivalent to the maximization of the average pay–off per cycle, i.e. the sum of the pay–offs over all stages of the cycle divided by the length of the cycle.

[14] Mixed strategies make it difficult to specify the history of a play because any observable action has to be seen as an instance of a random variable.

[15] Cf. Niemeyer [1977:71 ff.], Binmore [1992:361 ff.] or Binmore and Samuelson [1992].

[16] Even if the requirement for each and every machine is a finite complexity, there is an infinite number of finite automata.

Figure 3: Game–playing samaritan machines

We would have to consider 16 possible pairings of machines, finding the cycle into which each pairing settles, calculating the average pay–off for each machine (which is the pay–off to the player who selected the machine) and looking for Nash–equilibria. Nash–equilibria are characterized by combinations of machines where each is a best response to the other measured in terms of average pay–offs.

These possible combinations are:

1. Punishing Samaritan vs. Punishing Parasite: Cycling through (bn,rw) $\Rightarrow$ average pay–off of 1 to samaritan, 1 to parasite.

2. Punishing Samaritan vs. Tat-for-Tit Parasite: Cycling through (bn,w) $\Rightarrow$ average pay–off of 2 to samaritan, 2 to parasite.

3. Punishing Samaritan vs. Tit-for-Tat Parasite: Cycling through (bc,w) $\Rightarrow$ average pay–off of 4 to samaritan, 3 to parasite.

4. Punishing Samaritan vs. Naive Parasite: Cycling through (bn,rw) $\Rightarrow$ average pay-off of 1 to samaritan, 1 to parasite.

Figure 4: Game–playing parasite machines

5. Tit-for-Tat Samaritan vs. Punishing Parasite: Cycling through $(bn,rw) \Rightarrow$ average pay–off of 1 to samaritan, 1 to parasite.

6. Tit-for-Tat Samaritan vs. Tat-for-Tit Parasite: Cycling through (bc,rw), (bn,rw), (bn,w), $(bc,w) \Rightarrow$ average pay–off of 2.5 to samaritan, 2.5 to parasite.

7. Tit-for-Tat Samaritan vs. Tit-for-Tat Parasite: Cycling through $(bc,w) \Rightarrow$ average pay–off of 4 to samaritan, 3 to parasite.

8. Tit-for-Tat Samaritan vs. Naive Parasite: Cycling through $(bn,rw) \Rightarrow$ average pay-off of 1 to samaritan, 1 to parasite.

9. Testing Samaritan vs. Punishing Parasite: Cycling through (bc,rw), $(bn,rw) \Rightarrow$ average pay–off of 2 to samaritan, 2.5 to parasite.

10. Testing Samaritan vs. Tat-for-Tit Parasite: Cycling through (bc,rw), (bn,rw), $(bc,w) \Rightarrow$ average pay–off of 2.67 to samaritan, 2.67 to parasite.

11. Testing Samaritan vs. Tit-for-Tat Parasite: Cycling through $(bc,w) \Rightarrow$ average pay–off of 4 to samaritan, 3 to parasite.

12. Testing Samaritan vs. Naive Parasite: Cycling through (bc,rw), (bn,rw) $\Rightarrow$ average pay-off of 2 to samaritan, 2.5 to parasite.

13. Naive Samaritan vs. Punishing Parasite: Cycling through (bc,rw) $\Rightarrow$ average pay–off of 3 to samaritan, 4 to parasite.

14. Naive Samaritan vs. Tat-for-Tit Parasite: Cycling through (bc,rw) $\Rightarrow$ average pay–off of 3 to samaritan, 4 to parasite.

15. Naive Samaritan vs. Tit-for-Tat Parasite: Cycling through (bc,w) $\Rightarrow$ average pay–off of 4 to samaritan, 3 to parasite.

16. Naive Samaritan vs. Naive Parasite: Cycling through (bc,rw) $\Rightarrow$ average pay-off of 3 to samaritan, 4 to parasite.

The results can be represented in the matrix of figure 5.

	Punishing Parasite	Tat-for-Tit Parasite	Tit-for-Tat Parasite	Naive Parasite
Punishing Samaritan	1 1	2 2	3 4	1 1
Tit-for-Tat Samaritan	1 1	2.5 2.5	3 4	1 1
Testing Samaritan	2.5 2	2.67 2.67	3 4	2.5 2
Naive Samaritan	4 3	4 3	3 4	4 3

Figure 5: Possible combinations of game playing machines

As one easily can see, there are six Nash–equilibria in pure strategies: the Tit-for-Tat Parasite is the best response to all but the Naive Samaritan while all types of samaritans are equally good best responses to the Tit-for-Tat Parasite. The other three equilibria are characterized by the Naive Samaritan as best response to any type of parasite, and all parasites but the Tit-for-Tat Parasite are equally good best responses to the Naive Samaritan.

The number of Nash–equilibria may be surprising — and if the restriction on four types of machines for each player is relaxed, one may expect the number of Nash–equilibria to increase. In fact, the number of Nash–equilibria approaches infinity, as the complexity of the machines increases.

But all Nash–equilibria[17] of the infinitely repeated game have one property: the average pay–offs for the cycles generated by equilibrium–strategies are dense in the cooperative pay–off region of the stage game limited by the minimax pay–offs. This intuitively plausible result is known as the 'folk–theorem'[18]. An illustration is given in figure 6: all Nash–equilibria of the infinitely repeated stage game generate average pay–offs for the finite cycles which lay in the shaded area of the cooperative pay–off region.

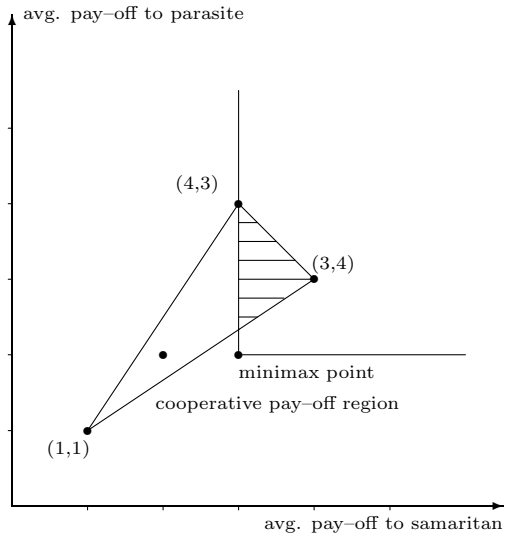

Figure 6: Nash–equilibrium outcomes for the active samaritan's dilemma

Up to now we have sidestepped the issue of credibility. Once a player has selected an automaton which plays the game on the player's behalf, no problem of credibility arises. The machines are programmed to generate a response to any stimulus — they are not rational in the sense of deciding the optimal, i.e. expected–utility maximizing, course of action at every stage. As long as a player can commit himself to the use of a strategy which can be represented by a finite automaton, all credibility problems are assumed away. But any such strategy

[17] This means: All possible Nash–equilibria without any restriction on the complexity of the automata involved.

[18] For a simple proof see Binmore [1992:369 ff.].

does not resemble rational behavior in the sense of opportunistic utility maximization. Rather it represents a form of boundedly rational behavior[19], which may be of advantage to the boundedly rational player: "[B]ounded rationality, in essence, empowers a meta–player to make *commitments* to stay within a certain strategy" (Binmore [1988:15]). Or, as Buchanan puts it, there are "genuine advantages to be gained by the samaritan from locking himself into a strategic behavior pattern" (p. 176).

The credibility problem, however, lures at the backdoor: if one cannot really delegate decision–making at every stage to an agent (here the machine) one must suspend one's own decision making capacity if one wants to follow the rule which is represented by the automaton's program. This is seemingly paradoxical. How can a rational agent decide not to decide rationally? The fact that rule–governed behavior may be advantageous to a possible rule–follower[20] is not a sufficient condition for the adoption of rule–following (cf Sugden [1991]). Buchanan is right in his claim that "having once adopted a rule, the samaritan *should not* be responsive to the particulars of situations that might arise." (p. 177) Players must *be able*, however, to resist the temptation of violating the rule if it seems better to do so.

If B expects his opponent not to be a game–playing machine but rather a rational player, the advantages from the infinite repetition seem to vanish. To be sure: if B expects the threat of eternal punishment by the choice of bn as represented by the Punishing samaritan he never will deviate from his choice of w. But is this threat credible if the opponent is not a programmed automaton which really has no other option than to punish in response to rw? The answer clearly is: no. If B deviates, the best A can expect to get if he implements his punishment schedule is a pay–off of 2 which is worse than the pay–off of 3 if he accomodates and play settles down at (bc,rw). There is no incentive to implement the threatened punishment when it would be the time to punish. Credibility could be restored if failure to punish would be punished. But this punishment in turn is only credible if the failure to punish the failure of punishment will be punished and so on.

The old question of 'who guards the guardians' is at the heart of the problem. And the answer seems to be an infinite chain of punishment for failure to punish failure to punish .... for failure to punish deviations from the cooperative sequence (cf. Binmore [1992:377 ff.]). Only an infinite time horizon allows for this answer to be implemented. The guardians may indeed guard themselves, as long as there is always the opportunity for additional punishment, i. e. the possibility of an infinite repetition. Therefore the fact that the folk theorem does also hold for subgame–perfect Nash equilibria comes at no surprise — as long as the time horizon is infinite.

[19] Marks [1992] employs finite automata to model boundedly rational behavior. This approach, however, is not without its critics (cf. Selten [1990]).

[20] Heiner [1983,1990], for example, shows why and when rule–following behavior is better for imperfect agents than flexible optimizing.

But rational players — as opposed to boundedly rational players — cannot escape the dilemma if no infinite future exists. They cannot surrender their decision making capacity unless this option is at hand. And as soon as the probability for continuation or the discount factor becomes too low, the credibility problem arises again and the whole series of threats unravels. This will be the case if one relaxes the assumption of a constant probability of continuation but instead assumes a probability decreasing with the increasing number of stages already played. Then the prospect for rational individuals seems to be not very promising...

Boundedly rational individuals, on the other side, can escape dilemmas. And perhaps man is better modelled as a boundedly rational decision maker than as a cognitive Goliath, who is beaten sometimes by those small–minded Davids who are unable to violate their personal norms of conduct when it would be better to do so. The power which lies in the ability to stick to rules cannot be overlooked. Frank [1988] gives impressive examples for the 'strategic role of the emotions'.

But if rule–governed behavior, which is either hard–wired in the inividuals or rooted in internalized norms of behavior, makes up an important part of interactions, then evolutionary concepts may be fruitfully employed. As Buchanan tries to trace back the dilemma to the development of modern societies, at the end a few remarks on this topic should be appropriate.

4 Which Way Now?

"A species that increasingly behaves, individually and collectively, so that it encourages more and more of its own members to live parasitically and/or deliberately to exploit its producers faces eventual self–destruction". (p. 185) Is this the future of an evolutionary process that led man out of the necessity to act strategically and thus trapped him by the temptation of short–term utility maximization?

Perhaps this view is overly pessimistic. If the samaritans are the prey and the parasites are the predators, one might expect a stable situation where every deviation is self–correcting: too many predators reduce the prey they live on, and subsequently their proportion decreases (and vice versa). The issue at hand is surely one of evolutionary stability. The problem can be analyzed by assuming a population consisting of finite automata — not necessarily of the simple type sketched above. Individuals here are seen as a "host for 'memes'. Dawkins uses the term meme to include rules-of-thumb, social norms, conventions or other more complex idea systems that a human being may use in translating a stimulus into a response. Evolution is seen as being responsible for a selection being made from the pool of possible memes. After evolution has operated, non–selected memes play a role in interpreting counterfactuals much like that played by trembles in traditional refinement theory. In brief, the non–selected

memes serve as 'explanations' for what would happen if selected memes were to deviate from equilibrium play." (Binmore and Samuelson [1992:284])

If one looks only for strategies which are, appropriately defined, evolutionarily stable, the number of possible solutions to the infinitely repeated game can be reduced. Evolutionary stability is like a refinement criterion for Nash–equilibria in settings where the selection of different automata is driven by the relative performance of one type as compared to all other types.

The theoretical setting that allows for this analysis is as interesting as it is complicated; too complicated to be reproduced here[21]. At the heart are the so called 'replicator dynamics' which model how the relative performance influences the changes in the population.

For our purposes it may suffice to look at the question of which Nash–equilibria, resulting from the interaction of rule–following agents playing the repeated samaritan's dilemma, will survive in the evolutionary struggle. For this purpose one doesn't even need to specify the dynamic aspects, one merely has to concentrate on the question of evolutionary stability of equilibria.

As Binmore and Samuelson [1992] have shown, in some plausible setting only utilitarian outcomes "are sustainable as evolutionary viable equilibria (...) The idea behind our argument is of ancient vintage. An initially nonutilitarian population is vulnerable to invasion by mutants who recognize each other by means of what Robson calls a 'secret handshake'. This private signal allows the mutants to form an insider group who cooperate among themselves but treat outsiders as outsiders treat each other. (...) Only utilitarian machines can be immune to such invasions and so they are the only possible candidates for evolutionary viability." (Binmore and Samuelson [1992:280])

The utilitarian outcome can be determined by maximizing the sum of the individual pay–offs[22]. In our graphic representation this means that only those Nash-equilibria are evolutionary viable which lie on the tangent to the cooperative pay–off region with the slope of -1 and the greatest possible distance to the origin. For our active samaritan's dilemma these are all points at the north-eastern frontier of the cooperative pay–off region in the diagramm of figure 6. This means that either of the Nash–equilibria resulting in pay–offs of (3,4) or (4,3) as well as any combination therefrom is evolutionary viable. There is at least a chance for the situation where the samaritans behave charitably *and* the parasites work to be an equilibrium which survives the evolutionary struggle.

For the passive samaritan's dilemma the situation looks worse. An analysis of figure 7 shows that the utilitarian outcome is given by (3,4). Thus, only equilibria in which the samaritan always behaves charitably and the parasite always refuses to work are evolutionary viable.

But these last results should not been taken too seriously: one reason is that

[21] The interested reader should refer to Binmore [1992, chapter 9] for an easily understandable introduction into the concepts of evolutionary game theory.

[22] This requires the pay–offs to be interpersonally comparable. But as in evolutionary settings the pay–offs are interpreted as fitness, this should not be seen as a problem.

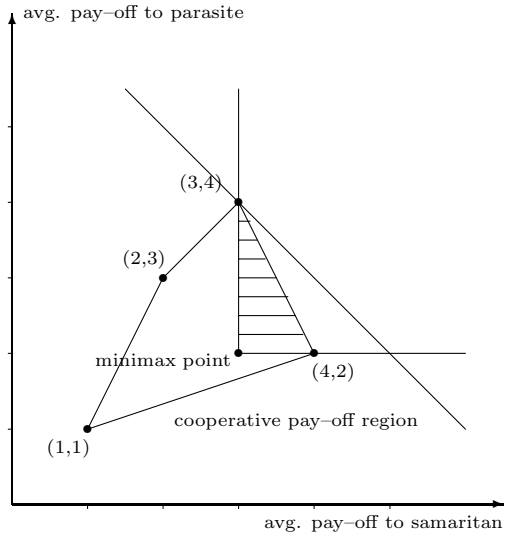

Figure 7: Nash–equilibrium outcomes for the passive samaritan's dilemma

even in an evolutionary viable equilibrium the samaritans will survive and will not be wiped out by the parasites. The second, more important reason is that the results are extremley sensitive with regard to the assumptions about the pay–offs[23]. If only the additional benefit to the parasite from being supported is less than the additional cost to the samaritan, the result will be the reverse.

And as long as the pay–offs are as stated, at least for a utilitarian, there seems to be nothing wrong with the results: if the parasite's gain outweighs the samaritan's loss, then the parasites should go on with the exploitation. But the coexistence of samaritans and parasites could then be regarded as a symbiotic relationship rather than as a dilemma.

References

Bicchieri, Cristina (1988): Strategic Behavior and Counterfactuals, Synthese Vol. 76.

[23]It should be obvious that *all* assumptions of the model have an influence on the results. But as Binmore and Samuelson [1992] are themselves "anxious that the paper not be seen as a stylized defense of utilitarianism valid for all societies without qualification" this point is stressed again, perhaps far beyond necessity.

Bicchieri, Cristina (1989): Self–Refuting Theories of Strategic Interaction: A Paradox of Common Knowledge, Erkenntnis Vol. 30.

Binmore, Ken (1987): Modelling Rational Players — Part I, Economics and Philosophy Vol. 3.

Binmore, Ken (1988): Modelling Rational Players — Part II, Economics and Philosophy Vol. 4.

Binmore, Ken (1992): Fun and Games, Boston.

Binmore, Ken and Samuelson, Larry (1992): Evolutionary Stability in Repeated Games Played by Finite Automata, Journal of Economic Theory Vol. 57.

Bonanno, Giacomo (1989): The Logic of Rational Play in Games of Perfect Information, Economics and Philosophy Vol. 7.

Buchanan, James M. (1977): The Samaritan's Dilemma, in: James Buchanan: Freedom in Constitutional Contract, Texas A&M–Press.

Dixit, Avinash and Nalebuff, Barry (1991): Thinking Strategically: The Competitive Edge in Business, Politics and Everyday Live, New York.

Frank, Robert H. (1988): Passions Within Reasons: The Strategic Role of the Emotions, New York and London.

Gilbert, Margaret (1989): Rationality and Salience, Philosophical Studies Vol. 57.

Heiner, Ronald (1983): The Origin of Predictable Behavior, American Economic Review Vol. 73.

Heiner, Ronald (1990): Rule–Governed Behavior in Evolution and in Human Society, Constitutional Political Economy Vol. 1.

Koboldt, Christian (1993): Zeitinkonsistenz, Teilspielperfektheit und Separabilitt: Vom mglichen Segen der "beschrnkten" Rationalitt, forthcoming in Homo Oeconomicus.

Kreps, David M. und Wilson, Robert (1982): Reputation and Imperfect Information, Journal of Economic Theory Vol. 27.

Kreps, David, Milgrom, Paul, Roberts, John and Wilson, Robert (1982): Rational Cooperation in the Finitely Repeated Prisoner's Dilemma, Journal of Economic Theory Vol. 27.

Marks, Robert (1992): Repeated Games and Finite Automata, in: John Creedy, Jeff Borland und Jrgen Eichberger (eds.): Recent Developments in Game Theory, Aldershot.

Milgrom, Paul und Roberts, John (1982): Predation, Reputation and Entry Deterrence, Journal of Economic Theory Vol. 27.

Niemeyer, Gerhard (1977): Kybernetische System– und Modelltheorie, Mnchen.

Pettit, Philip and Sugden, Robert (1989): The Backward Induction Paradox, The Journal of Philosophy Vol. 86.

Rasmusen, Eric (1989): Games and Information, Oxford.

Selten, Reinhard (1975): Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games, International Journal of Game Theory Vol. 4.

Selten, Reinhard (1978): The Chain Store Paradox, Theory and Decision Vol. 9.

Selten, Reinhard (1990): Bounded Rationality, Journal of Institutional and Theoretical Economics Vol. 146.

Sugden, Robert (1991): Rational Choice: A Survey of Contributions from Economics and Philosophy, The Economic Journal Vol. 101.

Sugden, Robert (1992): Inductive Reasoning in Repeated Games, in: Reinhard Selten (ed.): Essays in Honor of John C. Harsanyi, Berlin u.a.