The rationality of rational fools —*

The role of commitments, persons and agents in rational choice modeling

Werner Güth, Strategic Interaction Group, Max Planck Institute Jena and Hartmut Kliemt , Philosophy Department, University of Duisburg-Essen, Campus Duisburg

Abstract (95<100 Words): Subjective payoffs that represent given preferences "all things considered" together with strictly uncommitted opportunity taking cannot account for the behavior of personal actors. It is shown how agent based approaches can explicitly capture internal commitments of persons while sticking to conventional utility cum probability representations of desires and beliefs. However, if rational choice modeling is taken to this extreme, conventional analyses in terms of reasoning become implausible since sub-personal agents are not persons endowed with higher cognitive faculties. Starting from preference representations without looking into the black box of mental processes will hinder theoretical progress.

1. Introduction

Driving non-co-operative game theoretic modeling to its extreme subsequently we do not intend to deny that forms of rationality other than those based on fully opportunistic choice making *exist*. But we insist that *conceptually* opportunistic rationality should be clearly distinguished from other forms of rationality. It is illegitimate to deviate from the very precise assumptions of opportunistic rationality (under conditions of common knowledge) without acknowledging the deviation. One cannot have it both ways, on the one hand, use assumptions other than those characterizing opportunistic rationality (e.g. by smuggling in some kind of commitment into the concept of rationality itself) and, on the other hand, claim that one is still dealing with the same animal of opportunistic rationality as before.

Those who are in disagreement with non-co-operative game theoretic modeling as based on opportunistic rationality can legitimately suggest competing forms of modeling. But it is illegitimate not to acknowledge that they are doing something other than conventional non-co-operative game modeling. At the same time it is again legitimate to classify non-opportunistic forms of rationality as "bounded". But it needs to be done without implying any judgment to the effect that "boundedly rational" is inferior to "opportunistically rational" choice making in some sense or other. Bounded rationality is not less rational than – at least not in any intelligible pre-theoretic sense of that term – but different from opportunistic rationality.

Distinguishing between rational choice theory (RCT) that contains substantive assumptions about rational choice makers – about who is a player and what is on that player's mind – and rational choice modeling (RCM) that does not we will focus primarily on the latter here. In a move towards "picoeconomics" (see (George Ainslee, 1992)) we show that all sorts of personal commitments can be modeled explicitly as part of the rules of the game. Since non-co-operative game theory as opposed to (partly) co-operative game theoretic modeling is characterized by the background assumption that all aspects that are beyond the causal influence of choices of agents in plays of the game are explicitly captured as rules of the game the main part of the paper demonstrates how far

* Kliemt is most grateful to the Center for Study of Public Choice and his friends there for providing a pleasant, inspiring and interesting environment when working on the paper. We have profited greatly from comments by Geoffrey Brennan, Erik Davis, Dan Hausman, Ron Heiner, David Levy, Alex Tabarrok and Bruno Verbeek and even more so from extended written comments by Susanne Hahn and Bernd Lahno. The conventional disclaimer applies.

non-co-operative game theoretic modeling can conceivably be pushed (2.)[1]. However, eliminating personal rational actors from the account and substituting them by "rational fools" leaves practically no room for the conventional eductive (i.e. classical) interpretation of game theory in terms of reasoning. So there is a fundamental tension between the explicitness assumption of non-co-operative game theory and the classical eductive way of interpreting non-co-operative game theoretic reasoning as something that can be on the players' minds (3.).

[1] There may be, and presumably should be a revival of partially co-operative game theoretic modeling that leaves certain rules implicit to choice making itself or to the concept of a player rather than making them explicit as part of the rules of the game (which, of course, include the preferences). But there is no justification for confusing non-co-operative with co-operative modeling and, for that matter, opportunism with non-opportunism.

2. Substantive and formal issues of rational choice modeling

Relying on the distinction between what is and what is not subject to the causal influence of decision makers we can broadly say that "the rules *of* the game" comprise everything that is beyond the causal influence of choice making within a given game while the choice making itself concerns the causal influence as can be exerted by opportunistically rational choice making *within* plays of the game. One might note explicitly in passing that the material payoffs, the preferences of actors (or the utility functions representing those preferences), the knowledge conditions, the probabilities representing beliefs etc. are all represented as part of the rules of the game – and so are commitments that are center stage of our concerns here.

2.1. Two basic forms of personal commitment

Commitment power shows up in one of two ways *in* the rules of the game: payoff modification or modification of the move structure of games (elimination or addition of moves). In the most simple and pure cases commitment manifests itself either through the choice of one sub-game from a class of sub-games with identical game form but different payoffs or through the choice of one of a class of sub-games that all can be derived by removing some of the moves from one "master game form".[2] Take the example of Ulysses. If he employs some sailors to crack the whip should he make the slightest move towards going overboard when hearing the Sirens, he is using the first commitment technique. If Ulysses orders to be bound to the mast he uses the second technique since the option of jumping in is cut off the game tree so to say (at least as long as Ulysses is not cut off "the tree"). In the first case, Ulysses in a move preceding the sub-game chooses between one sub-game with the original and one with modified payoffs, in the second case he is choosing between a sub-game based on a full tree and one of which at least one branch is cut off.

To illustrate, consider commitment options for simple trust games as depicted in the two games depicted in figure 1. In both games the upper payoff is that of player A and the lower one that of B. The basic trust game is the sub-game ensuing in both games of figure 1 after B chose in an initial move the "do not commit" alternative. In the trust (sub-)game, if the second mover B is to move at all his preferences are such that he should choose to move down, D_B. The first mover, A, foreseeing that should choose down, D_A, as well. Among opportunistically rational actors the obvious solution of the sub-game therefore is (D_A, D_B). Though both actors could have been better off had they both chosen to "continue" according to C_A, C_B, in this sub-game, the Pareto superior result corresponding to the strategy combination (C_A, C_B) is closed off by opportunistic rationality itself.

With an additional commitment option for player B a Pareto superior result could be reached in ways compatible with opportunistically rational choice making. This would be the case if the player moving second in the original trust sub-game could, by a prior commitment move, make moving down less attractive for himself than to continue (corresponding to Ulysses engaging the sailors to crack the whip) and could perfectly inform the other player about that move. This first possibility is depicted in the upper game of figure 1 as "relative commitment". By payoff modification in the sub-game reached after an initial choice to commit by B it is less attractive for B to choose D_B than C_B. Player A who knows that B has chosen to become relatively committed foresees this and therefore a play of (C_A, C_B) emerges once the sub-game after an initial "commitment" is reached. Since the solution of the sub-game after B chooses to commit is superior to the solution of the original trust

[2] There are other possibilities but we confine ourselves to these paradigm cases here since they suffice to make all crucial points about commitment in RCM.

sub-game also in B's terms, B obviously should choose to commit and the Pareto superior result would be reached due to the presence of relative commitment options.

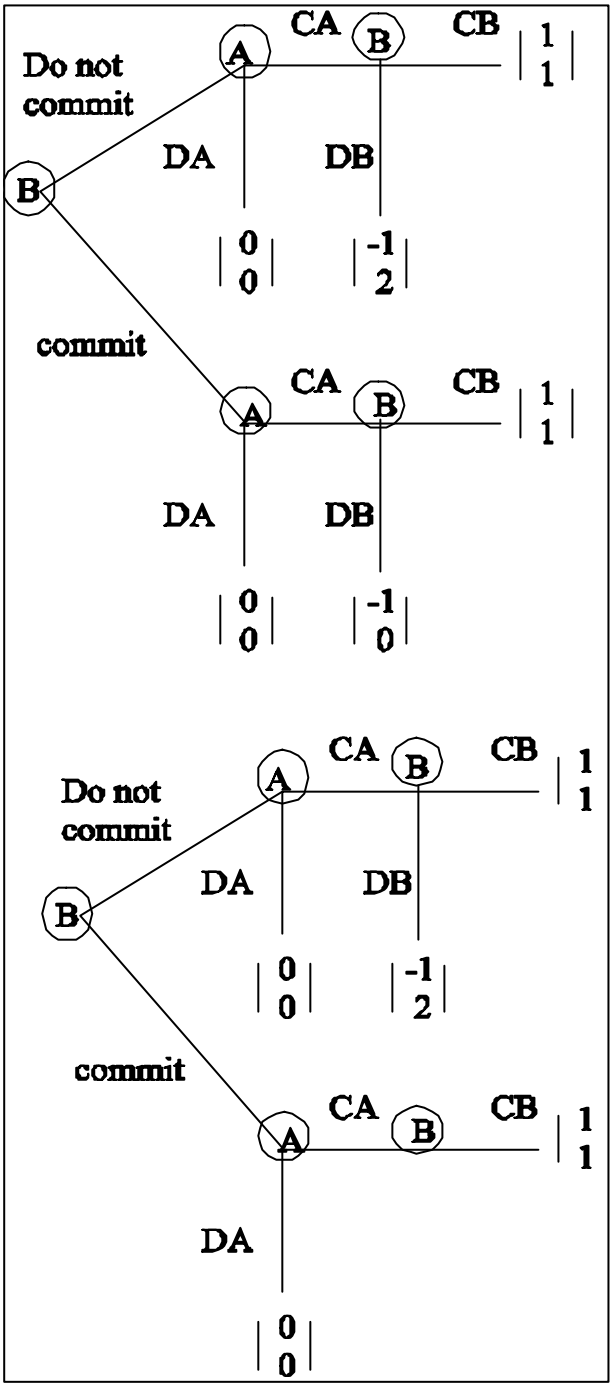

Figure 1 Relative and absolute commitments

The Pareto superior result could be reached in opportunistically rational ways also if by a prior move player B of the original trust sub-game could become absolutely committed (corresponding to Ulysses being bound to the mast) and could perfectly inform player A about that commitment. This second possibility is depicted as the second game tree in figure 1. Both trees include the original or basic trust game as reached if the actor does not and the modified game if he does commit by his initial move.

"In a strictly noncooperative game the players do not have any means of cooperation or coordination which are not explicitly modeled..." ((Reinhard Selten, 1975), sec. 2). All the rules of the game – everything that is beyond the opportunistic choices of the actors at any stage of a game – are explicitly modeled. In particular, if players become committed to certain courses of action this should show up in the rules of the game as something that is beyond further direct choices of the actor (what else could commitment mean?). From an analytical or modeling point of view this explicitness requirement rather than any assumption about the preferences and interests involved distinguishes non-co-operative rational choice modeling from co-operative rational choice modeling.

The explicitness condition is the reason why non-co-operative game theory may justly be regarded as more fundamental than (partly) co-operative modeling efforts. For instance, processes of coalition formation that are assumed to be operative behind the scenes in many co-operative game theoretic approaches, can always be modeled explicitly. What it means for several individual actors to act as a "corporate actor" can be resolved into inter-individual relations in a game internal to the "corporate actor" (or to a coalition). In this game all the internal commitments of the corporate actor are explicitly modeled as rules of the game.[3] For instance, if the corporate actor decides according to simple majority rule internally and then acts accordingly in its outside relations the voting game itself can be explicitly modeled as a way to "predict" the actions of the corporate actor in outside relations.

For a specific illustration (see for this and more generally on the logic of this strategy of committing by going collective (H. Geoffrey Brennan and Hartmut Kliemt, 1994)) imagine individuals with commonly known finite life spans of five periods. Assume they are of ages 1, 2, 3, 4, 5 respectively. Imagine that they participate as a team as player in a game created by repeating a simple trust game indefinitely. After each round of play the oldest individual exits and a new youngest member is accepted on the team such that the age distribution in the team remains the same. On each round of play players on the team decide as a team by simple majority whether or not they are going to exploit a trusting first mover. As a group they are committed to accepting the majority as binding for each of them (leaving open for the time being how this commitment to the majority rule and group decisions emerges and is maintained). Clearly, in such a situation the preference of the oldest individual on the team to exploit on his last round of play can conceivably be outvoted if the individuals in the team do not have too high discount rates and if continued co-operation is sufficiently rewarding to compensate a single deviation. The end game effect that would apply to each individual can therefore conceivably be avoided for the corporate actor (under appropriate conditions).

2.2. Modeling inner commitments explicitly

The logic of becoming committed by "going collective" is not restricted to personal players. It can be applied on the sub-personal level as well. To see that in somewhat more detail consider the following example of "take it or leave it" which contains both relative and absolute commitments as possibilities.

B A	+	-
+	1,0	-2,-2
-	0,1	0,1

Figure 2 Take it or leave it in strategic form

[3] Including for instance also joint randomization etc.

It is obvious that take it or leave it in strategic form has two equilibria in pure strategies, namely (+, +) and (-, -). However, if we go on to its extensive form it becomes immediately clear also that only one of the two pure strategy equilibria is plausible in terms of opportunistically rational choice making.

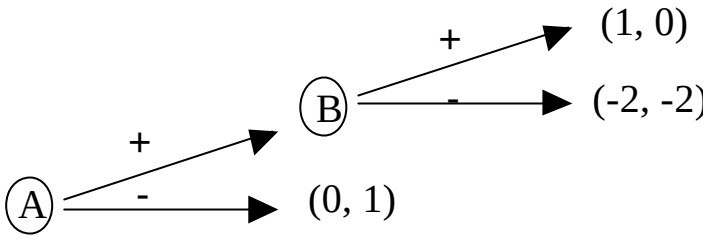

Figure 3 Take it or leave it in extensive form

Let us assume now that the second moving player commands both relative and absolute internal commitment power. As a personal actor he may for instance be intrinsically motivated to execute threats as uttered in pre-play communication. Like a corporate actor the personal player possibly may be dissolved into several agents who play a *game internal to the personal actor*.[4] To say that this is conceivable does not amount to claiming that it must be the case. As there may be ships with and without a mast to which Ulysses could be bound there may or may not be commitment power internal to a personal player. But if there is such commitment power then it should and – as we claim – can be modeled explicitly in non-cooperative game models. After such efforts a game like "take it or leave it" with explicitly modeled commitment options internal to the second moving personal player might emerge. For the sake of specificity we split the first moving player, A, of the original game into three agents (corresponding to the three possible decisions to be taken in the sub-games of the enlarged game) and the second moving player into four agents (corresponding to the four possible decisions of the personal player B in the new sub-games). The "splitting" yields two "classes", A and B, of agents representing the two personal players A=(A_{ii}, A_{iv}, A_{vi}) and B=(B_i, B_{iii}, B_v, B_{vii}). With payoff vectors representing the payoffs according to (B_i, A_{ii}, B_{iii}, A_{iv}, B_v, A_{vi}, B_{vii}) we arrive at the "modified take it or leave it" game as depicted in the next figure:

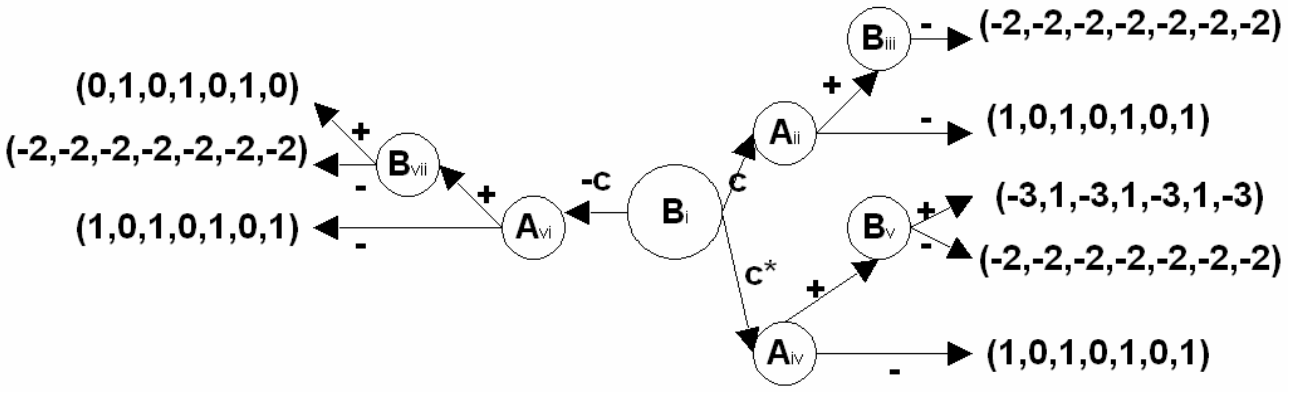

Figure 4 Modified take it or leave it game

[4] Alluding again to George Ainslee's work we might speak of pico-game-modeling here.

The nodes corresponding to the seven agents are numbered by Roman numerals. At the first node, agent i, of player B must decide whether to play the old (sub-)game played by agents A_{vi}, B_{vii}. This old sub-game is reached when no commitment is made by choosing –c. Making an absolute, c, or relative, c*, commitment two modified sub-games are reached. The first agent of B has an incentive to choose either c or c* since in both cases the backward induction solution of the sub-game would yield 1 for that agent (as well as all other agents of the personal player). For instance, if the agent B_v of B would come to move then she would use "-" since -3<-2. But then the agent A_{iv} of A would have to opt for "-" as well since 1>-2. Likewise we would see that the sub-game that could be reached by the choice of the absolute commitment option c yields the corresponding result.

We have assumed here that all agents of a personal player would evaluate end-results in the same way and at each node on a path would be motivated only according to preferences over the whole path. The assumption that a *personal* player at different decision nodes is endowed with "identical" preferences is thereby explicitly modeled. It is an empirical assumption about the decision making process of a personal player which may or may not be true.[5] RCM does not rule out in principle that the utilities could be different for different agents of the same personal player at different locations of the tree. Since in the "agent form" representation of the game a separate agent for each information set is included this is obviously possible. By virtue of the agent form RCM allows for an explicit modeling of commitments, changing preferences etc. of a personal player. At the same time it does not imply the strange view that persons of the real world are split into as many real players. Relying on the agent form representation does not force us to accept an empirical assumption about multiple or split selves in a real world meaning of those terms (see also (George Ainslee, 1992, Jon Elster, 1987)). Quite to the contrary, in the commonsense meaning of "player" there is only one "personal player" still. Only for theoretical purposes of making explicit all assumptions about the personal identity, the internal commitment technology and about the decision making process of players each player is split into several choice making agents.[6]

At each information set of the tree, we must know the value order relevant for *the decision at that node*. The lists containing one payoff for each agent of each player, at the ends of the game tree show how each agent of each player is affected by any decision (and in particular as seen from the point of view of the node where she is to move).[7] Those game models in which personal players who make several decisions are represented by one agent, implicitly make an empirical assumption which may be true or false. We have to read that assumption as the requirement that evaluations or preferences of a personal player remain constant independently of how the game is played. In those cases in which the assumption applies not much is lost by the simplification. However, there is a tendency to forget that an empirical assumption has been made.

Using the agent form, rational choice modeling, RCM, can take into account deviations from the specific theory of personal identity requiring that all agents of a personal player always have identical evaluations of all results of the process of choice making. Vice versa, if we assume all evaluations to coincide in this sense by not splitting the personal choice maker into agents then we

[5] In fact the same results of analyzing the game would emerge if the agents would have identical preferences over the plays of the game still possible after some stage and an initial agent's decision node have been reached.

[6] More traditionally speaking, by the agent form homo oeconomicus is dissolved into (a team of) homunculi oeconomici.

[7] We could go even further here and modify information conditions such as to include agent forms without perfect recall as well as games with incomplete information (see **Harsanyi, John C.** "Games with Incomplete Information Played by Bayesian Players." *Management Science*, 1967-8, *14*, pp. 159-82, 320-34, 486-502.). In the latter case, as a matter of fact, only one of the types of an agent exists when the game is actually played. The multiplicity of types is a consequence of closing the game informationally by a fictitious initial random move. As such the types are assumed to be on the players' minds when analyzing the game.

implicitly assume that all agents (or at least all that have to make decisions still) have the same evaluations independently of the path through a tree.

Nothing in the rules of non-co-operative game modeling precludes that we dissolve personal actors into agents. Analyzing the interaction situations fully into what, respectively, is beyond and what is subject to causal influence in separate acts of choice it seems rather natural to identify each instance of choice making with a separate agent. For each instance of choice we have a separate utility function representing the preferences of the agent who is active in that particular instance of choice making. It should be noted that this view is almost implied, at least very strongly suggested by the modern interpretation of the utility function as representative of preferences "all things considered".[8] For if utility represents the *preferences at a specific node all things considered* then the assumption of having *everything* considered at *that* node separates *that* decision from decisions at other nodes. *This applies even if decisions at different nodes are taken by the same person.* All that is relevant at any node is included in the evaluation at that node and only that.

As long as utility was a quality in itself – for example measuring levels of well-being – it could figure amongst the reasons for preferring one option over another one. As such it could conflict with other reasons for positioning options in a hierarchy of relatively better or worse options. But according to the modern concept of utility an option is never preferred because it has higher utility. The higher utility is assigned to the option because it is preferred (for what reasons ever). In forming the preferences over options all the reasons (whether that be material interests in well-being, aspirations to reach pecuniary advantages, some high ideals etc.) are taken into account and the utility function "afterwards" represents these preferences *all things considered*.

What appears as a kind of (rational) foolishness of opportunistically rational actors is a direct consequence of relying on "representative utility all things considered". Whatever can make human action intelligible and reasonable in a substantial sense is concealed from our views by representing preferences through utility. At least if it is correctly interpreted utility just indicates the fact that an option has a certain relative rank in a set of options after all the reasoning and in particular the "weighing of goods" has taken place.[9] What shows up at the end nodes of a game tree is a ranking information without substantial content, a pure ranking or a measure without dimensionality (though "cardinality" as needed to deal with risk applies). The ranking may or may not be reasonable as evaluated in substantial terms and therefore behind "the veil of the utility representation" the wise can look exactly like the fool (on how that relates to the concept of revealed preference including very telling citations of the original behaviorist interpretation of the concept by e.g. Hicks and Little, see (Amartya K. Sen, 1973/1982)).

As has often been stated the only remaining substantial assumption is "consistency". Rationality as consistency requires that evaluations must be such that we avoid to become victims of a money pump or a Dutch book. Otherwise utility is a stenographic device for presenting orders (over lotteries) as formed after taking into account all dimensions of value (whatever they may be). As such it gives us hardly any information about why and how the inclusion of these dimensions of value has affected the ordering that is represented. Neither does it – beyond the ranking itself – inform us about intentions "behind" utility functions.

---

[8] Daniel Hausman is particularly willing to take the "all things considered" concept as seriously as it deserves; see yet unpublished on this, **Hausman, Daniel M.** "Sympathy, Commitment and Preference," University of Wisconsin-Madison, 2004.

[9] Of course, we intend here to allude to **Broome, John.** *Weighing Goods. Equality, Uncertainty and Time*. Oxford: Basil Blackwell, 1991.

2.3. "Signaling" and executing intentions in plays of games

Nobody intends to maximize a utility function. Intending other things and having reasons other than utility per se the actor behaves *as if* maximizing utility. Starting from representative utilities sets economics free of the necessity to consider motives and intentions explicitly (they have all been considered in forming preferences as represented by the utility index). Economists are led to endorse the illusion that (cognitive) psychology is irrelevant for economics since everything that this science could contribute is "considered" in the utilities. Whatever it is that motivates humans and what they intend to do they behave *as if* they are maximizing their utility function. With the utility function in hand the economist seems to have all he needs to know about choice makers, their motives and intentions. However, though the elimination of human motives and intentions is at root of the specific style of modern rational choice modeling many economists "want to have it both ways". They try to bring in motives and intentions again. Again efforts to modify game theoretic rational choice modeling can serve as an instructive example here. In game theory intentions could conceivably show up in moves during a play of the game or in the strategies of players of which such moves are a part. We think that in both cases "commitment" power is smuggled in. To see how let us discuss the two closely related possibilities in turn.

2.3.1. Forward induction and signaling of plans

In game theory there is a long tradition of assuming "omnipotent players" (see (John von Neumann and Oskar Morgenstern, 1944) and (Elon Kohlberg and Jean-Francois Mertens, 1986)). "Forward induction" becomes viable among such actors. By their prior choices – that form first steps of a strategy that potentially contains later ones – such actors can allegedly signal how they intend to go on in a later choice making.[10] Consider the following game

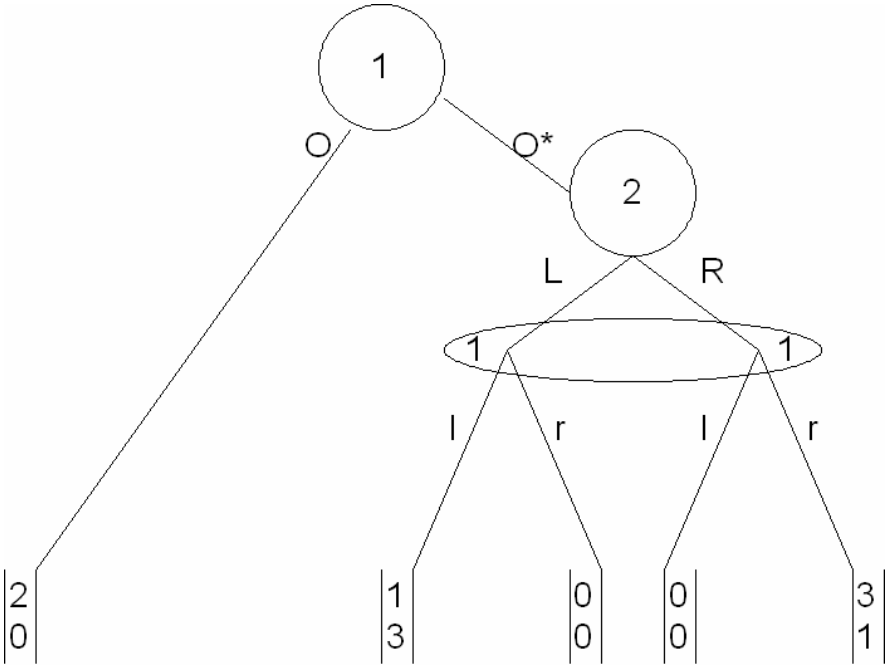

Figure 5 An embedded battle of the sexes

[10] This is what is driving also **Verbeek, Bruno.** "The Feasibility of Rational Self-Commitment," Leiden, 2004.

In the game of figure 5 there are two pure strategy sub-game perfect equilibria, namely ((O, l), L) and ((O*, r), R). The sub-game starting at the single instance of choice of player 2 contains two sub-game perfect pure strategy equilibria: (l, L) and (r, R). By choosing O* the choice maker 1 indicates that he intends to go for the second sub-game perfect pure strategy equilibrium. According to the standard argument the choice of O* would not make sense unless player 1 would have the firm intention to choose r along with the expectation that player 2 would understand this and act accordingly by in fact choosing R. Therefore, according to those accepting forward induction the equilibrium selection problem for the "battle of the sexes" sub-game that is starting with the decision node of player 2 is solved. If equilibrium selection is impossible in the standard battle of the sexes game (see (Robert Sugden, 1991)) then, if we accept that there is a solution in the larger game, it seems to follow that the past matters. For, if the game were not embedded in a larger game tree no such selection could be made. According to adherents of forward induction, after O* a selection can be made for a good reason. Moreover, since in the sub-game no violation of such principles like backward induction is committed the selected equilibrium solution of the sub-game is fully in line with forward looking rational choice in that sub-game.

We do not deny that the preceding argument if made in terms of objective rather than subjective payoffs is psychologically compelling. But note that in standard game theory such an argument is *not* made in objective payoff terms. It is formulated in payoffs which represent preferences that include *all* considerations relevant with respect to the future. If the future is indeed all that matters and if we take the notion of a saturated utility function seriously then the sub-game that starts with the decision of player 2 must have a solution independently of any preceding history of the play. Whether that game would be analyzed all on itself or as embedded in a game as in figure 5 should, once the sub-game is reached, not influence the choice making of players who choose in view of the causal consequences of their acts in that sub-game. Therefore the presence of the option external to the sub-game ensuing after O* should not have any influence on the solution behavior in that sub-game.

Still there is another argument according to which the strategy (O*, l) is dominated by both (O, l) and (O, r) and should therefore be eliminated. So, if move O* is observed and if a rational player will not choose dominated strategies then this observation shows that player 1 must have adopted (O*, r). But this is true only if the player is omnipotent in the sense of being able to choose whole strategies. As we will illustrate in some more detail now, the latter is contrary to basic assumptions of opportunistically rational choice making (or, for that matter, incompatible with modeling choice behavior in such terms).

2.3.2. Choosing strategies as actions and as plans

Strategies can be "chosen" only as plans *for* the game (being complete in specifying a move for each contingency that might conceivably arise in the game). Strategies are *not* choices that can be made *in* the game. If intentions are ascribed to another player by means of strategies then it must be possible to choose strategies in a stronger sense than forming a *plan for* choices to be made. It must be possible to choose strategies "*in*" the game. Note that in that case there must exist corresponding moves in the game (after all, by assumption, in a non-co-operative model all possible moves show up explicitly). Such moves show up as "commitment options" in the tree. The language of RCM as such does not commit us to any substantial assumption about whether or not such commitment options exist. Yet, according to the explicitness condition the ways to signal intentions in the game must be modeled explicitly if they exist. A personal player may or may not have these capacities. But if he can commit to a conditional strategy and thereby signal his intentions then we should model this option explicitly as an additional choice to be made in the game and that is the end of it.

– Again a specific example may be helpful. Consider what may be called a "sequential PD" with perfect information:

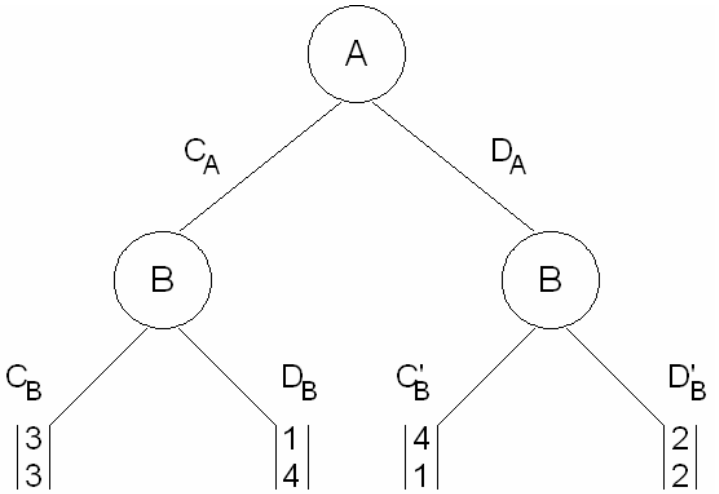

Figure 6 sequential PD

Strategies are plans that specify an action for each contingency that might arise in any play of the game. If an actor in the role of a second mover in a sequential PD has the option of actually *choosing* the behavioral program for each contingency before the game is played the following agent normal form bi-matrix game is an adequate representation of the strategic situation. It contains four rather than two choice options for the second mover.

Actor B Actor A	C_B/C_A C'_B/D_A	C_B/C_A D'_B/D_A	D_B/C_A C'_B/D_A	D_B/C_A D'_B/D_A
C_A	3, 3	3, 3	1, 4	1, 4
D_A	4, 1	2, 2	4, 1	2, 2

Figure 7 Strategic form of sequential PD

This strategic form of the sequential PD with perfect information is categorically different from the conventional tabular representation of the original PD in strategic form with ordinal payoffs as preference representations. It shows clearly how easily we are led astray by commonly used phrases like "choosing a strategy". Strictly speaking strategies are plans. We can choose to form one *plan* rather than another *for* a game. But we cannot choose the execution of the plan in one single act *in* the game. This is not among the options of the original sequential PD as presented in the extensive form. To choose a strategy in one act rather than merely planning to make several choices consecutively presupposes that there are "higher order" options to *make* such choices. Vice versa, the strategy as plan states how B intends to react to *both* C_A and D_A. It is, however, impossible to actually react to both whereas a reaction *function* that contains pre-programmed responses to both can be selected only if the option to make that choice (the choice of a reaction function or behavioral disposition in fact) exists.

The extensive game representation of the preceding strategic form representation would be the following one (leaving out the "primes" distinguishing equivalent moves C_B and C'_B etc.):

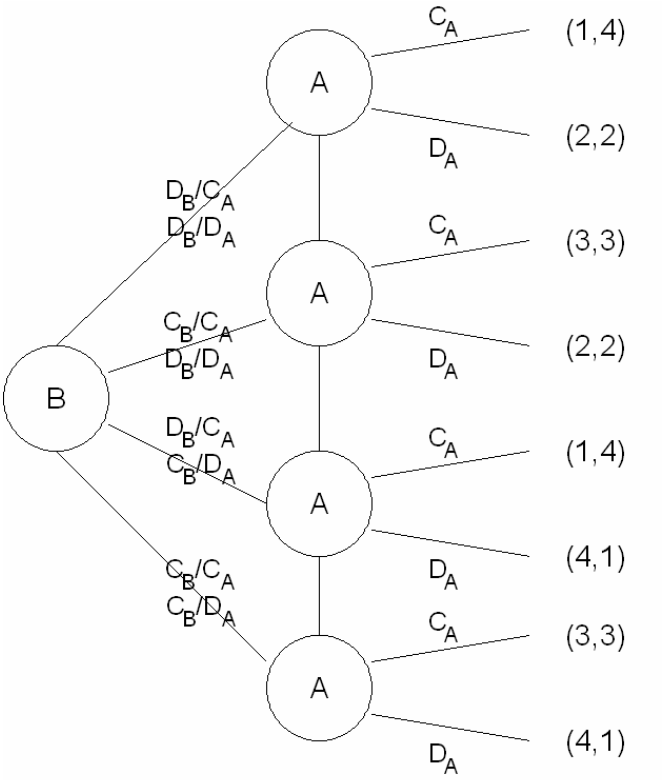

Figure 8 Standard PD with power to commit to strategies as programs

Due to the presence of the commitment option the relevant choices of B are made before A makes his. In the PD situation itself B does not make any choice anymore. After the information about A's choice transpires, B's programmed response follows.

In the preceding figure player B cannot inform player A bout his commitment choices. All decision nodes of player A are included in one information set. If contrary to that A's information partition would consist of four separate singleton sets (i.e., if in the preceding graph the connecting lines between A's nodes would be removed), then B could make his strategy choices not only beforehand but A would be perfectly informed about them. Under such conditions of perfect information the commitment power of player B becomes sufficient to solve the problem of avoiding the Pareto dominated result and the strategy combination $(C_A, (C_B/C_A \& D_B/D_A)$ would be chosen.[11]

It may well be that individuals command the faculty to modify their future behavior by simply planning and intending to do certain things in the future. It may also be that conventional rational choice modeling does not provide the best or most adequate tools to model such phenomena. But if we use the signs and the semantics that rational choice modeling supplies we should stick to the rules that go along with that "language game". If we do so we must model commitments as part of the rules of the game explicitly. Therefore, if "commitments to plans" affect the future then this must be modeled as a causal effect.[12]

[11] The case of the original prisoner's dilemma with imperfect information would be as follows: If actors – before that game is played – can choose a behavioral disposition to cooperate then the Pareto dominated result can be avoided by rational choice makers if the commitment to cooperate can be chosen such that it will be "binding" if and only if the other one is committed likewise. If players cannot make their commitments contingent on each other the PD problem will not go away but rather resurface on the commitment stage.

[12] Obviously we could rely on additional tools like creating a type distribution by forming a plan. The plan would affect the future by creating a set of more or less committed or uncommitted agents with different utility functions one of which in a fictitious random move would be chosen. A prior agent conceiving a plan could then causally affect the type composition and probability distribution over the set. – It seems unnecessary to go into these complications since the

Payoffs are construed for players who do intend to discriminate in their view of the world between what is and what is not subject to their causal influence at each instance of choice. So if game theoretic logic ever does apply then it should apply also with respect to the rest of the game tree. If it is assumed that players by reaching nodes that should not be reached, signal an intention of further play what can that mean if the payoffs at any future node express the preferences operative at that node? In a theory that is based on forward looking opportunistic choice according to given preferences (operative at the very instance of choice making), signaling an intention requires that players had the option of choosing a strategic commitment. But once that option is explicitly modeled the game tree is changed either in the ways of Figure 8 or by introducing additional players in ways akin to Figure 4.[13]

Analogous arguments quite straightforwardly apply to the notorious backward induction problem. If we strictly stick to the requirement that for the rational choice maker only the expected future matters, that his expectations are fully represented in his subjective probabilities/utilities, and that all influences on the future must be causal and as such be explicitly modeled backward induction seems rather obviously implied. For the sake of completeness, let us turn to the somewhat confusing debate even though from our point of view everything necessary has been said already.[14]

2.4. Having it backward

According to a common view violations of backward induction on preceding stages of games should lead to the conclusion that backward induction should not be used for analyzing the remainder of the game. The standard argument being, if you reach a node at all that you should not have reached according to backward induction arguments how can you then still assume backward induction to apply from where you are to the future?

In discussing backward induction it seems reasonable to ask how a game tree is to be *derived* in the first place. If it was written down correctly in agent form under the explicitness condition then the preferences and branches of the tree at each decision node must have been fixed *factoring in* the fact that there was a preceding history that may have influenced the evaluation of the future at *that* node. The payoffs for the future as fixed in *construing* the tree take into account that the node can be reached only by a play of the game that violates backward induction. If that violation does not affect the rules of the game, in particular the payoffs, at later rounds of play *all things considered* then the payoffs are what they are "all things considered".

Assumptions of forward looking rational choice seem to *separate* any informationally closed sub-game of a larger game from its preceding history. Edward McClennen who is presumably the philosopher who has been objecting to backward induction most strongly over the years (see (Edward F. McClennen, 1990)) is clearly right in insisting that a requirement of "separability" is crucial for backward induction to emerge. He rejects separability as intuitively implausible. But separability is almost implied in RCT if we take seriously the view that rationality requires the intention to distinguish between what is and what is not a causal effect of an act (along with the desire to improve one's situation). But it is certainly implied by the way RCM is set up. Without

---

basic approach seems obvious and an analysis of such complications, though showing how flexible RCM as a "language game" is, would not add much of an insight.

[13] As indicated in the previous footnote such players could be selected either by nature in a fictitious random move or strategically by a choice to modify payoffs.

[14] It is basically said once the modern utility notion as representing all things that are relevant at a decision node is taken seriously.

separability the whole enterprise of game theoretic analysis by "separate" parts of a game would become non-viable.

To see more specifically what is involved here let us introduce a somewhat coarsened form of the separability principle (see (Edward F. McClennen, 1998), sec 3 in particular pp. 21-22):

> Let T be a decision or game tree. Consider a node s in a singleton information set and an informationally closed sub-tree T/s that emerges after all nodes preceding node s are cut off while s together with all its subsequent nodes remain as a complete sub-tree. Then, according to separability, the solution remains the same:
>
> a. if T/s is considered separately as a game
> b. if s is reached in T when playing the larger game.

Backward induction appears so paradoxical because we intuitively tend to forget that according to the semantic rules of RCM utility is representing preferences all things considered (including causal influences on preferences by going through a game tree along a specific path). We always tend to relate to utilities as if they were objective payoffs and to treat them as if they were reasons for action. But once we argue that we *prefer* a higher utility to a lower one we apply the concept of preference to the representation of preferences and thereby are already using categories akin to meta-preferences. We treat utility as if it were money and thus as something we have preferences about. In the next step we are deceived into thinking that we prefer an act "because" it leads to higher utility. But utility represents preference and is not itself an object of preference. If we make it an object of preference then we are reasoning about the tree as if we preferred to get the more preferred results in the tree. According to the foes of backward induction we signal this alleged preference for higher utilities by violating backward induction. But this is exactly what an opportunist can *not* signal under conditions of perfect information about utility functions that represent preferences *all things considered.*

We cannot have it both ways, treat utility as representing preferences after *all* reasons for the ranking of alternatives have been considered – thereby avoiding "psychology" and the necessity to speak of human motives – and then treat utility as merely *one* consideration among *all* others. We cannot meaningfully assume that preferences describe completely what is preferred at each decision node of the tree – because in writing down the tree we have fully analyzed what it is like to be at *that* node *all* things including the path leading to the node considered – and then renege on this assumption if a decision node is in fact reached. Either our modeling tool is RCM or not. If it is RCM then backward induction seems to apply for the almost trivial reason that the game tree must be set up in ways that factor in the past into the utilities relevant at each node. Then the causal influence of the past is "in" the payoffs representing preferences relevant for forward looking choice making at any node of the tree.

Let us apply this to the famous (or perhaps notorious) centipede game which seems to be the standard counter example used by foes of backward induction. Though still unnecessarily complicated from our point of view, let us discuss the argument with the help of a specific decipede game as presented in the next figure.[15]

[15] It seems to form a particularly tough case for backward induction since at any node after the first the actor who is to make a choice knows that the principle must have been violated in reaching that node.

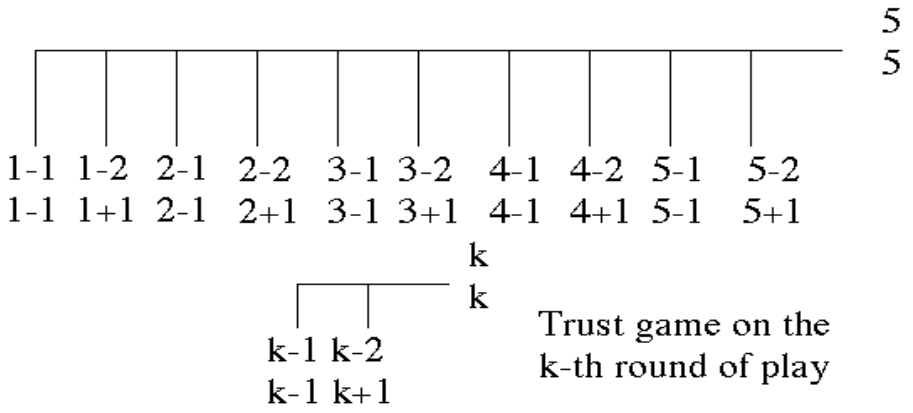

Figure 9 Decipede and generic constituent trust game

Since we believe that backward induction seems so counter-intuitive because we all tend to treat the subjective payoffs of RCM as if they were objective payoffs, imagine a decipede in *objective* payoffs, *first*. Refer to that decipede game as o-d-1. Next, imagine an exact replica of o-d-1 and refer to it as o-d-2. The compound structure which emerges after o-d-2 is "linked" to o-d-1 is o-d-3=(o-d-1&o-d-2). Here o-d-2 is put in as an adjunct after the last node of o-d-1. Let (o-d-3/o-d-1) denote the sub-game of o-d-3 that starts after o-d-1 ends. None of the *objective* payoffs is affected by combining or separating any of the games and, thus, clearly, o-d-2=(o-d-3/o-d-1). However, within RCM separability must apply to a context with subjective payoffs. So turn *second* to games proper. Let us call the game in subjective payoffs that emerges from o-d-1 the subjective decipede game or s-d-1. Let again s-d-2 be an exact replica of s-d-1. As long as s-d-1 and s-d-2 are considered separately the solution should remain the same. Finally refer to the subjective payoff game corresponding to o-d-3 as s-d-3. Now it is well possible that the sub-game corresponding to o-d-3/o-d-1 namely s-d-3/s-d-1 is different from s-d-2; i.e. s-d-2 ≠ s-d-3/s-d-1. The history leading to a node may influence payoffs that apply at that node and after. But since the influence is on the payoffs this does not violate separability. The history is so to say "in the payoffs" (absorbed as a causal influence on preferences). If we then cut the sub-game with the modified payoffs off (fixing those payoffs) the solution would remain the same as when embedded in the game even though the solution would not be the same as that of s-d-1=s-d-2. Vice versa, *if* we assume that a longer tree s-d-3 has in fact a sub-tree in subjective payoffs as s-d-2 then what we are talking about are games for which s-d-2=s-d-3/s-d-1. If it is in fact *the same game in subjective* payoffs why should it have a different solution? The payoffs being representations of preferences that have been formed *all things considered* it seems quite reasonable to assume that the solution remains the same. (That it may be exceedingly unlikely that the latter ever be the case is a different matter altogether!)

Each of the games of Figure 9 ends after playing down. This makes it hard to imagine the rules of the decipede game as emerging from identical repetitions of the same constituent game. However, not the solution theory based on backward induction but rather the assumption that super-games can be construed as repetitions of an *identical* constituent game in *subjective* payoffs is extremely

implausible. In RCT, that individuals are rational fools in the sense of having preferences that are independent of former histories is indeed almost a complete counterfactual. But taking into account path dependency of the *preferences* backward induction is plausible – at least if we are willing to concede that game theoretic analysis of models formulated in the language of RCM is possible at all. Either the analysis of a game can be based on the rules of the game alone or not. Game theory as understood here requires that the analysis of games be based on the rules of games. The implication of this is, that we must describe a game such that the analysis on the basis of that description becomes viable. But then we also buy into something akin to separability and backward induction.[16] (Otherwise some form of holism according to which all games are different and cannot be analyzed by parts would emerge anyway.)

According to the preceding argument backward induction in the decipede (in subjective payoffs) is fully coherent. To solve the backward induction puzzle in RCM does not justify the conclusion that everything is right with a conventional rational choice approach. RCT seeks to explain behavior in terms of given preferences. And therefore quite the contrary is true. First, if all commitments are explicitly modeled and preferences are "given" all things considered then the analysis of a game is basically complete after writing down the tree. Second, the whole notion of representative utility was from the beginning *un*suitable for any kind of eductive analysis. Strategic thinking is much more naturally invoked if the questions are formulated in terms of objective or material payoffs that are *desired* by an actor and may show up among the *reasons* for action. Third, it is not clear at all whether we can separate the specification of the rules of the game, in particular the fixing of the preferences, from the strategic analysis of the game in terms of "given" preferences. Obviously "all things considered" must not include the strategic considerations that lead to preferring an action x over some action y at some node. But then, how exactly can we draw the line between formulating and analyzing the game tree? Fourth, starting from given preferences that are merely represented stenographically by a utility function leaves in the dark practically everything that is of explanatory interest. Why care about such a rational fools' world at all?

Though we do think that RCM and conventional RCT were in fact a tremendous success as instruments of reflection on and understanding of interaction in strategic situations the preceding account leaves us with the riddle why that was so and what exactly social theory in general and economics in particular learnt from going through that extended exercise. By starting from "given" preferences all things considered it becomes practically impossible to account for human action in terms of real processes operating in the human psyche. Casting over human motivation the veil of the utility representation we abstract from too much that is relevant for motivating human behavior. Introducing such constructs as the agent form we can capture some of the inner constraints of choice making without giving up the (as if) maximization of utility. In this paper we illustrated that and how this can be done but is it worth it? Or to put it more bluntly what are the merits of going to the extremes of RCM?

[16] It is true that heroic efforts notwithstanding (see in particular **Harsanyi, John C. and Selten, Reinhard.** *A General Theory of Equilibrium Selection in Games*. Cambridge, Mass.: MIT Press, 1988.) it is at least contested whether there is a definite solution theory singling out a unique mode of rational play in all fully specified games as formulated in terms of RCM. So, if we assume that there is no definite solution of a sub-game could not a preceding history lead to the selection of a solution for the embedded sub-game? Since this is again a forward induction argument – typically for equilibrium selection – we have already dealt with it. Moreover, if we had set valued solution theories that select a set S as solution of a game – instead of a unique strategy combination – we would still have to explain why embedding an informationally closed sub-structure into different games should *systematically* affect solutions.

3. Eductive analyses, objective success and psychology

The chief merit of modeling the world as the scene of rational choice making of "rational fools" consists in putting the focus clearly on the delicate interplay between opportunism and commitment. Economics always had a clear perception of opportunism and was willing to insist on its presence. This gave economics the edge over the well-intentioned normativism of much of social science and social philosophy.[17] Non-cooperative game theory has been instrumental to spelling out all the implications of the traditional economic focus on opportunistic rationality. But in spelling out the implications it became completely clear, too, that the traditional (in particular also the traditional philosophical) criticisms of the model of opportunistic rationality as applied to personal players were justified. As a matter of fact human personal actors are not rational fools and a theory that makes an empirical claim to that effect is patently wrong. However, at the other extreme, a theory that eliminates the human faculty to act opportunistically completely will not lead to adequate accounts of human social interaction either.

Rational choice modeling as such cannot tell us anything about the real world but can possibly help us to state more clearly the theories and views that we may hold about the world. Again the test of the pudding will be in the eating and it would be very interesting to see whether RCM besides its obvious philosophical merits has some empirical, too[18]. But quite apart from such possibilities of using rational choice modeling as a tool of formulating empirical theories about the world, whenever we observe phenomena as being fully in line with assumptions of opportunistically rational choice we better acknowledge that we have an explanandum not an explanans. We are provoked to ask how in the world it was possible that things appear *as if* brought about by opportunistically rational choice. Convincing answers to this question like Armen Alchian's model of market selection or Vernon Smith's experiments on how markets clear are formulated in terms other than rational choice – or homo oeconomicus (see (Armen A. Alchian, 1950), (Vernon L. Smith, 2000)). Markets are substitutes for individual rationality and work even with rational fools (see (Dhamanjay K. Gode and Shyam Sunder, 1993)).

The assumption that preferences are "given all things considered" (rather than opportunism as possibly conceptualized in terms of objective payoffs) is behind the impression that economics perceives the world as if populated by rational fools. It simplified economics and rational choice theory greatly. But the widely shared hope that economics, relying on the concept of representative utility, might not need a foundation in (cognitive) psychology is completely mistaken. It seems surprising that theorists of social interaction could ever have believed that they could understand the interaction of rational *persons* by reducing them to a utility cum probability function. If any real progress is to be made economists will have to re-consider their basic explanatory strategy and will have to look behind the veil of the preference and belief representation by "utility cum probability".

The ability to distinguish between what is among the causal consequences of our choices and what is not as well as our faculty to seize opportunities are clearly related to those higher faculties of the mind that we commonly associate with "human reason". Without the ability to act opportunistically

---

[17] Of course, not all social theorists were "normativists". Hobbes, see in particular chapters 10-17 in, **Hobbes, Thomas.** *Leviathan*. Harmondsworth: Penguin, 1651/1968., and Spinoza, see in particular chapter 16 in, **Spinoza, Benedikt de.** *A Theologico-Political Treatise. A Political Treatise*. New York: Dover, 1670/1951., are, of course, more hard-nosed adherents of homo oeconomicus than most economists.

[18] Alex Tabarrock raised the issue whether the models formulating the internal commitment structure of an actor might lead to empirical predictions that could be tested. They might, but then we would presumably have to put in so much empirical knowledge from cognitive psychology that we perhaps should just go for that kind of modelling entirely.

rational it would hardly be conceivable to speak of human rationality as we know it. Saying this we are quite willing to concede that there may be other forms of rationality besides "opportunistic rationality". These other forms of rationality may include all sorts of rule following behavior. Since the latter exhibit certain types of "boundedness" they should, however, be classified as "boundedly rational" behavior. Calling them boundedly as opposed to opportunistically rational we are not implying that they are inferior forms. They may well be superior in leading to superior outcomes as measured in objective terms. But a theory of social interaction in terms of the cognitive psychology of boundedly rational choice making will in any event lead to insights superior to any theory starting from the rational fools' assumption of given preferences.

In sum:

Rational Choice *as*			
Maximization			**Non-Maximization**
Local maximization "opportunism"	Global maximization	Dual maximization global & local	Rule guided bounded behavior
Advantages: 1. Human faculty to seize opportunities is prominent feature 2. One off PD solution in dominant strategies stands 3. Clear statement of all assumptions about commitments	Advantages: 1. No backward induction and chain store paradoxes 2. Personal players represent persons	Advantages: 1. Fundamental intuitions that involve some trade-off between local and global maxi-mization are respected 2. Opportunism and commitment in the same model	Advantage: Cognitive science psychology etc. can be used to explicate "rational choice" as close to real choice
Disadvantages: 1. Backward induction and chain store paradoxes emerge 2. Dissolution of persons into agents eliminates eductive interpretation of game theory	Disadvantages: 1. Elimination of sub-game perfect-ness problems 2. One off PD solution in dominant strategies does not stand 3. Unclear distinction between committed and uncommitted choice	Disadvantage: There is no criterion when global should dominate local considerations and vice versa	Disadvantage: There is no way to find a standard of ideally rational behavior that is not open to revision by the facts of choice

References

Ainslee, George. *Picoeconomics*. Cambridge: Cambridge University Press, 1992.
Alchian, Armen A. "Uncertainty, Evolution, and Economic Theory." *Journal of Political Economy*, 1950, *Vol. 58*, pp. 211-21.
Brennan, H. Geoffrey and Kliemt, Hartmut. "Finite Lives and Social Institutions." *Kyklos*, 1994, *47*(4), pp. 551-71.
Broome, John. *Weighing Goods. Equality, Uncertainty and Time.* Oxford: Basil Blackwell, 1991.
Elster, Jon ed**.** *The Multiple Self*. Cambridge: Cambridge University Press, 1987.
Gode, Dhamanjay K. and Sunder, Shyam. "Allocative Efficiency of Markets with Zero Intelligence Traders: Markets as a Partial Substitute for Individual Rationality." *Journal of Political Economy*, 1993, *101*, pp. 119-37.
Harsanyi, John C. "Games with Incomplete Information Played by Bayesian Players." *Management Science*, 1967-8, *14*, pp. 159-82, 320-34, 486-502.
Harsanyi, John C. and Selten, Reinhard. *A General Theory of Equilibrium Selection in Games*. Cambridge, Mass.: MIT Press, 1988.
Hausman, Daniel M. "Sympathy, Commitment and Preference," University of Wisconsin-Madison, 2004.
Hobbes, Thomas. *Leviathan*. Harmondsworth: Penguin, 1651/1968.
Kohlberg, Elon and Mertens, Jean-Francois. "On the Strategic Stability of Equilibria." *Econometrica*, 1986, *54*(5 (September)), pp. 1003-37.
McClennen, Edward F. *Rationality and Dynamic Choice - Foundational Explorations*. New York / Port Chester / Melbourne / Sydney: Cambridge Univeristy Press, 1990.
____. "Rationality and Rules," P. A. Danielson, *Modeling Rationality, Morality and Evolution.* New York and Oxford: Oxford University Press, 1998, 13-40.
Neumann, John von and Morgenstern, Oskar. *Theory of Games and Economic Behavior*. Princeton: Princeton University Press, 1944.
Selten, Reinhard. "Reexamination of the Perfectness Concept for Equilibrium in Extensive Games." *International Journal of Game Theory*, 1975, *4*, pp. 25-55.
Sen, Amartya K. "Behaviour and the Concept of Preference," *Choice, Welfare and Measurement.* Oxford: Basil Blackwell, 1973/1982, 54-73.
Smith, Vernon L. ed**.** *Bargaining and Market Behavior*. Cambridge: Cambridge University Press, 2000.
Spinoza, Benedikt de. *A Theologico-Political Treatise. A Political Treatise*. New York: Dover, 1670/1951.
Sugden, Robert. "Rational Choice: A Survey of Contributions from Economics and Philosophy." *The Economic Journal*, 1991, *101*(July), pp. 751-85.
Verbeek, Bruno. "The Feasibility of Rational Self-Commitment," Leiden, 2004.