Rational Choice Theory and Bounded Rationality[1]

Till Grüne-Yanoff

1. Introduction

Rational Choice Theory (RCT) has dominated economics for more than 50 years, and it is becoming increasingly important in other social-science disciplines. At the same time, the critical voices against it are increasing in volume. One particularly successful research effort in this direction is the modelling of 'bounded rationality'. The idea is that it offers psychologically more plausible models of human decision-making without giving up on the notion of rationality altogether. In many cases, the research is supported by experimental findings that document deviations from standard RCT. However, rational-choice theorists are often not convinced, and most economists have not yet exchanged their mainstream models for a more 'boundedly rational' behavioural economics. This article gives some reasons why this may be so, and more generally why bounded rationality, although an important research programme in its own right, is not likely to replace RCT altogether.

Section 2 of the article briefly sketches the main features of the formal RCT framework, and discusses various normative and positive interpretations. Section 3 gives a taste of the empirical results that seem to contradict RCT, and considers how these results could be interpreted. Section 4 introduces a few of the bounded-rationality models offered in response to the empirical evidence discussed in the previous section. A categorisation of such models is offered, and their interpretation and usefulness in each of these categories are discussed. Section 5 concludes the paper.

1 Thanks to Aki Lehtinen for his helpful comments and to Mette Ranta for bibliographic assistance.

2. Rational Choice Theory and Economic Behaviour

RCT is the dominant theoretical approach in microeconomics.[2] It is also widely used in other social-science disciplines, in particular political science. In these disciplines the term rational choice theory is often used in association with the notion of economic 'imperialism', implying that the use of RCT is an extension of economic methodology into their fields.

Explicit theories of rational economic choice were first developed in the late 19th century. These theories commonly linked choice of an object to the increase in happiness an additional increment of this object would bring. Early neoclassical economists (e.g., William Stanley Jevons) held that agents make consumption choices so as to maximise their own happiness. In contrast to them, 20th-century economists disassociated RCT and the notion of happiness: instead of explicating rationality of choice as a happiness-maximising effort they presented rationality as maintaining a consistent ranking of alternatives. Such a ranking is commonly interpreted as agents' desires or values.

Without a foundation in an ultimate end the notion of rationality is reduced to the consistent ranking of choice alternatives, the consistent derivation of this ranking from evaluations of the possible outcomes, and the consistency of beliefs employed in this derivation. Thus, 'rationality' explicated in rational choice theory is considerably narrower and possibly sometimes at odds with colloquial or philosophical notions of rationality. In such contexts 'rationality' often includes judgments about ends, the prudent weighting of long-term versus short-term results, and insights into purportedly fundamental moral principles. Nothing of this sort is invoked in rational choice theory. It simply claims that a rational person chooses actions in a manner consistent with her beliefs and evaluations. Accordingly, a person considered 'rational' in this sense may believe that the moon is made of green cheese, may desire to waste her life, or may intend to bring widespread destruction.

2 The term 'rational choice theory' is rarely used in economics, but became the term of choice in other disciplines, signifying the core theoretical assumptions of microeconomics.

2.1. Formal Framework

At the core of RCT is a formal framework that (i) makes the notion of preference consistency precise and (ii) offers formal proof that 'maximising one's utility' is identical to 'choosing according to a consistent preference ranking'. A brief sketch of this framework follows.[3]

Let $A = \{X_1, \ldots, X_n\}$ be a set of alternatives. Prospects are either pure prospects or lotteries. A pure prospect is a future event or state of the world that occurs with certainty. For example, when purchasing a hamburger from a well-known international restaurant chain I may expect with near certainty the pure prospect of certain taste experiences. Lotteries, also called prospects under risk, are probability distributions over events or states. For example, when consuming 'pick-your-own' mushrooms an agent faces the lottery $(X_1,p; X_2,1\text{-}p)$, where X_1 denotes the compound outcome (which has probability p) of falling ill due to poisoning and X_2 (with probability $1\text{-}p$) the compound outcome of not doing so. More generally, a lottery X consists of a set of prospects $X_1,\ldots,X_n$ and assigned probabilities $p_1,\ldots,p_n$, such that $X = (X_1,p_1;\ldots X_n,p_n)$, where $p_1+\ldots+p_n=1$. Obviously, the prospects $X_1,\ldots,X_n$ can be lotteries themselves.

RCT takes preferences over actions to be evaluations of lotteries over action outcomes. Its main contribution is to specify the relationship between preferences over actions, and preferences as well as beliefs over the compound outcomes of the respective lottery. It does so by proving *representation theorems*. Such theorems show that under certain conditions, all of an agent's preferences can be represented by a numerical function, the so-called utility function. Furthermore, the theory shows that the utility numbers of an action (i.e. lottery) $X = (X_1,p_1;\ldots X_n,p_n)$ and its compound outcomes $X_1, \ldots, X_n$ are related to each other through the following principle:

$$u(X) = \Sigma_i\, p_i \times u(X_i) \quad (1)$$

In other words, the utility of a lottery is equal to the sum of the utilities of its compound outcomes, weighted by the probability with which each outcome comes about. This is an important result that significantly constrains the kind of preferences an agent can have. Of course, because

3 The framework presented here is based on von Neumann and Morgenstern (1947). Alternative formal frameworks are to be found in Savage (1954) and Jeffrey (1990).

the representation result is a formal proof, all the constraining information must already be present in the theorem's assumptions. I will sketch the main features of these assumptions here.[4]

RCT assumes that at any time, there is a fixed set of prospects $A = \{X_1, \ldots, X_n\}$ for any agent. With respect to the agent's evaluation of these prospects, it assumes that agents can always say that they prefer one prospect to another or are indifferent between them. More specifically, it assumes that the agent has a preference ordering $\geq$ over A, which satisfies the following conditions. First, the ordering is assumed to be *complete*, i.e.

$$\text{either } X_i \geq X_j \text{ or } X_j \geq X_i \text{ for all } X_i, X_j \in \text{A}. \quad (2)$$

Second, the ordering is assumed to be *transitive*, i.e.

$$\text{if } X_i \geq X_j \text{ and } X_j \geq X_k, \text{ then also } X_i \geq X_k \text{ for all } X_i, X_j, X_k \in \text{A}. \quad (3)$$

Completeness and transitivity together ensure that the agent has a so-called weak ordering over all prospects.

The second domain in which RCT makes consistency assumptions concerns beliefs. In particular, it assumes that each rational agent has a *coherent set of probabilistic beliefs*. Coherence here means that beliefs can be represented as probability distributions, which satisfy certain properties. In particular, it is assumed that there is a probability function p over all elements of A, and that this function satisfies the following assumptions: first, for any X, $1 \geq p(X) \geq 0$; second, if X is certain, then $p(X) = 1$; third, if two alternatives X and Y are mutually exclusive, then $p(X \text{ or } Y) = p(X) + p(Y)$; finally, for any two alternatives X and Y, $p(X \text{ and } Y) = p(X) \times p(Y|X)$ – in other words the probability of the alternative 'X and Y' is identical to the probability of X multiplied by the probability of Y *given* that X is true.

The third domain in which rational choice theory makes consistency assumptions concerns preferences over lotteries. In particular, it assumes the *independence condition*. If a prospect X is preferred to a prospect Y, then a prospect that has X as one compound outcome with a probability p is preferred to a prospect that has Y as one compound with a

4 For a detailed discussion, see the references in footnote 3. For more in-depth overviews, see textbooks such as Luce and Raiffa (1957); Mas-Collel et al. (1995, chs. 1&6) and Resnik (1987). Hargreaves Heap et al. (1992, 3–26) give an introductory treatment.

probability p and is identical otherwise: i. e. for all X,Y,Z: if $X \geq Y$ then $(X,p;Z,1\text{-}p) \geq (Y,p;Z,1\text{-}p)$.

These assumptions (together with a few others that are not relevant here), imply that preferences over lottery prospects $X = (X_1,p_1;\ldots X_n,p_n)$ are represented by a utility function such that for all X,Y:

$$X \geq Y \Leftrightarrow \Sigma_i\,[p_i \times u(X_i)] \geq \Sigma_i\,[p_i \times u(Y_i)] \quad (4)$$

This formal result has been given different interpretations, which are discussed in the next two subsections.

2.2. Normative Interpretations

RCT is often interpreted as a theory of how people *ought* to form their preferences (and by extension how they ought to choose). Accordingly, people who violate RCT in their actual deliberations or behaviour might still be subject to a normative standard of preference consistency spelled out in RCT.

Such a normative standard has been justified in various ways. The most prominent justifications are pragmatic: they seek to show that agents who fail to retain consistency in the way RCT prescribes incur certain losses. Two well-known examples are the *money pump* and the *Dutch book* arguments.

The money pump (Davidson et al. 1955) can be illustrated as follows. A stamp collector has preferences with respect to three stamps, denoted *A*, *B*, and *C*. She prefers *A* to *B*, *B* to *C*, and *C* to *A*, hence violating the transitivity assumption of RCT. She is willing to pay 10 cents for a prospect that she prefers to the status quo. She goes into a stamp shop with stamp *A*. The dealer offers to trade *A* for *C* if she pays 10 cents. She accepts the deal. The dealer then offers to trade *C* for *B*, which she again accepts, paying another 10 cents. The dealer then offers to trade *B* for *A*. Again, given her preferences, she will accept the deal and pay 10 cents. Thus her preferences leave her open to being 'pumped': she leaves the shop with the same stamp she had when she entered it, but 30 cents poorer. It seems that violations of transitivity yield a certain loss for the violating agent.

The Dutch book argument (Ramsey 1931) can be illustrated as follows. An agent *A*'s degrees of belief in *S* and $\neg S$ (written $p(S)$ and $p(\sim S)$) are each .51. Their sum is 1.02, and hence *A* violates the axioms of RCT: according to RCT a person with degree of belief p in sentence

S is assumed to be willing to pay up to and including $\$p$ a bet on the truth of *S*, and is willing to sell it for any price equal to or greater than $\$p$. Now a bookmaker sells a bet on *S* to *A* – *A being* willing to pay $\$p(S)$ – and also sells a bet on $\neg S$ –A being willing to pay $p(\neg S)$). A's beliefs make her pay $1.02 on a combination of wagers guaranteed to pay exactly $1. *She* would thus have a guaranteed net loss of $.02. It seems that violations of probability laws yield a certain loss for the violating agent.

Interpreted literally, neither the money pump nor the Dutch book is very convincing. An agent could simply refuse to accept money-pumping trade or Dutch-booking bets. Thus, rationality does not literally require that one is willing to wager in accordance with the RCT assumptions described above. It is more plausible to interpret these arguments hypothetically. Both could be conceived of as heuristic in determining when one's preferences or degrees of belief have the potential to be *pragmatically self-defeating*. Given any reasonable way of translating one's mental states into action, preferences or degrees of belief that violate RCT motivate one to act in ways that make things worse than they might have been when, as a matter of mere logic, alternative actions would have made things better.[5]

2.3. Positive Interpretations

RCT is often conceived of as a formalisation of folk psychology (e.g., Ferejohn 2002; Cox 1999; Coleman 1990). 'Folk psychology' here refers to pre-theoretical psychology based on intentional states of belief and desire. People who use this term commonly believe that our everyday or 'folk' understanding of mental states constitutes a theory of mind, which could also be used to explain purposeful action. According to this interpretation, RCT models the folk notion of belief as probabilities and the folk notion of desire as preferences, and becomes a formally exact basis for explaining intentional action.

Many economists would disagree with such an interpretation. In 1938, Paul Samuelson showed that all RCT assumptions could be reinterpreted as constraints on choices, and that the whole theory of consumer behaviour could thus be 'freed from any vestigial traces of the

5 For more on this and other normative justifications, see Hansson and Grüne-Yanoff 2009, sec 1.

[hedonistic] utility concept' (Samuelson 1938, 71). Choices reveal preferences, and choices over lotteries reveal subjective probabilities. Without attributing any mental states to the subject, RCT merely uses the consistency of choice as an explanatory device.[6]

Consistency, in both its psychological and its behavioural interpretation, offers only relatively weak constraints on actual preferences or choices. The problem is that without definite descriptions of preference content or choice options it is easily possible to provide *ad hoc* explanations of seemingly anomalous behaviour. For example, players who cooperate in a one-shot Prisoners' Dilemma (behaviour that violates standard utility maximisation) could be said to be acting from altruistic motives, or to believe in playing an indefinitely repeated game. People who mount such a defence commonly see RCT as a mere conceptual scheme that needs to be filled with content for specific tasks. In their view, the framework is not then open to empirical refutation.[7]

Mainstream economists have mostly taken a different route. They conceive of RCT not as a mere conceptual frame, but as a substantial theory of rational self-interest. In their theories of consumer choice they supplement its consistency requirements with the assumptions of self-interest (all agents' preferences are independent of each other), non-satiation (more is always preferred to less) and the marginal rate of substitution (for all goods *X*, *Y*, all individuals are willing to exchange more of *Y* for a unit of *X* as the amount of *Y* is increasing relative to *X*) (Hausman 1992, 30). Most microeconomic models are based on these assumptions, and they are used to explain a wide range of phenomena, from consumer demand, goods prices and bargaining behaviour to social conventions and legal institutions.

It is important to understand (i) that these additional assumptions are not strictly part of RCT, and (ii) that they have an explanatory or predictive purpose, but not a normative one. Indeed, as Sen (1987) points out, it would be absurd to assume self-interest or non-satiation as a normative rationality requirement. Nevertheless, these two dimensions are often confused in the debate about RCT.

6 See Wong 1978 for a critical analysis of the revealed preference approach.

7 For an argument to that end, see Gintis (2009, ch. 12).

3. Anomalies

According to Thomas Kuhn's account of scientific revolutions, anomalies are worrying puzzles for a scientific discipline that could lead to a loss of confidence in the discipline's paradigm. 'Anomalies' also was the heading of a regular column written by economist Richard Thaler in the *Journal of Economic Perspectives* (from 1987 to 1990), in which he documented instances of individual behaviour that seemed to violate RCT. These cases helped to raise economists' doubts about the theory, although many similar anomalies were documented before and have been since. This section discusses a few cases of this kind.[8]

A prominent example of an RCT anomaly is the so-called Allais' Paradox. Allais' (1953) idea was to find two pair-wise choices such that RCT would predict a specific choice pattern, and then check the prediction in the laboratory. This choice experiment is described in Figure 1.

Choice problem 1 – choose between:

A:	\$2500	with probability 0.33	*B:*	\$2400	with certainty
	\$2400	with probability 0.66			
	\$0	with probability 0.01			

Choice problem 2 – choose between:

C:	\$2500	with probability 0.33	D:	\$2400	with probability 0.34
	\$0	with probability 0.67		\$0	with probability 0.66

Fig. 1

RCT prescribes and predicts that agents choose *C* if they have chosen *A* (and vice versa), and that they choose *D* if they have chosen *B* (and vice versa). To see this, simply re-partition the prizes of the two problems as follows. Instead of '2400 with certainty' in *B*, partition the outcome such that it reads '2400 with probability 0.66' and '2400 with probability 0.34'. Instead of '0 with probability 0.67' in *C*, partition the outcome such that it reads '0 with probability 0.66' and '0 with probability

8 For more detail, see e.g., Kahneman et al. 1982; Thaler 1992; Gigerenzer et al. 1999.

0.01'. Of course, these are just re-descriptions that do not change the nature of the choice problem. They are shown in Figure 2.

Choice problem 1* – choose between:

A:	\$2500	with probability 0.33	*B*:*		
	\$2400	with probability 0.66		\$2400	with probability 0.66
	\$0	with probability 0.01		\$2400	with probability 0.34

Choice problem 2* – choose between:

C:	\$2500	with probability 0.33	D:		
	\$0	with probability 0.66		\$0	with probability 0.66
	\$0	with probability 0.01		\$0	with probability 0.66

Fig. 2

Through this re-description we now have an outcome '2400 with probability 0.66' both in *A* and in *B**, and an outcome '0 with probability 0.66' both in *C** and in *D*. According to the RCT independence condition, these identical outcomes can be disregarded in the deliberation. But once they are disregarded it becomes clear that option *A* is identical to option *C** and option *B** is identical to option *D*. Hence, anyone choosing *A* should also choose *C* and anyone choosing *B* should also choose *D*. However, in sharp contrast to this claim, in an experiment involving 72 people, 82 per cent of the sample chose *B*, and 83 per cent chose *C* (Kahneman and Tversky 1979).

Other major anomalies include Ellsberg's paradox (cf. Resnik 1987, 105–107), according to which a perception of ambiguity distorts rational belief formation. The framing effect (Tversky and Kahneman 1981) shows how background conditions and descriptions of alternatives influence choice, sometimes to the extent that it violates RCT. Status quo is the tendency for people to like things to stay relatively the same. It has been detected in various contexts, such as 'loss aversion', where the disutility of giving up an object is greater than the utility associated with acquiring it, and the 'endowment effect', where people often demand much more to give up an object than they would be willing to pay to acquire it (Kahneman et al. 1991).

3.1. Interpreting Anomalies

A valid anomaly is an observation of systematic behaviour (under appropriate laboratory conditions) that contradicts one of the deductive implications of RCT. According to superficial versions of falsificationism, any anomaly poses a serious threat to RCT.

However, section 2 identified various uses and interpretations of RCT. Depending on the envisaged use and the intended interpretation the theorist may be justified in continuing to use it in the face of certain valid anomalies.

The first case concerns the distinction between positive and normative use. Framing effects or status quo biases, for instance, may challenge explanatory or predictive uses of RCT. Yet there is little reason to believe that they pose problems for its normative use. If someone holds that RCT has normative force, the fact that many or even most people violate these principles is irrelevant. On the contrary, it is because these principles are often violated that the importance of their normative content increases.

The issue is more complicated with anomalies such as Ellsberg's and Allais' paradoxes. Not only do they constitute an apparent threat to positive uses of RCT, they are often thought also to affect its normative standing. According to Savage:

> If, after thorough deliberation, anyone maintains a pair of distinct preferences that are in conflict with the sure-thing principle [his version of the independence condition], he must abandon, or modify, the principle; for that kind of discrepancy seems intolerable in a normative theory (Savage 1954, 101).

According to anecdotal evidence, Savage himself was unsure about the normative validity of RCT after being confronted with Allais' paradox. Although there are no clear methodological guidelines for assessing fundamental normative claims, such evidence at least opens up the possibility of a normative rejection of RCT (on the methodological issues involved, see Guala 2000).

The second case concerns the distinction between RCT as a (positive) theory of cognition versus RCT as a (positive) theory of behaviour. The theory may be behaviourally realistic in the sense that it correctly describes human behaviour, and it may be psychologically realistic in the sense that the mental states and processes it evokes can be correctly attributed to decision makers. Thus it could be employed in conjunc-

tion with three different kinds of true claims about the world: (i) as both behaviourally and psychologically realistic, (ii) as behaviourally realistic but psychologically unrealistic, and (iii) as psychologically realistic but behaviourally unrealistic. As discussed in section 2, RCT in the social sciences is often interpreted in accordance with both the first and the second claims.

This complicates the question of how relevant observed anomalies are to the positive interpretation of RCT. Clearly, if the findings are valid and systematic they challenge claim (i), but it is more difficult to show that they also challenge claim (ii) concerning behavioural realism. Caution is particularly in order when specific assumptions of RCT are singled out and shown to be 'unrealistic'. Although such claims may be correct, they may not have any relevance for users of RCT who insist on its behavioural realism.[9] Friedman (1953) insisted on this possibility in his highly influential article, arguing that the unrealisticness of its assumptions is no reason for complaint or worry about a theory. This is often interpreted as an argument based on predictive instrumentalism: underlying assumptions, especially psychological ones, need not be realistic as long as the model results succeed in predicting behaviour well. Mäki (2009a), however, notes that Friedman often leaves predictive purposes aside when considering the benefits of unrealistic assumptions, suggesting instead that he conceived of theory construction as a matter of theoretical isolation whereby economists abstract essential features of complex reality.

This brings me to the third kind of claim, that RCT may be psychologically realistic but behaviourally unrealistic. Although it attracts less attention in the scientific literature, it offers a plausible interpretation, which in turn challenges the relevance of many of the anomalies. It is a long-standing tradition, going back at least to Mill and Marshall, to argue that successful theory isolates the workings of certain factors in the world. To take an example, wealth maximisation is an important causal factor of choices made in the economic domain, but it is not the only one. Rather, its influence on choice is compounded by other causal factors. Thus, in the real world we should not expect to observe the unobstructed operation of wealth maximisation: what one could hope for at best is to observe its unobstructed operation in con-

9 The examples mentioned in this section do challenge even this interpretation, however, as they show that people systematically *choose* in a way that is inconsistent with RCT.

trolled experiments. Nevertheless, if causal factors such as wealth maximisation could be neatly separated from other factors, our theories may strive to represent the operation of these factors in their purity (Mäki 1994, 2009b). If RCT is interpreted in this fashion, the above anomalies may not challenge it. We should expect observed behaviour to deviate from the conclusions of an isolating theory: it may well also be influenced by other factors. What matters is whether the theory successfully isolates the actual operation of one of the contributing factors. Most behavioural experiments give no answers to such questions about the underlying cognitive mechanisms.

4. Bounded Rationality

The development of models of bounded rationality was largely triggered by dissatisfaction with the dominant RCT theories. According to Herbert Simon, who is commonly seen as its main pioneer, the point of bounded rationality is to

> designate rational choice that takes into account the cognitive limitations of the decision maker – limitations of both knowledge and computational capacity (Simon 1987, 266).

Simon's efforts were largely directed at finding an adequate formal characterisation of rationality. Today, the term 'bounded rationality' has acquired a more general meaning that includes all efforts at modelling choices with more cognitive and informational limitations than RCT assumes.[10] As I aimed to show in the last section, the way in which RCT is considered unsatisfactory depends on the interpretation and the kind of anomalies considered. Accordingly, at least three strands of bounded-rationality models can be distinguished, as shown in Fig. 3.

	Positive	**Normative**
Behaviourally realistic	Positive behavioural models	⋆
Psychologically realistic	Positive procedural models	Normative procedural models

Fig. 3

10 On the history of the concept of bounded rationality, see Klaes and Sent (2005).

Models of bounded rationality have been developed for the positive purposes of explanation and prediction as well as for the normative purpose of explicating what it is to be rational. On the positive side some models claim only to capture behavioural deviations from RCT, whereas others claim to capture them by modelling the underlying psychological mechanisms. On the normative side, in turn, the aim is to propose deliberation procedures that rational agents should follow. It is noteworthy that no normative behavioural models of bounded rationality have been put forward. This may be a contingent matter of fact, but one could speculate that it seems conceptually difficult to challenge RCT as a normative standard for choice results.

4.1. Positive Models

An influential family of positive behavioural models focuses on agents' probability misperception.[11] Rank-dependent expected utility theory (RDU, Quiggin 1982) has become the most popular member of this family. If the outcomes of a lottery are ordered so that $X_1 > X_2 > \ldots > X_n$, RDU is calculated as the weighted utility of the outcomes:

$$RDU(p_1,X_1;\ \ldots p_n,X_n) = \Sigma\ \pi_j \times u(X_j) \quad (5)$$

where the probability weight π_i of an outcome X_i depends on its probability *and* the ranking position of the outcome:

$$\pi_j = w(p_1+\ldots+p_j) - w(p_1+\ldots+p_{j-1}) \quad (6)$$

The intuition behind the theory is that the degree of attention agents give to an outcome depends not only on its probability, but also on its favourability in comparison to other possible outcomes (Diecidue and Wakker 2001). Pessimists, for example, tend to overemphasise 'bad' outcomes of a lottery, believing (irrationally) that unfavourable events tend to happen more often. Their attitude is characterised by the convex weighting function *w*. Optimists, on the other hand, tend to overemphasise favourable outcomes, hence their attitude is characterised by a concave *w*. Rank-dependent utility satisfies basic intuitions

11 For a wider 'sampler' of bounded rationality in economic models, see Conlisk (1996) or Starmer (2000); for an in-depth presentation of a selection of models, see Rubinstein (1998).

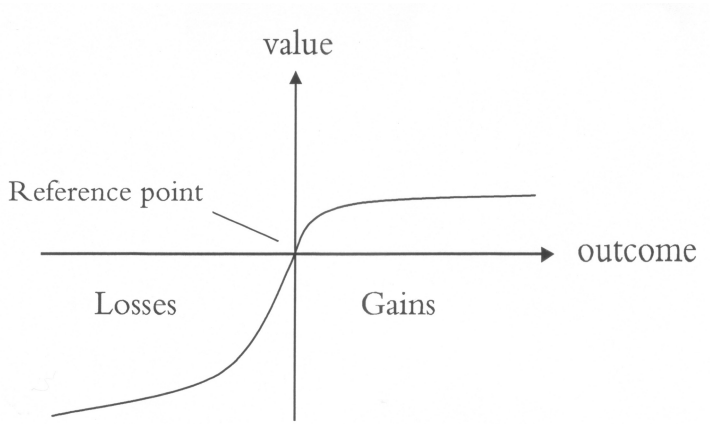

Fig. 4

about rationality.[12] Nevertheless, it accounts for the agents' behaviour in the Allais' paradox: a lot of people are pessimistic, and pessimists choose *B* because they overemphasise the possibility in *A* of not winning, whereas this does not make a big difference in comparison between *C* and *D*.

In contrast to such positive behavioural models, positive procedural models often attempt to spell out how agents *actually* reason and deliberate. One of the most prominent of these is Kahneman and Tversky's (1979) prospect theory. According to this theory the deliberation process is divided into two stages: editing and evaluation. In the editing stage the different choices are ordered following a variety of heuristics, and a reference point is determined. In particular, people decide which outcomes they see as basically identical. Then they set one such equivalence class as their reference point. In the evaluation stage prospects below the reference point are interpreted as losses, and prospects above it as gains. The value function (sketched in Figure 4) passing through this reference point is s-shaped. It is concave for gains and convex for losses, and it is steeper below the reference point.

12 In particular, it satisfies stochastic dominance, according to which within a given lottery, shifting positive probability mass from an outcome to a strictly higher outcome leads to a strictly higher evaluation of the transformed lottery.

Kahneman and Tversky (1979) interpret these two properties as *diminishing sensitivity* and *loss aversion*. Diminishing sensitivity implies that the psychological evaluation of an incremental increase of gain or loss will decrease as one moves further away from the reference point, hence the s-shape of the value function. Loss aversion holds that losses loom larger than corresponding gains, hence the increased steepness of the value functions below the reference point. The original version of prospect theory violates stochastic dominance. The editing phase may overcome this problem, but not necessarily so. A revised version, called cumulative prospect theory, uses probability weighting in a similar way to rank-dependent expected utility theory. Although this shows the closeness of prospect theory to positive behavioural models, the editing phase distinguishes it as a procedural theory.

These positive theories are subject to a number of criticisms. First, it is clear that none of the models capture human behaviour perfectly. Moving away from RCT at best gives us models that are a little less false. Thus the question arises whether the purported increased predictive and explanatory potential of bounded-rationality theories is enough to offset the undeniable decrease in parsimony of the new theories when compared to RCT. Such decreased parsimony has negative effects on explanation and prediction. Less parsimonious models are more difficult to grasp, and hence less likely to enhance understanding. They are also more likely to 'overfit' the data: the increased degrees of freedom produce a better fit to existing data, but they are more likely to pick up on irregularities in the sample that do not reflect the true trend. Formalising these trade-offs in order to facilitate proper theory choice is an important task that requires more attention (for a good example, see Harless and Camerer 1994).

Secondly, positive procedural models are open to instrumentalist critique. As Friedman (1953) argued, economic theories should not be judged by their assumptions but by their predictive implications. Yet procedural models focus precisely on the underlying cognitive mechanisms, possibly to the detriment of a more predictively powerful theory (in the sense argued in the preceding paragraph). Of course, this applies only to procedural models employed for predictive purposes, which arguably are rather rare.

Thirdly, it is a widespread misconception that the main goal of RCT is the explanation of individual behaviour. Most rational-choice theorists and, in particular economists, rather argue that the theory was designed mainly to explain aggregate-level phenomena. Specific

psychological assumptions in RCT models may not be relevant for such explanatory purposes, even if the same assumptions are explanatorily relevant to individual behaviour. What is relevant, however, is whether or not the aggregate-level model result (e.g., an equilibrium allocation or a dependency of variables derived from comparative statics) is an artefact of particular RCT modelling assumptions. In order to investigate this question modellers employ robustness analysis (Lehtinen and Kuorikoski 2007). They examine how hypothetical changes in the values of model variables or parameters would change the analytical results, and what they would leave intact. From this perspective bounded-rationality models have a role in that they suggest ways in which to vary model parameters in robustness analysis. At the same time it shows the limits of such models: although they identify deviations from RCT models in individual psychological or behavioural features, these deviations may turn out to be irrelevant in the analysis.

Fourthly, not all economic models are built with a view to prediction or explanation, and some rather serve as tools for conceptual exploration (Hausman 1992), or the investigation of possible explanations (Grüne-Yanoff 2009)., Such models are often interpreted as counterfactual worlds in which assumptions and hypotheses can be tested in the same way as a thought experiment, and neither predictive success nor explanatory realism is an objective. If the aim is conceptual exploration or possible explanation the introduction of a bounded-rationality assumption must somehow enhance understanding of the modelled counterfactual world. It is not always obvious that models such as the ones discussed above do indeed further understanding in these ways.

Finally, with the rapidly increasing number of competing bounded-rationality models, the danger of arbitrariness arises. Many disciplines and sub-disciplines that make use of such models only adopt the assumptions that suit their needs, and disregard others. This has aroused suspicion that bounded-rationality assumptions are employed as *ad hoc* remedies for deficient models, without any underlying theory providing clear guidance. Simon, in a letter to Rubinstein (1998), raised such a concern: "At the moment we don't need more models; we need evidence that will tell us what models are worth building and testing" (Simon, in Rubinstein 1998, 190).

Instead of constructing models from the armchair we need to develop a more general theory explaining why bounded-rationality assumptions are relevant in certain contexts and not in others, which is testable

and tested. Only on the basis of such a theory can a more principled model involving such assumptions be constructed.

4.2. Normative Procedural Models

The positive models discussed thus far focus largely on human cognitive limitations. Yet Simon's original account proposes two kinds of limitation on agent rationality, which operate together like a pair of scissors whose two blades are "the structure of task environments *and* the computational capacities of the actor" (Simon 1990, 7, emphasis added). Research on the first of these, the structure of the environment, focuses on establishing the dependence of computational or mental factors on environmental pressures, and on how environmental forces have selected simple heuristics for making decisions. The resulting concept of rationality differs substantially from any optimisation effort. Instead, the decision maker adapts the use of his choice rule to the environment in which he lives. On a wider scale this idea takes on an evolutionary perspective: biological evolution endowed humans with a multitude of special-purpose psychological modules for reasoning and decision-making.

This approach has contributed to both positive and normative procedural models. Its proponents argue that biological evolution has equipped humans with heuristics that make them *ecologically rational*. Ecological rationality is seen as conflicting with the demands of a normatively understood RCT, with its emphasis on maximising choices based on all available information.

> A computationally simple strategy that uses only some of the available information can be more robust, making more accurate predictions for new data, than a computationally complex, information-guzzling strategy that overfits (Gigerenzer et al. 1999, 20).

In stressing the procedural view on rationality defenders of ecological rationality argue that it may be disadvantageous to follow a procedurally understood RCT instead of employing intuitive heuristics that arose in the form of adaptation to specific environments. Thus, the approach assumes that agents have an *adaptive toolbox* at their disposal:

> …the collection of specialised cognitive mechanisms that evolution has built into the human mind for specific domains of inference and reasoning, including fast and frugal heuristics (Todd and Gigerenzer 2000, 740).

In the long ancestral history of humanity, the assumption goes, the selected features of the toolbox have been those that prove to be most useful for survival *in a specific environment*. In terms of normative rational assessment it is a question of how well these heuristics work in the experimental environments in which RCT anomalies are observed. Forcing people into environments that are irrelevant to them (say, expressing a relation to the world in terms of probability correlations in an experimental set-up) is not pertinent to normative assessment. This, defenders of ecological rationality claim, is exactly what the experiments devised by Kahneman and Tversky and others do: Subjects are placed in artificially created environments to which they are not adapted (Gigerenzer 1996). Against this, proponents of ABC insist that deliberation rules must only be tested in relevant environments, and that they work well in environments for which they have been adapted.

This adaptive argument faces two challenges, however. First, evolutionary arguments point to a disposition, not an actuality. Traits selected for fitness tend to be optimal, but there are various lacunae that provide causes why they are not. For example, suboptimal traits may be 'bundled' with traits that ensure survival, the environment may provide resources in such abundance that selective pressure is low, or competing traits may not be challenging. The fact that certain mechanisms are evolutionarily selected thus does not guarantee their optimality even for the environments for which they have been adapted. If this was the case, RCT could help humans improve on their adapted heuristics.

Secondly, competences adapted to pre-historic circumstances may be of no help in the modern world. The above claims imply that deliberation procedures are adapted to ancestral circumstances. To be normatively relevant, however, these procedures must also be adapted to current circumstances. Otherwise they may face the same fate as the Dodo when confronted with human settlers and their domesticated animals. Defenders of ecological rationality tend to suggest that adaptation to current circumstances follows from adaptation to ancestral circumstances, but no clear arguments are given to support this claim.

5. Conclusion

Without doubt, bounded rationality has proven to be a very fruitful and multifaceted research program. It has increased social scientists' sensitivity to the cognitive mechanisms underlying choice, and to systematic

behavioural deviations from the standard view. In many cases it has led to the development of innovative models that take account of these features. It has also cast some doubt on the normative adequacy of RCT, and motivated the search for alternative accounts of rationality.

Nevertheless, I hope to have shown in this article that the significance of these experimental results and modelling efforts are sometimes overemphasised. There is no reason for the social sciences to adopt bounded-rationality models across the board. Indeed, true appreciation of the multitude of different purposes for which RCT is employed makes it clear that it is better suited to some purposes than models of bounded rationality. At the very least, the question of which approach is better must be decided case by case, taking into account the available data, the scientific purpose, the results of robustness analysis, and general considerations of understandability.

References

Allais, Maurice (1953). Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école Américaine. *Econometrica* 21, 503–46.

Coleman, James S. (1990). *Foundations of social theory*. Cambridge: Harvard University Press.

Conlisk, John (1996). Why bounded rationality? *Journal of Economic Literature* 34, 669–700.

Cox, Gary W. (1999). The empirical content of rational choice theory: A reply to Green and Shapiro. *Journal of Theoretical Politics* 11(2), 147–69.

Davidson, Donald, John McKinsey, & Patrick Suppes. (1955). Outlines of a formal theory of value, I. *Philosophy of Science* 22, 140–60.

Diecidue, Enrico, & Wakker, Peter P (2001). On the intuition of rank-dependent utility. *Journal of Risk and Uncertainty* 23(3), 281–98.

Ferejohn, John A. (2002). Symposium on explanations and social ontology 1: Rational choice theory and social explanation. *Economics and Philosophy* 18(2), 211–34.

Friedman, Milton (1953). The methodology of positive economics. In *Essays in positive economics*, 3–43. Chicago: University of Chicago Press.

Gigerenzer, Gerd (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky. *Psychological Review* 103(3), 592–96.

Gigerenzer, Gerd, Todd, Peter M. & the ABC Group. (1999). *Simple heuristics that make us smart*. Oxford: Oxford University Press.

Gintis, Herbert. (2009). *The bounds of reason: Game theory and the unification of the behavioral sciences*. Princeton, NJ: Princeton University Press.

Grüne-Yanoff, Till. (2009). Learning from minimal economic models. *Erkenntnis* 70(1), 81–99.

Guala, Francesco. (2000). The logic of normative falsification: rationality and experiments in decision theory. *Journal of Economic Methodology* 7(1), 59–93.

Hansson, Sven Ove, & Till Grüne-Yanoff. (2009). Preferences. In *The Stanford encyclopedia of philosophy (Spring 2009 Edition)*, edited by Edward N. Zalta. <http://plato.stanford.edu/archives/spr2009/entries/preferences/>

Hargreaves-Heap, Shaun, Martin Hollis, Bruce Lyons, Robert Sugden, & Albert Weale. (1992). *The theory of choice: A critical guide*. Oxford: Blackwell.

Harless, David W., & Colin F. Camerer. (1994). The predictive utility of generalized expected utility theories. *Econometrica* 62, 1251–89.

Hausman, Daniel M. (1992). *The inexact and separate science of economics*. Cambridge: Cambridge University Press.

Jeffrey, Richard (1990). *The logic of decision*. 2nd ed. Chicago: University of Chicago Press.

Kahneman, Daniel, & Amos Tversky. (1979). Prospect theory: An analysis of decision under risk. *Econometrica* 47, 263–291.

Kahneman, Daniel, Paul Slovic, & Amos Tversky (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. New York: Cambridge University Press.

Kahneman, Daniel, Jack L. Knetsch, & Richard H. Thaler. (1991). Anomalies: The endowment effect, loss aversion, and status quo bias. *The Journal of Economic Perspectives* (American Economic Association) 5(1), 193–206.

Klaes, Matthias, & Esther-Mirjam Sent. (2005). A conceptual history of the emergence of bounded rationality. *History of Political Economy* 37(1), 27–59.

Lehtinen, Aki, & Jaakko Kuorikoski. (2007). Unrealistic assumptions in rational choice theory. *Philosophy of the Social Sciences* 37, 115–38.

Luce, R. Duncan, & Howard Raiffa. (1957). *Games and decisions: Introduction and critical survey*. New York: Wiley.

Mäki, Uskali. (1994). Isolation, idealization and truth in economics. In *Idealization VI: Idealization in economics*, edited by Bert Hamminga & Neil De Marchi. (*Poznan Studies in the Philosophy of the Sciences And the Humanities* 38, 7–68).

Mäki, Uskali. (2009a). Unrealistic assumptions and unnecessary confusions: Rereading and rewriting F53 as a realist statement. In *The methodology of positive economics: Reflections on the Milton Friedman legacy*, edited by Uskali Mäki, 90–116. Cambridge: Cambridge University Press.

Mäki, Uskali. (2009b). Realistic realism about unrealistic models. In *Oxford handbook of the philosophy of economics*, edited by H. Kincaid, & D. Ross, 68–98. New York: Oxford University Press.

Mas-Colell, Andreu, Michael D. Whinston, & Jerry R. Green. (1995). *Microeconomic theory*. New York: Oxford University Press.

von Neumann, John, & Oskar Morgenstern. (1947). *The theory of games and economic behavior*. 2nd ed. Princeton, NJ: Princeton University Press.

Quiggin, John (1982). A theory of anticipated utility. *Journal of Economic Behavior and Organization*. 3(4), 323–43.

Ramsey, Frank P. (1931). Truth and probability. In *The foundations of mathematics and other logical essays*, edited by Richard B. Braithwaite, 156–98. London: Routledge and Kegan Paul.

Resnik, Michael D. (1987). *Choices: An introduction to decision theory*. Minneapolis: University of Minnesota Press.

Rubinstein, Ariel. (1998). *Modeling bounded rationality*. Cambridge, MA: MIT Press.

Samuelson, Paul. (1938). A note on the pure theory of consumer behavior. *Economica* 5(17), 61–71.

Savage, Leonard. J. (1954). *The foundations of statistics*. New York: Wiley.

Sen, Amartya K. (1987). *On ethics and economics*. Oxford, Blackwell.

Simon, Herbert A. (1987). Bounded Rationality. In *The New Palgrave Dictionary of Economics*, edited by John Eatwell, Murray Milgate, and Peter Newman. London: Macmillan.

Simon, Herbert A. (1990). Invariants of human behavior. *Annual Review of Psychology* 41, 1–19.

Starmer, Chris (2000). Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. *Journal of Economic Literature*, 38, 332–82.

Todd, Peter M., & Gerd Gigerenzer. (2000). Simple heuristics that make us smart. *Behavioral and Brain Sciences 23*(5), 727–41.

Thaler, Richard H. (1992). *The winner's curse: Paradoxes and anomalies of economic life*. Princeton, NJ: Princeton University Press.

Tversky, Amos, & Daniel Kahneman (1981). The framing of decisions and the psychology of choice. *Science* 211, 453–58.

Wong, Stanley. (1978). *Foundations of Paul Samuelson's revealed preference theory: A study by the method of rational reconstruction*. London: Routledge.