



Bounded Rationality and Theory Absorption

Werner Güth

Max Planck Institute for Research into Economic Systems, Jena
(e-mail: gueth@mpiew-jena.mpg.de)

Hartmut Kliemt

Dept. of Philosophy, University of Duisburg-Essen, Duisburg, Germany
(e-mail: Hartmut.Kliemt@t-online.de)

Abstract In plausible theories of bounded rationality actors are not stimulus-response machines but human beings. As such they are guided by theories that predict the course of the world and prescribe how they should try to intervene in that course. Since boundedly rational human beings cannot only observe but can also modify their theories, in particular if they are not satisfied with the results, a self-application of concepts of boundedly rational behaviour to theory choice and an inquiry of theory absorption seems natural. The paper explores by means of specific examples some issues that are raised by combining the concept of satisficing behaviour with that of theory absorption.

JEL Classification A12, A13, B14, B52, C61

Keywords theory absorption, bounded rationality, satisficing behaviour, secrecy problem, equilibria

1. Introduction

Ever since concepts of bounded rationality have been introduced into economic theorizing efforts were made to ‘reduce’ the new approach to the old optimization under constraints approach. The argument was basically that there is a higher order optimization process of some sort or other going on behind the scenes. According to that view bounded rationality is not a new star under the sun of full rationality but merely a kind of moon that in the last resort gets whatever light it may shed on human behaviour from the source of full rationality. We think that this view is entirely mistaken.

To get them out of the way let us rehearse and reject some of the more conventional arguments in defense of optimization:

1. *Boundedly best replies can be and in very simple situations are as a rule optimal but that does not corroborate concepts of full rationality* If a decision problem is simple, even limited cognitive abilities will allow to assess what is optimal and what not. ‘Confirmation’ of the predictions derived from theories of perfect rationality in often artificially simple situations does not justify the rational choice approach in general. Since both, bounded and unbounded rationality imply the same predictions in simple cases evidence derived from these cases does not tell us anything about the relative merits of theories of bounded and full rationality respectively. Even worse, since theories of boundedly as opposed to those of full rationality always draw attention to the fact that real decision makers need to and tend to reduce the complexity of a decision situation to manageable proportions the focus on very simple decisions amounts to loading the dice in favor of theories of full rationality.

2. *Constrained optimization cannot account for what is achieved by bounded rationality* Optimization under (additional) constraints need not be easier than without. Imposing additional constraints may often make a problem more difficult to solve – especially when constraints make it necessary to compare boundary optima with interior ones. The reduction of complexity by constraints may render it more likely only in very special cases that optimality be reached. For instance if we focus on a repeated prisoner’s dilemma game and do not allow for recall at all then that constraint makes the choice simple since actors can only choose between cooperation and defection. But eliminating the shadow of the past entirely, the preceding argument from artificial simplicity applies.

3. *Aspiration formation and satisficing are elements of discrete optimization in disguise and therefore can be eliminated on a deeper level of analysis* This could be plausible only if we had some indications that on a deeper level some kind of optimization was going on. But there are no indications that some (conscious) optimization takes place in which setting aspiration levels and efforts at satisficing serve as instruments in the pursuit of optimality. To set goals whose achievement can be easily judged and related to a few relevant choice alternatives is what bounded rationality aspires to – full stop. Setting aspiration levels and satisficing behaviour with respect to them cannot be viewed as steps towards full optimization.

4. *Pure path dependence as, for instance, conceptualized in evolutionary theory or learning theory, e.g. reinforcement learning with low or (nearly) no cognitive demands may imply a kind of selective optimization but does neither involve rationality nor bounded rationality* Such approaches eliminate forward-looking selection of behaviour as made on the basis of a cognitive model of the situation altogether. They use selection via evolutionary competition (Darwinism) or via the law of effect. Although results might look as if rationally chosen there is no ‘rationality’ at all (Alchian 1950), for an early ‘as if’ justification of rationality).

Boundedly rational behaviour is a sub-species of *rational* behaviour. It is, first, forward looking though not perfectly so. Second, the boundedly rational actor is in general aware of the presence of other actors who are also forward looking. Third, the boundedly rational actor can try to put herself in other actors' shoes. And, fourth, the same faculty of the mind is also behind the ability to reflect on a situation, to manipulate or to modify mental models of a situation deliberately and to accept or to reject theories of how to behave and to expect others to behave. Bounded rationality does neither deny deliberation (the shadow of the future) nor path dependence (the shadow of the past). Based on past experiences, we usually try to develop a mental model of our decision environment and construct a few (in view of their past success) relevant decision alternatives. But on the basis of the mental model we do look into the future and are – within the bounds of our cognitive capacities – motivated by expected causal effects of present behaviour on future results as presently conceived.

That there is theorizing of a bounded rather than a perfect form, that humans do know that they know things and that others do so albeit within their cognitive limitations is undeniable. The 'reflexive aspects' of boundedly rational behaviour are the focus of the present 'reflections' on the absorbability of theories of bounded rationality. Here we distinguish between unilaterally, partially and fully absorbable theories. A theory of boundedly rational decision-making is *unilaterally absorbable* if a decision-maker: 1. considers only her – or himself as being in command of the theory and 2. after consequently following the theory, will be *satisfied* with the results of the theory's predictive (what the theory predicted as course of the world did occur to a sufficient extent) and prescriptive uses (the results of using the theory as guidance in determining what choices the actor should make were satisfactory). If a theory is unilaterally absorbed then for this actor there exists – other things being equal – no reason to change the theory or the behaviour since the actor is satisfied with the result. In its descriptive uses the theory predicts what happens 'in the world' independently of the choices of the actor. In its prescriptive uses it tells the actor how to act in ways that lead to satisfactory results. Obviously, in the prescriptive use a predictive component is involved in that the prescribed actions (the do and do nots of the theory) lead to satisfactory results. That the latter will happen is at least implicitly predicted whenever a prescriptive theory recommends an action such that a satisfactory result be reached.¹ In the other extreme, in inter-active decision-making of *n* actors, a theory is *fully absorbable* if the assumption that *all* *n* actors follow its advice would not violate assumptions of satisficing as made by the theory and thus

¹ There is a close relationship between prediction and prescription in that the latter rests on the former in a teleological conceptualization of action. Nevertheless the two remain conceptually distinct.

would not make revisions of the theory's advice recommendable for any of the n actors.

It would be of great interest to study concepts of theory absorption in interactive decision making that do not assume full but only partial absorption. But in view of the difficulties of this we will start to study the largely unexplored territory of boundedly rational theory absorption or the absorption process of boundedly rational theories by focusing on the extremes first. We begin with unilaterally absorbable theories of boundedly rational behaviour which we approach by means of the specific example of the so-called secretary problem (2.). In a next step we turn to interactive decision making and multilateral theory absorption among boundedly rational actors in such situations (3.). Finally some rather tentative conclusions of a more speculative nature are drawn (4).

2. Absorbable routine behaviour in the secretary problem

A non-inter-active decision problem where optimality is difficult to conceptualize is the so-called *secretary problem* (Bolle 1979). In its standard version a potential employer does not know the characteristics of the available secretaries. The employer would have to search for information by inviting secretaries, by speaking to them etc. but this is costly. Moreover, as it may be assumed here, secretaries who are not hired on the spot will be hired by somebody else before one can reconsider them and make another offer. It is in each and every case an 'all or nothing decision'. Therefore, though possible in principle, it is not advisable to screen the field of potential candidates completely. After complete screening one would be stuck with the last candidate. Since that candidate more likely than not will not be the best one, full search cannot be a good strategy for the secretary problem.

To represent the problem in a form that could also lend itself to experimentation imagine $N (\geq 2)$ cards (representing the 'secretaries' with different qualifications). The cards lie face down on the table. Each of the cards bears a different monetary value. In a first step we (re)shuffle the stack of the N cards such that all orderings are equally likely and form one pile of all the cards. Then in each of possibly N successive rounds $t=1, \dots, N$ the uppermost card is turned and shown to the decision-maker. The decision-maker has two options, either to accept the card or to reject it. If she accepts she earns the monetary value noted on the card and the search-process ends (the secretary is chosen). If she rejects the card on round t then the next card $t+1$ will be turned. If the final round $t=T$ is ever reached the monetary reward on this card is paid out without any further choice (the decision-maker is stuck with the last card corresponding to being stuck with the last secretary in the original problem).

A traditional rational choice approach would have to specify some prior beliefs about the possible values of the N cards and the distribution of those values among cards. In a next step, in such an approach, one would determine the optimal among the many possible stopping rules. If it were not so sad it would be quite amusing that some hard-nosed rational choice theorists would indeed suggest to survey the field of all possible stopping rules and to ask which of those, if any, would fit best with observable behaviour. Pointing to the routine with the best fit to their theoretically derived stopping rule alternatives they would then typically claim that they found an explanation for the observed behaviour. After all, one of the stopping rules fits the observational bill best. But even if the results of observed behaviour were as expected under assumed prior beliefs if one of the stopping rules would be applied this would tell us nothing about the real decision-making process. Even if the prediction were extremely good the claim that priors together with the stopping rule 'explain' the behavioural result is quite absurd. Quite to the contrary we would rather wonder how on earth real people could by what kinds of mechanisms have brought about results that are in conformity with optimization.

It is hardly conceivable that any human being ever would do such a thing as choosing an optimal stopping rule first and then apply it even when the choice is as important as choosing a secretary. The process of decision-making by boundedly rational actors would be rather different. A bounded rationality approach could (and would in all likelihood) specify the following typical elements:

1. a process of aspiration formation,
2. a search process in which an effort is made to satisfy the aspirations,
3. an adaptation process in which aspirations are, for instance, lowered towards the end of the search process if no satisfactory result has been found yet (when t is close to N).

More specifically boundedly rational actors seem to go through an initial phase $t = 1, \dots, n (< N)$ of n rounds of 'testing the waters'. Assume that the initial phase would be leading to n values $v_1, v_2, \dots, v_{n-1}, v_n$ of some variable measuring the degree in which some aim was achieved. These values can be used to determine (before round $n+1$)

1. an upper aspiration $\bar{V}_{n+1}(v_1, \dots, v_n)$ and
2. a lower aspiration $\underline{V}_{n+1}(v_1, \dots, v_n)$

If n is small as compared to N , e.g. $n = N/4$, it seems very plausible to assume

$$\bar{V}_{n+1}(v_1, \dots, v_n) = \max\{v_1, \dots, v_n\}$$

and

$$\underline{v}_{n+1}(v_1, \dots, v_n) = \sum_{i=1}^n v_i / n$$

From round $n+1$ on, until some end-phase beyond a threshold of, say, $t^* \geq N - N/4$ is reached, a boundedly rational actor might for example decide as follows on each round t :

1. If $v_t > \bar{v}_t(\cdot)$, card t is chosen yielding v_t ; stop!
2. If $\underline{v}_t \leq v_t$, and $t \geq t^*$, card t is chosen yielding v_t ; stop!
3. If $v_t \leq \bar{v}_t(\cdot)$, the upper and the lower aspirations might be adapted in some boundedly rational manner or other as long as $\underline{v}_{t+1} \leq \bar{v}_{t+1}$ can be fulfilled and provided that some adequacy conditions like the following are fulfilled:

$$\bar{v}_{t+1}(\bar{v}_t, v_t) \in [v_t, \bar{v}_t), \underline{v}_{t+1}(\underline{v}_t, v_t) \in [v_t, \underline{v}_t) \quad \text{if } \underline{v}_t > v_t$$

and

$$\bar{v}_{t+1}(\bar{v}_t, v_t) \in (v_t, \bar{v}_t], \underline{v}_{t+1}(\underline{v}_t, v_t) \in [\underline{v}_t, v_t] \quad \text{if } \underline{v}_t \leq v_t;$$

with functions $\bar{v}_t, \underline{v}_t$ such that $\underline{v}_{t+1} > \bar{v}_{t+1}$ is ruled out.

Assume now that somebody would suggest an intuitively plausible procedure π fulfilling the preceding requirements as a theory of boundedly rational behaviour. The theory would *predict* that rational actors show behaviour in conformity with the theory π . If beyond that it is claimed that the theory π is *unilaterally absorbable* then it is also predicted that, first, actors who have been informed about the theory's content will still behave in conformity with the theory and, second, that such actors will stick to its prescriptions.

If the card-staple game (as an experimentally tractable representation of the secretary problem) would be played repeatedly then the specific aspiration adaptation theory of boundedly rational behaviour proposed before would inform us that players will behave accordingly. We might not know the specific heuristic π that the players use. But if it is claimed that some specific heuristic π underlying a behavioural theory is in fact accepted as a standard of behaviour and that it is absorbable then it is implied that players behave in conformity *because* they follow the theory's prescriptions of how they should behave and that they in doing so do not have reason to be dissatisfied and to deviate from the theory and its underlying heuristic π .

Whether or not a theory is (in general) true, accepted and absorbable can be experimentally tested. Imagine in the simplest case that a player has to play the card staple game twice. In the first game it can be observed what the player as a

matter of fact does. The player is then informed about that theory and its prescriptions. Then the player has to play again. The theory is potentially accepted if the player's behaviour does not deviate from it on the second round. It should be noted, though, that it is hard to tell from observing overt behaviour whether a theory is accepted. Conformity with complicated theories may simply emerge because the theories are ignored rather than 'obeyed'. Actors act in conformity with the theory but not because of their knowledge of the theory and its prescriptions. So some additional testing and questioning on the second round may be necessary if one intends to apply the strong notion of absorbability according to which the theory must be among the reasons for action.

Besides the sampling in aspiration formation and the updating in aspiration adaptation other processes could be proposed as standards of boundedly rational behaviour. Some of the possible alternatives might indeed conform better with the facts. Again it could be tested what is unilaterally absorbable and what not by suggesting alternative aspiration formation and adaptation mechanisms or heuristics π to the boundedly rational players. Contrary to classical optimization based 'predictions' of outcomes there is no uniqueness implied if we go for boundedly rational theories. Still the criterion of unilateral absorbability is a possible criterion of theory selection. Only such theories of boundedly rational behaviour that can be accepted by the players and be maintained after they are informed about the theories will 'survive' that test. In the example of the card game aspiration adaptation can be such that boundedly rational actors could indeed follow at least those normative prescriptions that respect their cognitive limitations without having good reason to deviate. Of course, if there would be additional information and the like this would have to be factored in as well and might lead to absorbable processes other than the one sketched here.

We do not know which of the many possible boundedly rational decision making procedures will be applied. It seems clear, however, that for instance Bayesian updating is not among the plausible approaches. After all an absorbable theory that is really guiding behaviour must be accepted as a standard of behaviour and boundedly rational individuals will simply not accept a standard that – like Bayesian updating – is too complicated or goes against the grain of their intuitions. Going beyond unilateral theory absorption this should apply to inter-active decision-making of more than one actor as well.

3. Absorbable theories for inter-active situations

In a fully absorbable theory for inter-active situations – whether fully or boundedly rational – it must be possible that all $n \geq 2$ of a collectivity of n actors comply with the theory without thereby providing a good reason for any of the actors

to deviate from the theory. If all n use the same theory and use it to predict the behaviour of the other (boundedly) rational actors whom they assume to follow the prescriptive component of the theory – as they do themselves – then no individual player should have a reason to go against the prescriptions of the theory and thus a result akin to an equilibrium should have emerged. The latter does not show that the theory based on such assumptions applies to the real world. It is only assumed to apply so in theory. However, if one can show that the theory can be absorbed in the kind of equilibrium condition described before, then this at least demonstrates that the theory is ‘coherent’ in some minimal sense of that term. It indicates that minimum conditions for reaching an inter-personally sustainable ‘reflective equilibrium’ in theory formation are fulfilled.

A complete theory of boundedly rational behaviour in interactive situations would have to describe how actors are influenced by the theory itself and how they expect others to be so influenced etc. As of now it seems outrageously unrealistic to present such a theory. However, theories of fully rational behaviour do not fare better in that regard. It only seems so since they do away with rather than solve the problem by assuming ‘rational expectations’. Thereby rational actors expect the theory to be true and by behaving accordingly make it true. The theory of fully rational behaviour assumes itself to be fully absorbed in this sense. If we grant the same heroic premise of full theory absorption then the alleged advantages of theories of full rationality over those of bounded rationality vanish to a large extent. Solving the problem of predicting the behaviour of other actors (as influenced by the theory itself or rather the knowledge thereof) is as easy in theories of bounded rationality if we grant them the premise that boundedly rational actors assume other boundedly rational actors to behave according to the theory and themselves behave according to that theory.

If we have a fully absorbable theory, assume that it is in fact fully absorbed and known to be absorbed strategic uncertainty, i.e. the problem to predict others’ behaviour, is not an issue anymore. This applies regardless of whether we use a theory of full or of bounded rationality. In both, the case of full rationality and that of bounded rationality we then ‘solve’ the problem of strategic uncertainty ‘in theory’ by assuming it away ‘through theory absorption’. The common knowledge assumption involved may seem much more extreme, though, for a conception of bounded rationality than for a theory of full rationality. But this only shows how far removed from anything realistic conceptions of full rationality are. That they simply do not care whether or not real individuals could and would in fact apply the theory which allegedly explains their behaviour and is used by them to predict the behaviour of others is not a merit of such theories. It rather makes them useless if we are interested in the true behavioural laws guiding human behaviour. On the other hand, if we grant the premise of full absorption for theories of boundedly rational behaviour – an admittedly heroic

assumption – many of the niceties of theories of fully rational behaviour emerge. That they emerge without having to assume unlimited cognitive abilities of all actors shows that the much admired theoretical coherence of theories of full rationality can to some extent be had in theories of boundedly rational behaviour (though at the same price of assuming full theory absorption).

3.1 An example of full theory absorption with bounded rationality

Consider the following two-person game without an equilibrium in pure strategies (Holler and Illing 2003). Interests conflict but not completely so since there exists no positive linear transformations of the payoffs such that the transformed payoffs sum to zero.

	L	R
T	(4, 1)	(2, 3)
B	(1, 6)	(3, 2)

The unique mixed strategy equilibrium rendering both Row and Column indifferent between the choice of alternative T, B and L, R, respectively, requires that Row chooses T with $prob(T)=2/3$, and Column chooses L with $prob(L)=1/4$. Adopting these equilibrium strategies leads to a payoff expectation of $5/2$ for Row and of $8/3$ for Column.

A theory of full rationality that prescribes behaviour would have to be such that none of the players would have a reason to deviate from the theory's prescription under the assumption that the other player adopts the very same theory. It is common knowledge among the players that each player predicts other and chooses own behaviour on the basis of the prescriptions of the same theory of fully rational behaviour. In view of this, if players put themselves in the shoes of the other and think through what the other might do they immediately come to the conclusion that whenever one of their alternatives had a higher expectation given their prediction of other behaviour they should choose that alternative with probability one. Assuming for the sake of the argument that mixing is not only ignorance in the eye of the beholder (i.e. the co-player) the preceding argument indeed implies a behavioural probability of action such that the other is exactly indifferent between all her alternatives. Obviously only theories that dictate that strategies be chosen such that both players become *simultaneously* indifferent between alternatives can survive in reflective equilibrium of fully rational players.

The focus on absorbable theories also shows that the old query that making

the other player indifferent between alternatives there is no reason for the co-player anymore to mix her strategies is rather pointless. Of course, foreseeing that the co-player chooses according to the theory will make the player indifferent between his own choices. Then the player has no incentive actually to behave according to the theory's prescriptions and vice versa. However, both players know that only a theory that dictates mixing to both will not provide a reason to deviate to any. In that sense the theory of fully rational play is determined by the absorbability condition (regardless of the fact that after adopting the theory it is not strictly self-enforcing).

Now, let us consider maxmin strategies. The pure maxmin strategies would guarantee each player a payoff of 2. Assuming again that behavioural mixing is possible mixed maxmin strategies become an option. Now players are not made indifferent by the co-player but render themselves indifferent about what the other might choose by mixing their own choices appropriately. Whatever the other does by mixing according to a theory of maxmin mixing they would get their own maxmin expectation. Again that expectation is $5/2$ for Row and $8/3$ for Column. It emerges by fixing

$$\text{prob}(T) = 1/2 \text{ for Row}$$

$$\text{prob}(L) = 1/6 \text{ for Column.}$$

Since the maxmin strategy does not render the co-player indifferent the best reply to the maxmin strategy is, however, not the maxmin strategy but rather a pure strategy, namely *B* for Row and *L* for Column. If a theory would suggest maxmin behaviour to fully rational individuals then a reflective equilibrium could not be reached. The theory would not be absorbable among fully rational players. This problem would vanish, however, if best replies to given expectations could not or would not be observed or would remain unknown. Boundedly rational players might even understand that the expectation of maxmin behaviour by others could conceivably induce them to deviate but may still stick with the payoff expectation they can guarantee for themselves by maxmin behaviour. After all they are merely boundedly rational and if they are appropriately bounded in their opportunity seeking behaviour they will stop with maxmin.

According to the preceding discussion the chances that a theory can be fully absorbable and will in fact be fully absorbed seem better under assumptions of bounded rationality than under assumptions of full rationality. Considerations that start from the premise of full absorption of theories of bounded rationality are strikingly similar to equilibrium analyses for theories of full rationality (Güth 2000).

3.2 Equilibrium in full and bounded rationality

Let $\varphi_i(\cdot)$ denote the advice which recommends the strategy $\varphi_i(G_i) \in S_i$ to a player i with strategy set S_i in a strategic choice problem G_i . Since the mental model may differ between players $i \in \{1, 2, \dots, n\}$ we use the subscript i when representing i 's strategic choice problem G_i .

In a behavioural theory the description of a strategic encounter G_i may differ from an orthodox game theoretic description fundamentally. On the one hand, only aspirations will be specified – full preference orders over results and expected utilities representing them are lacking –, on the other hand, aspects like game frames – that would be left out in a full rationality approach – may be incorporated.

Let us denote by (G_i, φ_{-i}) player i 's choice problem which results when all other players $j (\neq i)$ follow the advice $\varphi_j(G_j) \in S_j$ as derived from the theory φ . For all $i \in \{1, 2, \dots, n\}$ the recommendation of play, $\varphi_i(\cdot)$, may be listed in one vector

$$\varphi(G) = (\varphi_1(G_1), \dots, \varphi_i(G_i), \dots, \varphi_n(G_n))$$

The theory φ is fully absorbable if for all players $i = 1, \dots, n$ we have

$$\varphi_i(G_i) = \varphi_i(G_i, \varphi_{-i}) \quad \text{or} \quad \varphi(G) = (\varphi_1(G_1, \varphi_{-1}), \dots, \varphi_n(G_n, \varphi_{-n})).$$

In other words, the theory $\varphi(\cdot)$ of game playing must 'survive' its general acceptance. 'Survival' implies two things here: On the one hand, *before* making choices the knowledge that all other individuals act according to the fully absorbable theory will not keep any of them from following the theory. On the other hand, *after* all choices have been made and results emerged nobody will have a reason to give the theory up.²

As is obvious from the preceding remarks, full absorbability of a theory is very closely related to the concept of an equilibrium. Equilibria, i.e. strategy vectors from which no player can gain by unilateral deviation, are absorbable predictions since they are characterized by optimality and true, respectively, rational expectations ((Aumann and Brandenburger 1995). An equilibrium is a fixed point of the best response mapping or to put it slightly otherwise, in equilibrium everybody has already given his best responses to the best responses of all others. Likewise, under full absorption of a theory, all have adopted the same theory and

² Of course, the non-deterrence and no-regret criteria of conventional accounts of equilibrium look at the same thing, equilibrium, merely through two different windows.

playing according to that theory none has reason to change plans or to be dissatisfied with the theory. As long as everybody is ascribing the theory to everybody else and behaves accordingly herself there is no reason to revise the predictive or prescriptive structure of the theory or to deviate from its recommendations.

Fully absorbable theories of bounded rationality in strategic encounters will differ from theories recommending equilibria only in a rather subtle way. The description of states of full absorption of theories of boundedly rational behaviour can be derived from descriptions of equilibria in theories of full rationality very easily. This is accomplished by substituting theories and assumptions of bounded rationality for the corresponding assumptions of fully rational optimization. If full absorption is assumed and assumed to be known then all individuals who command that knowledge and share that assumption know what to expect from their co-players. Given this knowledge they have no reason to behave otherwise than according to the prescriptions of the theory. This is exactly the same as in traditional full rationality approaches that assume rational expectations. Still, there is a difference between the full rationality and bounded rationality approaches since at least cognitive limitations may matter in the latter. We must make psychologically realistic assumptions about what rational actors can understand and what they can do on the basis of that understanding and this will restrict the scope of those theories that can be absorbed unilaterally or multilaterally. Nevertheless, if made, the assumption of full absorption of a theory of bounded rationality leads to results akin to equilibrium notions in theories of full rationality. To put it slightly otherwise, the equilibrium notion is, in a way, more fundamentally related to full absorption than to full (unilateral) rationality. Therefore the question naturally arises what might happen if full absorption would be given up in favor of partial absorption.

3.3 *An example of partial theory absorption*

A specific example may provide a first glimpse on the relationships between bounded and full rationality on the one hand and full and partial absorbability of theories on the other hand. Think of a group of n people who participate in a repeated collective good game of the voluntary contribution type. Each player can contribute one unit or not. It remains fully anonymous who contributes and who does not. But after each round of play all players are informed about the total number, k , $0 \leq k \leq n$, of contributions. Each individual has a dominant strategy of non-contribution. However, if everybody follows the dominant strategy the result is Pareto inferior.

We model the situation in the conventional stylized way as an n -person

prisoner's dilemma game that is characterized by the following utility functions $f_i(\cdot)$, $g_i(\cdot)$ describing the payoffs of i as a co-operator and as a defector, respectively, for any given number k , $n-1 \geq k \geq 0$, of *others* who co-operate. For each round of play we assume that

1. the functions $f_i(\cdot), g_i(\cdot)$ are weakly monotonic in the number of *other* individuals ($\neq i$) who co-operate,
2. for all individuals i and $k > 0$: $f_i(k-1) < g_i(k-1)$
3. $f_i(n-1) > g_i(0)$.

Here $f_i(k-1)$ stands for the payoff of a co-operator, i , if $k-1$ other individuals cooperate; while $g_i(k-1)$ indicates the payoff of the individual i who is defecting if $k-1$ other actors cooperate. The condition $f_i(k-1) < g_i(k-1)$ models that free-riding is always, i.e. for any number of *other* contributors, better than contributing. The condition $f_i(n-1) > g_i(0)$ indicates that the result of universal non-contribution is Pareto-inferior to universal contribution.

Assume that a theory has been proposed that suggests that each player should condition her own co-operation on the next round of play on whether or not at least, \underline{k} , $1 < \underline{k} < n$ contributions have been made on the previous round of play. Assume that the theory has been accepted by a subset $M_h \subseteq \{1, 2, 3, \dots, n\}$ with h actors. All who accept the theory act accordingly.

Consider a strict maximizer in this situation. A maximizer would have an incentive to co-operate iff he expected exactly $\underline{k}-1$ other actors to co-operate and $f_i(\underline{k}-1) > g_i(0)$. In this single instance his action would not only affect the outcome but also the willingness of the boundedly rational actors who accept the theory to co-operate. If besides the individuals $i \in M_h$ nobody co-operates (there are in particular no individuals who co-operate no matter what or without any theory) then the theory of boundedly rational behaviour can be absorbed if at least $h = \underline{k}$ individuals accept the theory.

Maximizers have good reason to deviate from theories that recommend co-operation for a specified number of other co-operators *unless exactly* \underline{k} individuals (including themselves and $\underline{k}-1$ others) cooperate. In a satisficing approach things are different, however. If \underline{k} or more individuals have been cooperating on the last round of play, everybody will remain satisfied. If there is a theory of boundedly rational behaviour that predicts that at least \underline{k} individuals will cooperate on the next round of play if they did so on the previous round of play because that theory recommends such satisficing behaviour then the process may be going on indefinitely regardless of the fact that more than \underline{k} individuals cooperate. That isolated deviations would be possible without violating the condition for cooperation does not matter to those who are satisficing. One should note also, that partial absorption of a theory is completely sufficient here. There

may be people around who act upon completely different theories and consequently adopt different strategies. There may be actors who are acting in a completely random way. Nevertheless the theory recommending cooperation on condition that some threshold requirement is met can be partially and also fully absorbed among boundedly rational actors. As long as actors stick with outcomes that are satisfactory and expect others to do the same any positive number k of individuals will do.

The previous kind of behaviour seems to be rather plausible from a common sense point of view. As long as people do not have a suspicion that others might bring cooperation levels down they may be willing to cooperate and may go on to do so voluntarily as long as results are satisfactory. There are also some relationships to classical philosophical theories of human behaviour. The notion that political order is possible only if as Marsilius had it 'sufficiently many, sufficiently influential' actors follow the prescriptions of certain theories comes to one's mind immediately. Likewise it should be noted that Thomas Hobbes was thinking in terms of assurance games rather than prisoner's dilemma games most of the time. We may reformulate the Hobbesian view as saying that as long as we expect sufficiently many individuals to cooperate we might be able to reach satisfactory results as long as we all follow theories of boundedly rational behaviour that prescribe appropriate actions. If people start to maximize locally the process might unravel, though. Interestingly enough it will also unravel if people feel a very strong resentment against free riding. Once people start to cultivate resentment against those who behave in ways deemed unjust they may become dissatisfied with results that would otherwise seem satisfactory (de Jasay 1995). Then what Hobbes called 'defensio' against the exploitation by others rather than greed induces people to behave as if maximizing.

More generally speaking, if the theater of social inter-action is dominated by individuals who are rational but only boundedly rational and thus try to meet aspirations rather than seeking the maximum that may be in for them, then prescriptive and predictive theories of action may be quite robustly absorbed from some threshold on. But the lack of assurance that results will persist may be very dangerous even among boundedly rational individuals who managed to reach satisfactory results. The theory must be absorbed by sufficiently many sufficiently influential individuals who must remain under its spell and individuals must be assured that sufficiently many will remain so. As long as that is in fact the case the theories that accomplish this may in a way be self-supporting or self-fulfilling as would be typical in conventional equilibrium approaches based on full rationality as well. However, giving up the assumption of unlimited cognitive abilities, theory absorption cannot anymore imply conventional common knowledge assumptions.

3.4 *Bounded rationality and a-symmetry*

In a bounded rationality framework, knowing the general procedures of decision making which others apply does, of course, not imply knowing others' behaviour. Procedural concepts of rationality as those of bounded rationality may in particular leave open others' characteristics like their motives, financial endowments etc. A boundedly rational actor cannot derive by logical inference how other rational actors will decide. He does know, however, how such idiosyncratic characteristics as others may have can influence behaviour in general. He can know something about the procedures they apply; i.e. the boundedly rational actor can know the general characteristics of processes of boundedly rational decision making. He definitely knows that the others are at best boundedly rational actors. This may give him some weak indications of what the world might be like and how others might adapt in particular if combined with introspection. In that regard the assumption of symmetry between self and others may be helpful in deriving conclusions about adaptive behaviour of others.

There may, however, be situations where a-symmetry is a more realistic assumption. A brighter individual may know how to play optimally whereas less clever ones do not (a few classic board games, much simpler than chess, can even be solved!). The less clever cannot predict what the more clever one will do even when being aware of his superiority. Less clever decision makers cannot emulate what the more clever ones are going to think. But they may be clever enough to expect in such situations to be exploited by those who are brighter and avoid interacting with them. Such situations are, however, rare. Usually problem solving has many faces. And analytic skills as such may lead one astray. For instance in the first ultimatum experiment (Güth, et al. 1982), the few 'rational proposers' were typically students of operations research and mathematics who were clever in a way that violated Axelrod's maxim 'don't try to be too clever'. The clever were clever but at the same time much too naïve to be successful. Because of their failure to understand what motivates responder behaviour they were lost and lost out financially.

4. **Some tentative conclusions and speculations**

4.1 *Non-strategic procedures again*

In section 2 above we have seen that for specific decision tasks like the secretary problem, one may have an idea about the qualitative aspects of the adequate decision routine. But even for such special tasks and even when accepting this type

of decision routine, there exist many possible decision rules differing, for instance, in the length of the experimentation phase and the formation and adaptation of higher and lower aspirations. Since the secretary problem is just one type of decision task from a large universe of such problems it is exceedingly hard to characterize solutions in general terms. This illustrates the enormous challenge that we necessarily face when trying to define boundedly rational replies to given expectations. All we can offer are some rather speculative considerations. These ‘conclusions’ are not straightforwardly implied by the preceding discussion but suggest themselves in view what has been said before.

In the secretary problem after an experimentation phase with, for instance, $n = 10$ observations it is rather unlikely that all the 10 different values v_1 to v_{10} will be used as aspiration levels in the later satisficing phase. Rather one will try to form just two, as we have done in section 2 above, or three aspiration levels which in the latter case could be described as a good, bad or intermediate success. Determining the probabilities of achieving the aspiration levels requires much too complex probabilistic considerations. If, for instance, the higher aspiration $\bar{v}_{n+1}(v_1, \dots, v_n)$ is the best of the first n draws, one may simply judge the probability of finding a candidate better than $\bar{v}_{n+1}(v_1, \dots, v_n)$ as $(N - n)/N$. Thus a larger n will tend to increase the higher aspiration but decrease the ‘likelihood’ of its achievement.

It is still an open question how such rather anecdotal plausibility considerations will translate into more general insights about boundedly rational decision making. Here are some tentative views on the matter:

1. *Learning from analogous experiences* For the sake of specificity let us rely again on the secretary problem. If in previous secretary problems an experimentation phase with $n/N = 1/4$ yielded on average better results than other quotas n/N like $n/N = 1/2$ or $n/N \ll 1/4$, the analogy of secretary problems with each other justifies to rely on $n/N = 1/4$ also in an upcoming task. Note that such learning is different from pure path dependence because it is based on a cognitive assessment of the structural/qualitative analogy of former and actual decision problems and relies on adjusting the former experiences (with different N -parameter) to the present one (rather than just repeating former n -choices). The difficulty in specifying how to learn from analogous experiences is to assess the qualitative and sometimes even quantitative similarity of decision tasks.

2. *Setting priorities* Instead of choosing a point on a continuous trade-off curve one often will view a certain goal in a multi-objective choice problem as more decisive, at least for a certain range. For a family with children it may, for instance, matter most of all to live in a safe neighborhood as long as this is affordable. An example of choosing according to priorities is the ‘Take the best!’ – heuristic (Gigerenzer 2000; Gigerenzer and Todd 1999). Note again that decisions according to such simplified procedures are neither arbitrary nor un-

reflected. The underlying quasi-lexicographic ordering of value dimensions will as a rule result from a serious cognitive effort. The effort is guided by the desire to eliminate the necessity of trade-off considerations and to avoid cognitive dissonance (Festinger 1957). Asking what, for the situation at hand, matters most in general is a kind of individual constitutional decision that reduces the effort required in finding specific decisions.

3. *Mental modeling* Since life (of homo sapiens) is much too complex to be captured in full by tractable models a reduction of complexity is required. In response to that requirement we all became experts in mental modeling in the sense of capturing the crucial aspects of a real world situation by a much simpler mental model. Using the terminology of science it may finally assume the form of a theoretical model. Let us illustrate this again by an example: When discussing how taxes affect (un)employment, a complete model would have to specify how the incentives when hiring an employee are influenced not only by the taxes themselves but also by the way in which the government uses the tax revenues and the demand and supply effects implied by such government spending. In essence a full analysis of all this would mean to analyze a general equilibrium model with all its possible feedback effects and circular dependencies. This is clearly beyond what non-experts can and will do. Actually the simple arguments used in political debates illustrate that people argue more in linear ways rather than by appealing to circular reasoning (as has been shown in Reinhard Selten's work on 'qualitative reasoning', see Selten (2004)). Trade unions, for instance, usually claim that more taxes and thereby higher government expenditures create more demand and thus increase employment. Liberals mostly downplay such effects and focus only on the disincentives of higher taxes. Of course, such attitudes may be self-serving at root. But we typically find the same line of linear argumentation also in non-interested parties, e.g. when journalists discuss economic policy (Selten (2004) provides an example). 'Linear reasoning' is chiefly driven by cognitive requirements rather than interests and the proclivity to restrict ourselves to it even against interests seems rather strong.

4. *Aspiration adjustment and satisficing* Illustrates how boundedly rational decision makers can adjust to past experiences when deliberating their choices. Assume that a decision maker is not the only one who confronts a secretary problem repeatedly but is surrounded by many others whose practices, e.g. the relations n/N , he can observe. If others usually appear to be more lucky, he can react by forming a more modest higher aspiration $\bar{v}_{n+1}(v_1, \dots, v_n)$ out of n previously observed evaluations v_1 to v_n or by decreasing n so that (see the point above) achievement of $\bar{v}_{n+1}(v_1, \dots, v_n)$ appears more likely. Similarly, one will react to own experiences, e.g. by increasing, resp. decreasing n when nearly always achieving, resp. missing the higher aspiration. Generally, one will not form aspirations whose success is almost sure or is hardly possible.

4.2 *Absorbability again*

It is an interesting issue how the preceding would affect views on absorbability. Clearly if individual cognitive processes across the board would comply with our description then absorbable theories of boundedly rational behaviour would have to be of the same kind. The information of such theories would have to be processed in boundedly rational ways by the boundedly rational addressees of normative prescriptions as well as descriptive predictions of such theories. Advice which relies on effects of institutional or discretionary policy changes may prove to be wrong when given on the basis of a model of perfect rationality and when bounded and perfect rationality imply different behavioural reactions. Moreover, like the economic actors themselves the policy makers are at most boundedly rational. Explaining to boundedly rational policy makers on the basis of boundedly rational behavioural assumptions why and how certain measures may (or may not) work will render policy advice more acceptable than conventional advice based on welfare maximization.

According to classical perfect rationality approaches an absorbed theory of decision behaviour implies true or so-called ‘rational’ expectations. Since such expectations together with optimality define already the equilibrium of non-cooperative game theory (Nash, 1951), it became clear that the main problem of absorbable bounded rationality is to define a boundedly rational reply to given expectations. What this means is that strategic uncertainty (the problems resulting from not being sure what others will choose) is reduced to the extent that the theory is definite. A definite theory makes definite predictions of behaviour, possibly on the basis of specific and definite advice, once such a theory is absorbed. Others derive their choices in the same boundedly rational ways as I do etc. This is not a false consensus (Engelmann and Strobel 2000). I do not project my behaviour or my concerns onto others but can very much accept that they have different concerns and will therefore behave differently. What I know, however, is that others suffer from the same cognitive limitations as I do. Moreover others will act upon the same theory of boundedly rational behaviour as absorbed by them in their boundedly rational ways.

However, among boundedly rational individuals theory absorption will hardly ever lead to definite results as in theories of perfectly rational strategic interaction. There can be differences in the individual capabilities of problem solving (precluded by symmetry assumptions in theories of perfect rationality). I may be aware that others are more clever than I am although they are far from being perfectly rational. Such superiority may allow them to exploit me like an outsider would be exploited by insiders. My likely boundedly rational reply to such risks of exploitation by experts may be not to interact with them similar to the no trade – or no betting results of the rational choice approach (see for instance the

classical market for lemons argument in Akerlof (1984)). Still, if I have to interact with them I will have to cope with the problem of predicting behaviour of individuals who are asymmetrically superior to myself and therefore hardly predictable for me. Moreover, since as boundedly rational actors we may have only a common – and rather simple – procedural theory of how actors will decide we just do not know enough to predict the specific result of decision processes of others. In a way, we could say we know the *form* but *not the content* of the process. We do not know what they know idiosyncratically but we know how they process their information in principle.

All we can hope to offer are a few ‘do-nots’ and a few ‘may-helps’. Bolder attempts would have to combine the constructive ideas algorithmically (see, for instance, the framework suggested in Güth (2000)). But, in our view, such attempts should be guided by empirical results, e.g. by stylized facts from game playing experiments. Introspection may help but also may lead us astray, e.g. in the sense of a false consensus effect when neglecting heterogeneity in problem solving. Moreover, all the wonderful a priori assumptions of symmetry of rationality, rational expectations and the like that made life easier for theories of perfect rationality should be cast out as lacking any empirical justification – except for the most simple toy games.

Acknowledgments

We would like to express our gratitude to an anonymous referee whose helpful suggestions improved the paper. Of course, the conventional disclaimer applies.

References

- Akerlof, G.A. (1984). *An Economic Theorist's Book of Tales*. Cambridge: Cambridge University Press.
- Alchian, A.A. 1950. ‘Uncertainty, Evolution, and Economic Theory’. *Journal of Political Economy*, 58, pp. 211–221.
- Aumann, R., and Brandenburger, A. 1995. ‘Epistemic Conditions for Nash Equilibrium’. *Econometrica*, 63, pp. 1161–1180.
- Bolle, F. (1979). *Das Problem des optimalen Stoppens: Modelle und Experimente*. Frankfurt: Peter Lang.
- de Jasay, A. (1995). *Social Contract – Free Ride*. Oxford: Oxford University Press.
- Engelmann, D., and Strobel, M. 2000. ‘The False Consensus Effect Disappears if Representative Information and Monetary Incentives Are Given’. *Experimental Economics*, 3, pp. 241–260.

- Festinger, L. (1957). *Theory of Cognitive Dissonance*. Evanston (Ill.): Roar.
- Gigerenzer, G. (2000). *Adaptive Thinking: Rationality in the Real World*. New York: Oxford University Press.
- Gigerenzer, G., and Todd, P.M. (1999). *Simple Hueristics that Make us Smart*. New York: Oxford.
- Güth, W. 2000. 'Boundedly Rational Decision Emergence – A General Perspective and some Selective Illustrations'. *Journal of Economic Psychology*, 21, pp. 433–458.
- Güth, W., Schmittberger, R., and Schwarze, B. 1982. 'An Experimental Analysis of Ultimatum Bargaining'. *Journal of Economic Behavior and Organization*, 3, pp. 367–388.
- Holler, M.J., and Illing, G. (2003). *Einführung in die Spieltheorie*. Berlin Heidelberg New York: Springer.
- Selten, R. (2004). 'Boundedly Rational Qualitative Reasoning on Comparative Statics'. In: Huck, S. (ed.) *Advances in Understanding Strategic Behaviour*. New York: Palgrave.