Philosophy Compass 2/3 (2007): 534–563, 10.1111/j.1747-9991.2007.00074.x

Bounded Rationality

Till Grüne-Yanoff*
Royal Institute of Technology, Stockholm, Sweden

Abstract

The notion of bounded rationality has recently gained considerable popularity in the behavioural and social sciences. This article surveys the different usages of the term, in particular the way 'anomalous' behavioural phenomena are elicited, how these phenomena are incorporated in model building, and what sort of new theories of behaviour have been developed to account for bounded rationality in choice and in deliberation. It also discusses the normative relevance of bounded rationality, in particular as a justifier of non-standard reasoning and deliberation heuristics. For each of these usages, the overview discusses the central methodological problems.

Introduction

The behavioural sciences, in particular economics, have for a long time relied on principles of rationality to model human behaviour. Rationality, however, is traditionally construed as a normative concept: it recommends certain actions, or even decrees how one ought to act. It may therefore not surprise that these principles of rationality are not universally obeyed in everyday choices. Observing such rationality-violating choices, increasing numbers of behavioural scientists have concluded that their models and theories stand to gain from tinkering with the underlying rationality principles themselves. This line of research is today commonly known under the name 'bounded rationality'.

In all likelihood, the term 'bounded rationality' first appeared in print in *Models of Man* (Simon 198). It had various terminological precursors, notably Edgeworth's 'limited intelligence' (467) and 'limited rationality' (Almond; Simon, 'Behavioral Model of Rational Choice').[1] Conceptually, Simon ('Bounded Rationality in Social Science') even traces the idea of bounded rationality back to Adam Smith. Today, the term is used in various disciplines, notably economics, psychology and AI. Conceptually, however, the usage of the term often differs even within the same discipline. At least four important usages of 'bounded rationality' can be distinguished:

(i) To criticise standard theory;
(ii) To enrich behavioural models and theory;

(iii) To provide appropriate rational advice;
(iv) To explicate the concept of rationality.

This article traces these different usages, presents examples of each (mainly from economics) and discusses criticisms levelled against them. In particular, it discusses how bounded rationality differs from standard instrumental rationality (and how not), and whether it retains a bond between its descriptive and normative interpretations.

1. Standard Theory

The standard theory of individual rationality provides the backdrop against which bounded rationality is discussed. It identifies individuals as a set of well-defined preferences, and treats an action as rational if it is the one most likely to satisfy these preferences. More specifically, it treats individuals as choosing under *risk*, where outcomes of actions have a determined probability, but do not obtain with certainty. This theory, sometimes referred to as rational choice theory, has been widely accepted as the basis for various standard economic theories, and is increasingly used in other social and behavioural sciences. Its core has been axiomatised in expected utility theory (EUT). One such axiomatisation is sketched in section 1.1. One axiomatisation of probability, used in combination with EUT, is sketched in section 1.2. These are not the only axiomatisations of expected utility and probability calculus, nor do they cover all areas of rational deliberation. But they are two important hallmarks of the standard account of rationality, against which bounded rationality can be put in relief.

1.1. EXPECTED UTILITY THEORY

Expected utility theory, without doubt, is 'the major paradigm in decision making since the Second World War' (Schoemaker 529), both in its descriptive and normative interpretations. Various axiomatisations of EUT exist. For the discussion here, it will be useful to discuss one set of axioms from which EUT can be derived. These axioms were first presented by von Neumann and Morgenstern, and are still commonly referred to in economics.

Expected utility theory measures an agent's valuation of *prospects*. von Neumann and Morgenstern's theory measures the strength of a person's preference for a prospect by the risk she is willing to take to receive it. 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, the agent's consumption of goods and services are such pure prospects, as are the agent's illnesses, emotional or intellectual developments. Lotteries, also called prospects under risk, are probability distributions over events or states. For example, when consuming self-picked mushrooms, an agent faces the lottery $(X_1, p; X_2, 1 - p)$, where X_1 denotes the state (which has probability p) that she falls ill from poisoning and X_2 (with probability $1 - p$) the state that she will not. More generally, a lottery Y consist of a set of prospects $X_1, \ldots, X_n$ and assigned probabilities $p_1, \ldots, p_n$, such that $Y = (X_1, p_1; \ldots X_n, p_n)$, where $p_1 + \ldots + p_n = 1$. Obviously, the prospects $X_1, \ldots, X_n$ can be lotteries themselves.

At any time, EUT assumes that there is a fixed set of prospects $\;=\{X_1, \ldots, X_n\}$ for any agent. With respect to the agent's evaluation of these prospects, EUT 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 $\succeq$ over , which satisfies the following conditions. First, the ordering is assumed to be *complete*, i.e. either $X_i \succeq X_j$ or $X_j \succeq X_i$ for all $X_i, X_j \in$. Second, the ordering is assumed to be *transitive*, i.e. if $X_i \succeq X_j$ and $X_j \succeq X_k$, then also $X_i \succeq X_k$ for all $X_i, X_j, X_k \in$. Completeness and transitivity together ensure that the agent has a preference ordering over all prospects. Further, it assumes that $\succeq$ is *continuous*, i.e. for all X, Y, Z where $X \succeq Y$ and $Y \succeq Z$, there is some p such that the prospect $(X, p; Z, 1 - p) \sim Y$. In words, for any prospect Y, there is always a prospect X preferred to Y, and another prospect Z over which Y is preferred, such that a probability mixture of X and Z is equal in value to Y. Completeness, transitivity and continuity of $\succeq$ imply that these preferences can be represented by a numerical function $u(.)$, such that for all X, Y: $X \succeq Y \Leftrightarrow \mathrm{u}(X) \geq \mathrm{u}(Y)$. That is, an individual will prefer the prospect X to the prospect Y if and only if $u(.)$ assigns a higher value to X than to Y.

Standard EUT further assumes the independence of irrelevant alternatives. This axiom maintains that if a prospect X is preferred to a prospect Y, then a prospect that has X as one compound with a probability p is preferred to a prospect that has Y as one compound with a probability p and is identical otherwise. That is, for all X, Y, Z: if $X \succeq Y$ then $(X, p; Z, 1 - p) \succeq (Y, p; Z, 1 - p)$. If all three axioms are satisfied, preferences over lottery prospects $X = (X_1, p_1; \ldots X_n, p_n)$ are represented by a utility function:

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

such that for all X, Y:

$$X \succeq Y \Leftrightarrow \Sigma_i \, p_i \times u(X_i) \geq \Sigma_i \, p_i \times u(Y_i)$$

By this representation, it is insured that an agent who chooses what she prefers most maximises her utility $u(.)$, and vice versa.

1.2. PROBABILITY CALCULUS

Expected utility theory uses the concept of probability to characterise the situation of risk under which agents choose. Similarly as in the case of EUT, there is a widely accepted axiomatisation of probability judgements interpreted as subjective beliefs. This Bayesian axiomatisation consist of three parts.

1. The individual has a *coherent set of probabilistic beliefs*. Coherence here means compliance with the mathematical laws of probability. These laws are the same as those for objective probability, which are known from the frequencies of events involving mechanical devices like dice and coins.
   (i) $1 \geq P(p) \geq 0$
   (ii) $1 \geq P(p|q) \geq 0$
   (iii) If p is certain, then $P(p) = 1$
   (iv) If p and q are mutually exclusive, then $P(p \text{ or } q) = P(p) + P(q)$
   (v) $P(p \text{ and } q) = P(p) \times P(q|p)$
2. The Bayesian subject has a *complete set of probabilistic beliefs*. In other words, to each proposition he assigns a subjective probability. A Bayesian subject has a (degree of) belief about everything. Therefore, Bayesian decision-making is always decision-making under certainty or risk, never under uncertainty or ignorance.
3. When exposed to new evidence, the Bayesian subject *changes his (her) beliefs in accordance with his (her) conditional probabilities*. Conditional probabilities are denoted $p(\ |\ )$, and $p(A|B)$ is the probability that A, given that B is true. (As usual, $p(A)$ denotes the probability that A, given everything that you know.)

2. Bounded Rationality Phenomena

A massive amount of evidence has been gathered that seems to show that people often violate predictions derived from these two axiomatic systems. More importantly, individual as well as group behaviour seems to indicate that people *systematically* violate them. The most interesting and convincing of these behavioural phenomena have been tickled out by increasingly sophisticated experimental designs. Others are not identified by experiments, but are indirectly inferred from anomalies of economic theories.

The term 'bounded rationality' is often employed to denote any evidence of the deficiencies of the standard models. Presumably, standard models are thought of as assuming 'full' rationality, so that evidence against them is considered evidence for some not necessarily further specified 'bound' on rationality. The term 'bounded rationality' can of course be used this way, but needs to be clearly delineated from its other uses to avoid confusion. It should therefore be stressed that this usage of 'bounded rationality' refers to a set of *phenomena*, and possibly also to the scientific practices of their elicitation. It refers to research that provides us with 'a large body of empirical material that provides a rich qualitative description of the phenomena' (Simon, 'Bounded Rationality in Social Science' 349). Of philosophical interest here are the methods by which these phenomena are elicited, the types into which they are categorised, and the arguments that justify criticism of standard theory on their basis.

2.1. BOUNDED RATIONALITY AS PHENOMENA OF INDIVIDUAL BEHAVIOUR

The axioms of EUT and probability calculus, if used in a predictive theory, certainly have empirical content. However, testing them is not trivial. It requires clever experimental set-ups to identify behaviour that violates the axioms. Consequently, the discovery of bounded rationality phenomena has moved more and more from the philosopher's armchair to the behavioural laboratory.

Bounded rationality phenomena can be categorised according to the hypotheses that they violate. Examples of behavioural phenomena that seem to violate EUT include Allais' Paradox and the preference reversal phenomenon. Allais' idea was to find two pair-wise choices such that EUT would predict a specific choice pattern, and then check the prediction in the laboratory. The choice experiment designed, by Kahneman and Tversky ('Prospect Theory') according to Allais' idea, looks as follows (prizes were in Israeli pounds):

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

EUT predicts that agents will choose *C* if they have chosen *A* (and vice versa) and that they will choose *D* if they chose *B* (and vice versa). To see this, simply re-partition the prizes of the two problems as follows:

Choice problem 1⋆ – choose between:

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

Choice problem 2⋆ – choose between:

C⋆:	2500	with probability 0.33	*D*⋆:	0	with probability 0.66
	0	with probability 0.66		2400	with probability 0.34
	0	with probability 0.01			

Note that the prize '2400 with probability 0.66' occurs in both options *A*⋆ and *B*⋆ of the first choice problem. According to the independence axiom, this prize is irrelevant for the options evaluation. Further, substituting any other prize for it in both options will not – according to independence – affect the options evaluation. Substituting '0 with probability 0.66' for '2400 with probability 0.66' in both *A*⋆ and *B*⋆ will therefore not affect their evaluation. Hence *A* and *C* and *B* and *D* are evaluatively equivalent, according to EUT. However, in sharp contrast to this prediction, in an experiment involving 72 people, 82% of the sample chose *B*, and 83% chose *C*.

The preference reversal phenomenon occurs when people are offered choices between pairs of gambles, and consecutively state their reservation price for these gambles. Grether and Plott (1979) designed an experiment that controlled for many potentially disturbing factors. They drew up a menu of six pairs of gambles, such that one gamble in each pair would have a high probability of winning a small monetary prize, the other a low probability of winning a bigger monetary prize, and both include the possibility of a small loss. For example, a pair would consist of ($4, 35/36; –$1, 1/36) and ($16, 11/36; –$1.50, 25/36). In the first phase of the experiment, subjects were asked to choose a gamble from each pair on the menu. In the second phase, the same subjects were asked to state the smallest price they were willing to sell each gamble on the menu for (a special set-up insured that they revealed their actual reservation price).

Expected utility theory predicts that agents assign a higher reservation price for gambles they prefer. By definition, they are indifferent between obtaining the reservation price (and not playing) and playing the gamble (and not receiving the reservation price). If they prefer one of a pair's gambles over another, then by transitivity they prefer obtaining the reservation price for the one gamble (and not playing it) over obtaining the reservation price for the other gamble (and not playing it). Thus the reservation price of the former gamble must be higher than that of the latter. However, in sharp contrast to this prediction, of 99 choices of the high-probability-small-prize gambles, only 26 were consistent with the announced selling price. In 70% of the choices, agents stated a reservation price that was smaller for the gamble that they chose than for the gamble that they did not choose.

Examples of behavioural phenomena that seem to violate Bayesian probability calculus include experiments about people's use of base-rate information in making probabilistic judgements. According to the familiar Bayesian account, the probability of a hypothesis depends, in part, on the prior probability of the hypothesis. However, in a series of elegant experiments, Kahneman and Tversky ('On the Psychology of Prediction') showed that subjects often seriously undervalue the importance of prior probabilities. One of these experiments presented half of the subjects with the following 'cover story'.

> A panel of psychologists have interviewed and administered personality tests to 30 engineers and 70 lawyers, all successful in their respective fields. On the basis of this information, thumbnail descriptions of the 30 engineers and 70 lawyers have been written. You will find on your forms five descriptions, chosen at random from the 100 available descriptions. For each description, please indicate your probability that the person described is an engineer, on a scale from 0 to 100.

The other half of the subjects was presented with the same text, except the 'base-rates' were reversed. They were told that the personality tests had been administered to 70 engineers and 30 lawyers. Some of the descriptions that were provided were designed to be compatible with the subjects' stereotypes of engineers, though not with their stereotypes of lawyers. Others were designed to fit the lawyer stereotype, but not the engineer stereotype. And one was intended to be quite neutral, giving subjects no information at all that would be of use in making their decision. Here are two examples, the first intended to sound like an engineer, the second intended to sound neutral:

> Jack is a 45-year-old man. He is married and has four children. He is generally conservative, careful and ambitious. He shows no interest in political and social issues and spends most of his free time on his many hobbies which include home carpentry, sailing, and mathematical puzzles.
>
> Dick is a 30-year-old man. He is married with no children. A man of high ability and high motivation, he promises to be quite successful in his field. He is well liked by his colleagues.

As expected, subjects in both groups thought that the probability that Jack is an engineer is quite high. Moreover, in what seems to be a clear violation of Bayesian principles, the difference in cover stories between the two groups of subjects had almost no effect at all. The neglect of base-rate information was even more striking in the case of Dick. That description was constructed to be totally uninformative with regard to Dick's profession. Thus, the only useful information that subjects had was the base-rate information provided in the cover story. But that information was entirely ignored. The median probability estimate in both groups of subjects was 50%.

Further phenomena in this category include people violating probabilistic independence, neglecting the importance of sample sizes and exaggerating confirming over disconfirming evidence relative to initial beliefs.

Last, there are important phenomena of people violating other aspects of standard rational choice and reasoning accounts. These include making errors in deductive inferences, ignoring relevant information, using irrelevant information, displaying overconfidence in probability judgements, and making false inferences about causality. Overviews of these and many more phenomena and their elicitation in experiments can be found in Arkes and Hammond, Hogarth, Kahneman, Slovic and Tversky, or Nisbett and Ross.

The claim that these experimental observations reveal phenomena relevant for behavioural theories based on the rationality concept has been criticised from various angles. The first criticism suggests that there are various potential defects in the experimental techniques. Take for example the experiment concerning the use of base-rate information. Schwarz et al. (1991) argue that the subject's behaviour depends on their interpretation of the communicative intention of the experimenter. The task description invites participants to consider it as a request to exhibit psychological intuition about what the experimenter is trying to communicate, rather than to perform a disengaged computation of the information that is actually presented. In his overview of experiments of this sort, Koehler points out that even though the difference in cover stories between the two events has *almost* no effect, the small effect it does have is systematic, and hence shows that probability judgements *are* sensitive to base rates. If any of these allegations were correct, the results of such an experiment would be artefacts of that defect, not phenomena behavioural scientists are interested in. This, however, is a question that can be directed only at the specific experiment, and thus is difficult to discuss from a philosophical point of view.

But even without any 'defects', it has been questioned whether experimental observations capture anything relevant for our understanding of the real world. Behavioural models, most scientists would agree, are supposed to be applicable to real economies, not to the 'artificial' conditions implemented in the economic lab. But what guarantees, so the criticism goes, that the observations made in an experiment can be generalised to behaviour in the real world? This question of the *external validity* of experiments has indeed not been sufficiently discussed by methodologists, and no definite answer has been found to meet the criticism. Guala and Mittone make an interesting suggestion. They point out that in many cases, experimenters do not seek to prove external validity of their experiments. This does not make these experiments redundant, though, because experiments are not exclusively designed to test theories that make claims about the real world (Roth). Experiments also are designed to contribute to the 'library of phenomena that the applied scientists will borrow and exploit on a case-by-case basis' (Guala and Mittone 511), and the construction of such a collection does not require so much external validity but robustness of the phenomena elicited. Such libraries, similar to a chemist's cabinet of synthesised substances, may help understanding specific behavioural cases at some future point in time.

One attempt to strengthen the argument for the limited relevance of behavioural experiments has been to distinguish failures of competence from performance errors, and to claim that experiments can only identify behavioural phenomena of the latter kind (Cohen is an early advocate of this approach). Such 'cognitive illusions' are then often argued to be inherently unstable and lacking exactly the robustness that would make them interesting candidates for a library of phenomena. People may thus be boundedly rational, but they learn to achieve optima through experience, and any boundedly rational behaviour will disappear in the long run (if the environment remains stable). A reply to such a claim, however, can simply point out the two conditionals on which it is based: that the long run is not 'too long', and that the environment in which people behave is sufficiently stable for them to effectively learn to obtain optima. There is little reason to believe that these two 'ifs' are always true; hence identifying bounded rationality phenomena through experiments may well enhance our understanding of the real world at least in some cases.

Last, a common reply to the results from behavioural experiments has insisted on the irrelevance of bounded rationality phenomena on the aggregate level. Such phenomena may exist on the individual level, but for the aggregate level (in particular for market phenomena) they can be neglected. Haerle and Kirman present an interesting example of a market that exhibits the standard downward sloping demand curve (hence representable by a utility function of a representative agent), while its individual members do not at all exhibit this property. This claim is sometimes supported by a pragmatic argument. Individuals whose rationality is bounded loose money, which allows rational agents to take over wealth and dominate the overall market. Even if a substantial number of individuals' rationality is bounded, market forces provide strong monetary incentives for rational decisions. For market outcomes to be efficient, moreover, it is sufficient if only some agents act rationally and exploit arbitrage possibilities. Hence bounded rationality phenomena allegedly are not relevant for aggregate behaviour. This theoretical argument fails on two grounds, as Conlisk argues. First, boundedly rational agents may well survive on lower, more wasteful level, and then will continue participating (and influencing) aggregate outcomes. Second, boundedly rational agents can prevent to be exploited by tricksters 'by such simple devices as slamming the door and hanging up the phone' (648) – without ever reaching full rationality.

2.2. BOUNDED RATIONALITY AS PHENOMENA OF GROUP BEHAVIOUR

Not only is the theoretical arguments against the occurrence of bounded rationality phenomena on the aggregate level be flawed, but there is plenty of aggregate behavioural data that cannot easily be explained by EUT and Bayesian probability calculus. Camerer ('Prospect Theory in the Wild') lists phenomena from a wide range of applied economics topics, from the macroeconomic fact that consumers do not cut consumption when they get bad income news, through consumer behaviour that is more sensitive to price increases than cuts, to the finance anomaly that stock returns are too high relative to bond returns. Surprisingly, the stock market is a particularly rich source of interesting bounded rationality phenomena. Odean, for example, investigates whether investors weight losses more heavily than gains. His empirical study finds that around 15 per cent of all gains are realised by investors, but only ten per cent of all losses. This behaviour comes at an economic cost; it is therefore surprising, given that investors face strong monetary incentives to make decisions in accordance with EUT. A host of other anomalies has been identified in financial markets. For an overview, see Shleifer.

All of these observations, however, have to be taken with a grain of salt. These phenomena are identified through the anomaly of some economic theory (a good collection of such anomalies is found in Thaler). They are not only dependent on EUT, but also on other assumptions made in the theories of the respective fields. That certain aggregate data is an anomaly to a specific theory that is based on EUT is therefore at best indirect evidence for the existence of a bounded rationality phenomenon. This criticism will be continued in section 4.

3. Bounded Rationality as a Menu of Modelling Assumption

Economics is a largely model-based science. It develops models of concepts or phenomena from axiomatic assumptions. Models of rational choice that take into account limitations of human capacities are often grouped under the label of 'models of bounded rationality'. Typical assumptions pertaining to the limitations of human cognitive abilities are (qtd. in Simon, 'Bounded Rationality in Social Science' 25):

- Limited knowledge of the world;
- Limited ability to evoke this knowledge;
- Limited ability to work out consequences of actions;
- Limited ability to conjure up possible courses of action;
- Limited ability to cope with uncertainty;
- Limited ability to adjudicate among competing wants.

Models that include such assumptions present only a small subset of the totality of economic models. Nevertheless, they can be found in many economic sub-disciplines. Here, only a few areas can be addressed. For a wider 'sampler' of bounded rationality in economic models, see Conlisk; for an in-depth presentation of a selection of models, see Rubinstein. The following three may serve to illustrate how bounded rationality assumptions are often picked for the sub-disciplines' specific needs, rather than for a general goal of 'realism'.

Transaction cost economics aims at explaining the existence, size, structure and workings of organisations. In particular, it strives to account for the particular structure of a firm, most importantly, the extent of its vertical integration. As its main explanans function specific market failures, e.g. the non-availability of full information to all parties or imperfect competition. Departures from this perfection can result in firms incurring costs when they attempt to buy or sell goods or services. For example, lack of information about alternative suppliers might lead to paying too high a price for a good. Lack of information about a customer's creditworthiness might result in a bad debt. These are *transaction costs.* Research in this area was pioneered by Oliver Williamson, who explicates transaction costs as arising from agent's limited cognitive abilities: 'Economizing on transaction costs essentially reduces to economizing on bounded rationality' (110). The bounded rationality assumption makes complete contracting infeasible because not everything can be known and there are limits to the capabilities of decision makers for dealing with information and anticipating the future.

Despite this connection, Williamson's transaction cost economics remains squarely within the limits of standard economic theory. Its models assume that firms are profit maximizing, and that profit maximization involves costs minimization. These assumptions are crucial for the explanatory project: Williamson argues that firms minimize their total costs (made up of both production and transaction costs), and that this cost-minimisation drives the decision whether transaction takes place in an open market, or whether transactions are coordinated in an institutional structure. Transaction cost economics differs from standard economic theory only in its emphasis on transaction costs as distinct from production costs, and in its link between transaction cost and the boundedness of human rationality. Assuming bounds on rationality *only* with respect to transaction cost, Williamson's approach identifies an important potential explanans of institutional structure while retaining the ability to employ his models within the framework of standard economic theory.

Various strands of *game theory* also employ bounded rationality assumptions. Standard game theory assumes that agents have foresight, so that they can deliberate about all of time from the beginning of play. Some authors have proposed to eliminate the foresight assumption and replace it with a learning dynamics. Kalai and Lehrer, for example, show that in an infinitely repeated game, subjective utility maximisers will converge arbitrarily close to playing Nash equilibrium. The only rationality assumption they make is that players maximise their expected utility, based on their individual beliefs. Knowledge assumptions are remarkably weak for this result: players only need to know their own payoff matrix and discount parameters. They need not know anything about opponents' payoffs and rationality; furthermore, they need not know other players' strategies, or conjectures about strategies. Knowledge assumptions are thus much weaker for Nash equilibria in infinitely repeated games than those required for one-shot game Nash solutions or rationalisability. To obtain these interesting results, however, strong assumptions have to be made about the agent's cognitive capacities – which therefore is hardly in line with a general bounded rationality approach. In contrast, considerations of boundedly rational in strategic interactions (for example at the end of Selten) have not been formally developed and have not achieved a lot of attention within the discipline.

Another approach of bounded rationality modelling in game theory restricts the available strategies. Rubinstein formalises this idea by extending the analysis of a repeated game to include players' sensitivity to the *complexity* of their strategies. He restricts the set of strategies to those that can be executed by finite machines. He then defines the complexity of a strategy as the number of states of the machine that implements it. Each player's preferences over strategy profiles increase with her payoff in the repeated game, and decrease with the complexity of her strategy's complexity (He considers different ranking methods, in particular unanimity and lexicographic preferences). Rubinstein shows that the set of equilibria for complexity-sensitive games is much smaller than that of the regular repeated game.

Interestingly enough, Aumann includes the evolutionary stable strategy approach in the bounded rationality fold, although he in the next sentence admits that 'no rationality is required at all to arrive at a Nash equilibrium; insects and even flowers can and do arrive at Nash equilibria' (4). To match this theory, standard game models have to be considerably reinterpreted. Instead of individual players (with more or less limited cognitive capacities), it is types that interact, characterised by their respective strategy. Interaction results in changes in the number of offspring for each type. A certain proportional division between the two types constitutes a Nash equilibrium if and only if the type proportions remain constant from generation to generation. The evolutionary reinterpretation of game theory has become very popular, because it frees the models from controversial rationality assumptions. However, it is not clear how much room it leaves for bounded rationality assumptions either.

The idea of *rational expectations* is to model agents forecasting abilities on the knowledge of economist modelling them. This strive for symmetric models between the forecasting abilities of economic agents and economists has been criticised from various angles. Quite obviously, it is open to criticism from a bounded rationality view: rational expectations are very costly to achieve and maintain, as they involve knowledge of an entire market or economy. But even measured by its own criterion, standard rational expectations models are sometimes considered to be deficient. Sargent has argued that when implemented numerically or econometrically, rational expectations models bestow more knowledge on the agent within the model than what is possessed by an econometrician, who faces estimation and inference problems that the model-agent was supposed to have solved. To restore the symmetry, Sargent suggests limiting the assumed knowledge of the agent and instead introducing some form of learning. Research then focuses on whether a specific kind of learning – adaptive learning – might lead rational agents to rational expectations. This depends on a number of conditions, many of which again require relatively strong assumptions about the agent's rationality, in particular concerning belief formation and updating.

Two diametrically opposed criticisms have been brought against bounded rationality models. The one, spearheaded by Friedman's extremely influential article, rejects modelling the actual cognitive procedures of boundedly rational agents in favour of models that predict how agents behave 'as if' they performed the computations assumed in standard rationality models. All that matters is whether the model predicts well, not whether the mechanism the model uses is a realistic representation of the agent's cognitive make-up. In this simplified form, the criticism can be easily rejected, as it depends on the predictive success of the standard models. As this is often deficient, alternative assumptions, also from the bounded rationality menu, may and often have shown to lead to superior predictions. However, not all economic models are built with the purpose of prediction. Some rather serve as tools for conceptual exploration (Hausman 1992). Often, such models are interpreted as counterfactual worlds, in which assumptions and hypotheses can be tested similar to a thought experiment. For those kinds of models, predictive success is not of prime relevance. Hence the better-prediction rationale for introducing bounded rationality assumption into these models disappears. For conceptual explorations, the introduction of a bounded rationality assumption must somehow enhance our understanding of the modelled counterfactual world. The 'as if' criticism has a point insofar as 'realism' in models is not a goal in itself; rather, for predictive models it must improve predictive success, and for conceptual-exploration models it must facilitate our understanding.

While the first criticism has been mounted as a defence of standard economic models, the second criticises bounded rationality models for not deviating sufficiently. In each of the brief presentations of such models above it was noted that the respective sub-disciplines only used those assumptions that would suit their needs, while disregarding others. This has led to the 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, 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' (Rubinstein 190).

Instead of constructing models from the armchair, a more general theory is required that addresses why bounded rationality assumptions are relevant in certain contexts and not in others, and that is testable and tested. Only on the basis of such a theory can a more principled model construction involving bounded rationality assumption take place.

4. Theories of Bounded Rationality

Any theory of bounded rationality that deserves its name should provide an account of (i) what the limits of human rationality are; (ii) what effects these limitations have on the process and the outcome of deliberation; and (iii) why these effects are realised in some contexts and not in others. Unfortunately, no unified theory exists that delivers these three desiderata, and it is not clear that any such unified theory will ever exist. Instead, partial, and sometimes competing, theories have been offered for the different questions. It is therefore apt to speak of *theories* of bounded rationality. This section discusses and categorises some of these efforts.

Simon's original account proposes two kinds of limitations on agent's 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' ('Invariants of Human Behavior' 7). In subsequent research, the second aspect has received the overwhelming amount of attention, so it will be discussed here first.

If theories of rational decision making, like EUT and Bayesian probability calculus discussed in section 1, are interpreted as theories of human deliberation processes, they assume that the deliberating agents have enormous cognitive capacities. For example, utility maximisation requires that agents have preferences over all outcomes, and probability updating requires that agents know the priors, as well as conditional probabilities, of all propositions. This requires an astounding memory capacity. Further, the operations of maximisation and updating are performed over these extensive memory contents, and hence require extraordinary computational capacities.

Three arguments have been proposed *against* assuming such strong cognitive capacities on human agents, and *for* assuming more limited 'human-size' capacities. Two of these are armchair arguments; the third is based on the empirical findings discussed in section 2.

The first armchair argument against utility maximisation as a cognitive procedure is that such maximisation is not *computable* in some cases. Binmore discusses this issue with respect to non-cooperative games. For this purpose, he casts both players in a two-player game as Turing machines. A Turing machine is a theoretical model that allows for specifying the notion of computability. Very roughly, if a Turing machine receives an input, performs a number of computational steps (which may be infinite), and gives an output, then the problem is computable. If a Turing machine is caught in an infinite regress while computing a problem, however, then the problem is not computable. The question Binmore discusses is whether Turing machines can play and solve games. The scenario is that the input received by one machine is the description of another machine (and vice versa), and the output of both machines determines the players' actions. Binmore shows that a Turing machine cannot predict its opponent's behaviour perfectly and simultaneously participate in the action of the game. Roughly put, when machine 1 first calculates the output of machine 2 and then takes the best response to its action, and machine 2 simultaneously calculates the output of machine 1 and then takes the best response to its action, the calculations of both machines enter an infinite regress. Attempts to outguess someone else by predicting her behaviour as a rational reply to one's own behaviour is thus not computable in this sense.

Computational impossibility, however, is very far removed from the realities of rational deliberation. Take for example the way people play chess. Zermelo long ago showed that chess has a solution – however, nobody has been able to say up to now what the solution is. While this may be fortunate for chess enthusiasts (finding the solution, after all, would make chess trivial), it is worthwhile to reflect for a moment how the practice of chess play shows how deeply ingrained the concept of bounded rationality is. Nobody expects a chess player to find the solution. The game theoretic concept of a strategy – a list that specifies the agent's choice for each of her possible choices – is completely preposterous when applied to chess. Someone who would try herself at such an approach would not only be bound to fail miserably, but also would be regarded as an utter fool.

This is where the second armchair argument comes in. It identifies cognition as *costly*, and sees agents as using those cognitive processes that are low in cost but high in effectiveness.[2] In chess, players do not calculate the solution of the game and choose their strategies accordingly, because such a procedure would be too costly. Instead, it seems that they typically 'check out' several likely scenarios and that they entertain some method to evaluate the endpoint of each scenario (e.g. by counting the chess pieces). But people differ in the depth of their inquiry, the quality of the 'typical scenarios' selected, and the way they evaluate their endpoint positions, so the idea goes, because they face different computation costs: what may be easy for a grand master to oversee and calculate may be too difficult and exhausting for an amateur.

Simon ('Behavioral Model') devised *satisficing* as an alternative decision procedure to that of optimisation. The satisficing agent examines one potential action at a time (e.g. one chess move). She evaluates the action according to a simplified payoff function. In the extreme, such a payoff function only distinguishes two values: satisfactory and unsatisfactory. In her serial evaluation of actions, the agent will choose the first that she evaluates satisfactory. The simple satisficing account has some plausible applications. A chess player who finds an alternative that leads to a forced mate of his opponent, as Simon points out, will generally adopt the alternative without worrying about whether another alternative also leads to a forced mate. But a forced mate, the agent knows, is the best that she can achieve. That again may be too much to ask. How does the agent set her aspiration level where the best cannot be achieved, or may not even be known? Two answers have been given in the literature. The one tries to account for aspiration level setting as an adaptation to the environment. It will be discussed below in this section. The other tries to account for aspiration level setting as an optimization under constraints. As a descriptive theory, this is problematic: the motivation to look for an alternative was that EUT, as a deliberation procedure, appeared too costly in its computational requirements. But the computational economy of the alternatives may well be illusionary, if the choice of an alternative decision procedure requires an analysis of the costs and benefits expected from that alternative.

The two armchair arguments thus indicate potential cognitive limitations, but give inconclusive answers as to what the effects of cognitive limitations are, or under which circumstances they are to be expected. For answers to those questions, one must turn to theories that try to incorporate established empirical phenomena. Because of the historical importance and the normative acclaim of EUT, such theories generally take their departure from the EUT framework. Depending on the extent to which they honour the EUT assumptions, they can be divided into two categories (Starmer). *Conventional* non-expected utilities generally are non-procedural theories that satisfy stochastic dominance (also called monotonicity). *Non-conventional* non-expected utilities generally are procedural theories that do not satisfy stochastic dominance.[3] Stochastic dominance says that 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. It is a very plausible property; hence some take it as the benchmark that any theory of rationality – including bounded rationality – has to observe.

An important family of conventional non-expected utility theories focuses on agents' probability misperception. Outcomes of an action are assumed to have objective probabilities, but individuals subjectively distort objective probabilities by weighing the utility of outcomes by subjective decision weights. Early versions of this theory conceived these weights to violate the standard probability axioms, but this implies that the resulting subjectively weighted lotteries are not stochastically independent. Instead, rank-dependent expected utility theory (RDU, Quiggin) has become the most popular 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:

$$RDU(p_1, \mathrm{X}_1; \ldots p_n, X_n) = \Sigma\, \pi_j \times \mathrm{u}(\mathrm{X}_j)$$

where the probability weight π_{i} of an outcome X_i depends on its probability *and* the ranking position of the outcome:

$$\pi_j = w(p_1 + \ \ldots\ + p_j) - w(p_1 + \ \ldots\ + p_{j1})$$

The intuition behind the theory is that how much attention agents pay to an outcome depends not only on the probability of the outcome, but also on the favourability of the outcome in comparison to other possible outcomes (Diecidue and Wakker). 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 a 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 stochastic dominance. Still, it accounts for agents' behaviour in the Allais' paradox: a lot of people are pessimist, and pessimists choose *B* because they overemphasise the possibility in *A* of not winning, while this does not make an important difference in comparison between *C* and *D*.

Another important family of conventional non-expected utility theories focuses on agents' aversion of regret. When choosing between different lotteries, agents can determine for each state of the world what they could have won if they had chosen differently. For example, comparing lotteries (1/3, 0; 1/3, 50; 1/3, 100) and (1/3, 100, 1/3, 0, 1/3, 50), a regret averse agent will choose the second, because the first offers the possibility of a state where she would win nothing when she could have won 100. More generally, if the outcome of a prospect is worse than expected, a sense of disappointment will be generated. An outcome better than expected will generate a feeling of elation. The value function then looks as follows: $V(\mathbf{L}) = \Sigma_i\, p_i \times [u(X_i) + D(u(X_i) - U)]$, where U is a measure of prior expectation and D a disappointment/elation measure (Loomes and Sugden). Trivially, regret theory satisfies stochastic dominance, and still accounts for some deep violations of EUT. In particular, it gives an account why people choose B and C in Allais' paradox (A harbours potential regret when compared to B, but C differs little in this way from D).

These conventional theories stay within the standard methodology of 'as if' models. Even when they try to account for deviations from EUT rationality, they do not hypothesise how the actual procedures work. Their frameworks only attempt to systematise behavioural phenomena in a plausible and parsimonious way. In contrast, non-conventional theories often attempt to spell out how agents *actually* reason and deliberate. Simon formulated the basic assumption of any such theory

> Agents use selective heuristics and means-ends analysis to explore a small number of promising alternatives. They draw heavily upon past experience to detect the important features before them, features that are associated in memory with possibly relevant actions. They depend upon aspiration-like mechanisms to terminate search when a satisfactory alternative has been found. ('From Substantive to Procedural Rationality' 136)

Some authors have argued that only the work on these theories properly belongs to the bounded rationality research program (see e.g. Cozic). As we have shown, the actual use of the term is much wider. Additionally, we think that the connections between the behavioural experimental research, the new style of economic model building and the development of conventional and non-conventional decision theories is too close to merit such a rigid terminological distinction.

The most discussed non-conventional non-expected utility theory over the last 20 years is Kahneman and Tversky's prospect theory. The theory treats the deliberation process as divided into two stages, editing and evaluation. In the first, the different choices are ordered following a variety of heuristics so that the evaluation phase is simpler. The evaluation of prospects starts from a *reference point*. Seen from this point, prospects below it are interpreted as losses, prospects above it as gains. The value function (sketched in Fig. 1) 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.

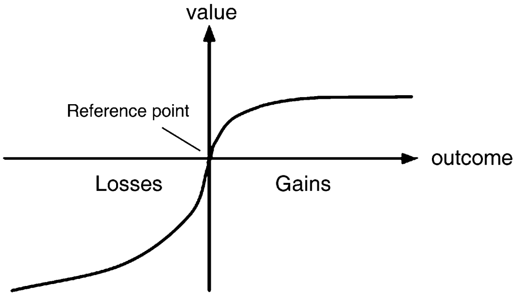

Fig. 1. The value function of prospect theory.

Kahneman and Tversky ('Prospect Theory') interpret these two properties as *diminishing sensitivity* and *loss aversion*. Diminishing sensitivity states 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 value functions increased steepness 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 function similar to rank-dependent expected utility theory. Cumulative prospect theory is therefore closer to conventional non-expected utility theories, but the editing phase distinguishes it as a procedural theory.

Experimental tests seem to show that prospect theory is a potent explanatory tool (Harless and Camerer). It also offers better and less *ad hoc* explanations of field data (Camerer, 'Prospect Theory in the Wild'). It is true that making special assumptions about the agents' motivations would allow for EUT to account for these phenomena, too. But as Camerer remarks, such modifications 'are truly ad hoc because a special modification to expected utility theory is needed for each phenomenon, which leads to an applied version of expected utility theory which is crusted with special features like a boat's hull is crusted with barnacles' (Camerer, 'Bounded Rationality' n5).

In addition to limitations on cognitive capacities, other, non-cognitive, limitations on the deliberation process are now sometimes included under the heading of bounded rationality. Some authors emphasise the numerous effects of emotions on behaviour (see Loewenstein and Lerner for an overview), or they stress the importance of intuitions. Kahneman in particular has emphasised that 'most judgments and most choices are made intuitively, [and] that the rules that govern intuition are generally similar to the rules of perception' (1450), and that intuition operates without conscious search or computation. He concludes that the

> central characteristic of agents is not that they reason poorly but that they often act intuitively. And the behaviour of these agents is not guided by what they are able to compute, but what they happen to see at a given moment. (1469)

Discussions about emotions and intuitions as limitations on rationality distinct from cognitive capacities, however, have just recently gained popularity. It remains to be seen how much attention will be paid to them.

The other blade of the pair of scissors, the structure of the environment, has attracted far less attention as a limiting factor on human rationality than the cognitive limitations discussed so far. Simon, in his article 'Rational Choice', tried to show with a simple model how an organism in its environment could survive with very simple, but well-adapted mechanisms of perception and choice. In later papers, however, Simon expressed doubts about the separate relevance of the environment:

> It is precisely when we begin to ask *why* the properly motivated subject does nor behave in the manner predicted by the rational model that we recross the boundary again from a theory of the task environment to a psychological theory of human rationality. The explanation must lie inside the subject: in limits of his ability to determine what the optimal behavior is, or to execute it if he can determine it. (Newell and Simon 54–5; also cf. Simon, 'From Substantive to Procedural Rationality' 147)

Nevertheless, a strand of recent research has focussed on establishing the dependence of computational or mental factors on environment pressures, and on how environmental forces have selected simple heuristics for making decisions. The underlying concept of rationality differs substantially from any optimisation effort. Instead, the decision maker adapts the use of his choice rule – EU calculations, satisficing, lexicographic choice rule – to the environment he lives in. 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. Ecological rationality, as it is called, 'suggests looking outside the mind, at the structure of environments, to understand what is inside the mind' (Gigerenzer 'The Adaptive Toolbox' 39). How human agents choose in particular situations then crucially depends on their evolutionary history and the rules that they were fitted with during this time. Internal cognitive architecture and their limits, on the other hand, are not so important for this approach.

While the adaptive programme is a fruitful way to explain observed phenomena, prediction is much harder. It requires an understanding of the cognitive mechanisms that resulted from adaptation to previous environments to assess the trajectory by which it will adapt to a new environment. To build a good case for such a prediction is as difficult as it is easy to tell an adaptive story about an observed phenomenon.

All of these theories are subject to two quite general criticisms. First, it is clear that none of these theories capture human behaviour perfectly.

Moving away from EUT at best gives us theories that are a little less false. So the question arises whether the increased predictive and explanatory potential of the bounded rationality theories is great enough to offset the undeniable decrease of parsimony of the new theories when compared to EUT. Formalising this trade-off to facilitate proper theory choice is an important task that requires more attention (for a good example, see Harless and Camerer).

Second, the question arises why these theories insist on rationality, now in bounded form, at all. Why not just construct a causal theory of behaviour, which allows any kind of irrationality, if needed? Some of the discussed researchers would without doubt agree: theirs is a descriptive project, and the concept of (bounded) rationality is useful only as a heuristic assumption. Few would today still insist that there is something irreducibly normative about behavioural explanation (for arguments against such a position, see Grüne-Yanoff). But still there is an argument to stick to the concept of rationality, derived from pragmatic considerations of scientific methodology.

Some behavioural theories are self-defeating. Because they correctly describe the deliberation on which past behaviour was based, they may provide incentives to the deliberating agents to change their deliberations for future behaviour. A notorious pessimist's choice between lotteries, for example, can be explained by rank-dependent utility theory. Explaining the agent's choice to her in terms of this theory involves clarifying to her the distinction between objective probability and subjective, irrational decision weights. This, in turn, may give the agent an incentive to evaluate future lotteries without her pessimistic bias. Because it was a correct account of past events, the theory caused itself to become incorrect: the next time, the now cured pessimist may choose in accordance with EUT. A theory that is based on more demanding notions of rationality, in contrast, is not self-defeating. It may not be a correct account of how agents choose now, but it will never cause its own incorrectness. Hence the question is whether the purposes for which the theory is used demand rather a momentary correctness at the risk of self-defeat, or whether a long-term stability of the theory is of more advantage for the theory users.

5. Bounded Rationality as Normative Principles

Expected utility theory and Bayesian probability calculus are often interpreted normatively. Such an interpretation sees them as prescribing how decisions should be made in order to be rational. In addition, EUT and Bayesian probability calculus have been defended against phenomena like those discussed in section 2 by pointing out that their real purpose lies in providing such normative guidelines.

However, bounded rationality does not only attack the standard theories of rationality on descriptive, but also on normative grounds. Already in 1955, Simon proclaimed his

> great doubts whether this schematized model of economic man provides a suitable foundation on which to erect a theory – whether it be a theory of how firms *do* behave, or of how they 'should' rationally behave. ('Behavioral Model' 99)

Accordingly, theories of bounded rationalities are often intended to provide normative principles and prescriptions, as well as descriptively accurate frameworks.

With respect to their normative function, two projects need to be distinguished. The *normative project* concerns itself with the question how humans should deliberate and choose in specific situations, what the normative standards are and which procedures most effectively satisfies these standards. The *evaluative project*, in contrast, asks whether people accord to the normative standards, and hence, ultimately, whether humans are rational. Particularly over the last project, there has been a lively dispute amongst partisans of bounded rationality. While advocates of the heuristics and biases program expressed a more pessimistic opinion about human rationality, evolutionary psychologists have insisted that human beings by and large are rational. As the analysis of Samuels et al. ('Ending the Rationality Wars') of both research traditions shows, however, these 'rationality wars' are mainly based on mutual misunderstanding and lack a deeper division. This article will not discuss the evaluative project any further, but instead concentrate on the underlying normative project.

Before discussing the normative standards, the target of these standards needs to be clearly identified. In particular, one needs to distinguish rationality as an attribute of choice from rationality as an attribute of deliberation processes. As an attribute of choice, rationality only makes claims how an agent should choose, and does not say how an agent should deliberate (or only says so because certain deliberation rules affect the rationality of choice). Rationality characterised this way is called *consequentialist* rationality.

On the other hand, there is a notion of rationality that evaluates the reasonability of a thought process itself. Such a process is reasonable if it adheres to certain rules, set by a normatively accepted theory:

> to be rational is to reason in accordance with principles of reasoning that are based on rules of logic, probability theory and so forth . . . principles of reasoning that are based on such rules are normative principles. (Stein 4)

Rationality characterised this way is called *deontological* rationality (for more on this distinction, see Samuels et al., 'Reason and Rationality').

Ideally, of course, the two standards would support each other. Deliberation rules would be rational because they lead to rational choice, and choices would be rational because they are guided by rational deliberation rules. Mutual support, however, must not turn into circularity. There must be a way to determine externally what rationality means, and it also must justify why this rationality is relevant as a norm. Here, consequentialist approaches seem to currently have the upper hand. They can argue that good reasoning is reasoning that tends to result in the possession of things that we value, like for example in obtaining true beliefs and avoiding false ones, or in efficiently satisfying one's personal goals and desires. Deontological approaches, on the other hand, have difficulties showing that there can be cases where it is more rational to do what does not further the attainment of our desirable ends. It then remains unclear why one should *ultimately* rely on principles, and not values, in explicating rationality. Instead, judging procedures to be rational may depend on judging the outcomes of procedures to be rational.

This brings us back to EUT and Bayesian propositional calculus. If it is rational to choose as if one maximised one's expected utility, why could it not be rational to adopt EU maximisation as a cognitive procedure? The first argument against EU maximisation as the rational deliberation process relies on some version of the principle 'ought-implies-can' (OIC). It states that one can only be obliged to perform actions that one has the capacity to perform. Even though this principle originates in ethics, it can also be applied to actions that one may be obliged to perform for rational, not moral reasons. Obviously, the relevance of this principle depends on what is considered within one's capacity. The clearest-cut cases involve issues of computability. Clearly, if a task is not computable, then a human agent cannot perform it. However, such cases are rare (see Cherniak on the impossibility to maintain the truth-functional consistency of our beliefs), and present only a small subset of the relevant deliberation processes. To extend the effect of OIC, one could include colloquial notions of 'can'. This includes capacities relative to a given time, effort or context. Uttering 'I can't do this' when faced with the annual tax report, for example, commonly does not mean that one does not have the capacity to do it. Rather, it means one cannot do it because one has other things to do, because one is hungry or tired, or because one needs to obtain further information before the task is settled. Such conditional incapacities, however, cannot be used under OIC, generally. Even those for whom filling in the tax form involves an unreasonable amount of time and energy are not exempt from reporting. Neither does sending in a faulty one get one off the hook. And so it is with many reasoning processes: it may be very difficult, people may make many mistakes, but in principle, if they only put in enough energy, they would be able to perform it correctly. So OIC does not lead to a straightforward rejection of EU maximisation: it may be unreasonably laborious, but impossible to perform it is not.

Instead, the idea that some reasoning processes may be unreasonably laborious leads back to the concept of deliberation costs encountered in section 4. Depending on the agent's cognitive abilities as well as on the circumstances she is in, a specific deliberation procedure will have certain costs. Conditional on these factors, the agent should choose the most *efficient* procedure – the one that gives the highest expected return for the lowest deliberation costs. Simon put it this way: 'The organism may choose rationally within a given set of limits postulated by the model, but it may also undertake to set these limits rationally' ('Behavioral Model' 115).

The claim is that for many environments and for many cognitive capacities, EU maximisation is not the most efficient procedure. Hence some other, simpler procedure, like maximin or satisficing, will be more rational to perform.

This argument, however, runs into a difficult obstacle. Let's start with a simple decision problem *P*. A deliberation procedure *D*(*P*) would return a best option. But it may itself be very costly. So by incorporating the deliberation costs, one can choose that deliberation procedure which is the most efficient one. This requires formulating a meta-decision procedure D', so that $D'(D(P))$ returns the most efficient procedure. But for that procedure D', one also has a choice, and one should find the most efficient one. And so on – so that one quickly ends in an infinite regress. What is the rational deliberation procedure therefore cannot be determined by reference to the procedure's efficiency. Further, EU maximisation cannot be excluded on the ground that it is less efficient than other procedures, as these procedures all need meta-procedures that determine under which conditions it is more efficient.

Members of the ABC group argue for yet a different kind of rationality assessment. Their argument is closely related to evolutionary psychology. Where evolutionary psychologists argue that biological evolution has endowed humans with a multitude of special-purpose psychological modules, members of the ABC group argue that biological evolution has equipped humans with an adaptive toolbox of fast and frugal heuristics. These heuristics make humans *ecologically rational*. Ecological rationality is concerned with a reasonable proportion between efforts and results of a deliberation:

> 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 and Todd 20)

Unlike notions of rationality that make reference to OIC or cost considerations, ecological rationality eschews universal rationality claims. Where the first two make rationality resource-dependent through a general principle, ecological rationality assesses the rationality deliberation procedures only with respect to the specific environment relevant for the deliberating agent. Ecological rationality takes seriously Simon's idea that the structure of the agent's actual environment has a decisive influence on his cognitive architecture. 'Rationality is defined by its fit with reality' (Gigerenzer and Todd 5) – not with any theoretically possible reality, however, but with the actual reality that is relevant for the agent.

The approach assumes that agents have at their disposition an *adaptive toolbox*: '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 740). In the long ancestral history of humanity, the assumption continues, those features of the toolbox were selected that proved to be most useful for survival *in a specific environment.* The question of normative rational assessment is now how well the structure of these heuristics matches with the structure of the relevant environment *today*. To force people into environments irrelevant to them (say, expressing a relation in the world in terms of probability correlations in an experimental set-up) is not pertinent to such assessment. This, however, 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, 'On Narrow Norms and Vague Heuristics'). Against this, ABC insists that deliberation rules must only be tested in relevant environments, and that they work well in environments for which they were adapted. Results from that, members of the ABC group claim, 'supports intuition as basically rational' (Gigerenzer, 'On Cognitive Illusions and Rationality' 242).

This adaptive argument, however, has two problems. First, evolutionary arguments only point to a disposition, not an actuality. Traits selected for fitness tend to be optimal, but there are various lacunae that provide causes for why they don't. 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 are adapted.

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

However, the convincing power of ecological rationality lies less in the theoretical adaptive argument than in the empirical investigations how well individual heuristics fare in present environments. The ABC group has published numerous studies in which they try to show that a specific 'fast and frugal algorithms' may be more efficient for a specified range of tasks than Bayesian computation. A nice intuitive example is the heuristic of ball catching. It's worth quoting Gigerenzer here at length.

> Imagine you want to build a robot that can catch balls – fly balls, as in baseball and cricket. (It's a thought experiment. No such robots exist yet.) For the sake of simplicity, consider situations where a ball is already high up in the air and will land in front of or behind the player. How would you build such a robot?

> One vision is *omniscience*: You aim at giving your robot a complete representation of its environment and the most sophisticated computational machinery. First, you might feed your robot the family of parabolas, because, in theory, balls have parabolic trajectories. In order to select the right parabola, the robot needs to be equipped with instruments that can measure the ball's initial distance, initial velocity, and projection angle. Yet in the real world, balls do not fly in parabolas, due to air resistance, wind, and spin. Thus, the robot would need further instruments that can measure the speed and direction of the wind at each point of the ball's flight, in order to compute the resulting path and the point where the ball will land, and to then run there. All this would have to be completed within a few seconds the time a ball is in the air.
>
> An alternative vision exists, which does not aim at complete representation and information. It poses the question: Is there a smart heuristic that can solve the problem? One way to discover heuristics is to study experienced players. Experimental studies have shown that players actually use several heuristics. One of these is the *gaze heuristic*. When a fly ball approaches, the player fixates the ball and starts running. The heuristic is to adjust the running speed so that the angle of gaze remains within a certain range. The angle of gaze is the angle between the eye and the ball, relative to the ground. In our thought experiment, a robot that uses this heuristic does not need to measure wind, air resistance, spin, or the other causal variables. It can get away with ignoring every piece of causal information. All the relevant information is contained in one variable: the angle of gaze. Note that a player or robot using the gaze heuristic is not able to compute the point at which the ball will land. But the player will be there where the ball lands.
>
> The gaze heuristic is a fast and frugal heuristic. It is fast because it can solve the problem within a few seconds, and it is frugal because it requires little information, just the angle of gaze. ('Fast and Frugal Heuristics' 65)

As a thought experiment, this will only have intuitive appeal. But with technical progress, such a robot can be built and its efficiency tested. In addition, empirical test of heuristics abound. For example, Borges et al. test how the 'recognition heuristic' fares in choosing stock market portfolios, Martignon and Blackmond Laksey test the 'Take The Best' heuristic against three Bayesian models in a variety of situations, and Bishop reviews fascinating evidence for the fact that simple and biasing heuristics for various predictive task outperform complex Bayesian models. These are interesting results for the question of normative assessment. Whether they will convince the majority of researchers in the field is still an open question.

Standard EUT, however, does not make a claim which deliberation procedure is the rational one. All it claims is that *the outcome* of any rational procedure must maximise expected utility. Certainly, the *procedure* of maximising EU is not the only procedure, which has this property – many other procedures may also yield this result, in particular if the range of environments in which those procedures are considered is rather narrowly defined. In fact, there are some important cases where the rule of maximising utility is self-defeating in the sense that its correct application does not lead to outcomes that maximise expected utility. Insomniacs, for example, know that an explicit decision to fall asleep ('the best thing to do given that it's 2 in the morning, I've been in bed for 3 hours and I have to get up at 7 in the morning') is highly counterproductive. Equally counterproductive is the explicit decision to forget an insult (Weirich n11). Some further argue that forming a friendship or falling in love would be impossible if chosen on the basis of an EU maximisation procedure. Hence, the procedure of EU maximisation is clearly distinct from EU maximisation as a procedure's measure of success. No clear arguments have been given that the above three arguments in any way exclude EU maximisation as a measure of success. In fact, OIC is inapplicable, because no prescription is given how EU maximising outcomes are reached, but people are obviously often capable of somehow reaching it. Cost considerations even require some measure of success in order to specify its concept of efficiency. And so does ecological rationality: without some such measure, it cannot detail how well its heuristics fare in specific environments. Until specific arguments are provided, it is not clear why EU maximisation should not be this measure of success.

Of course, a measure of success is different from a prescription or a norm. To prescribe that one ought to maximise the EU of one's actions' outcomes borders closely on the trivial, as it describes that one ought to choose what satisfies one's preferences best. Any normative theory of rationality that intends to capture at least part of the everyday concept of rationality will also have to say which methods ought to be used to reach such a goal. Insofar as standard EUT eschews such a statement, it is less an incorrect than an incomplete theory of rationality.

6. *Conclusion*

This overview distinguished the different usages of the notion of bounded rationality. It is used to denote research both into descriptive and normative theories of human behaviour. In the descriptive arena, bounded rationality sometimes refers to the elicitation of phenomena used to establish the deviation of human behaviour from the standard models. It also refers to a new modelling style in economics, which very selectively incorporates some of these phenomena into models that otherwise are very close to the standard. Most importantly, it refers to the development of new theories of behaviour, some of which try to account merely for behavioural deviations from the standard theory, while others also attempt to give a realistic account of the cognitive procedures underlying human choice.

In the normative arena, bounded rationality refers both to arguments for non-standard norms of reasoning and deliberation, as well as to the investigation whether humans actually satisfy some adequate set of such norms. In this latter discussion, bounded rationality research also gets inevitably involved in the attempt to explicate the pre-theoretical notion of rationality, and to bring the formal theories closer to it.

Beyond the description of its various usages, this overview also discussed some of the central methodological issues of bounded rationality research. It pointed out the problem of external validity of behavioural experiments, and the *ad hoc* criticism against the bounded rationality modelling style. It further addressed the trade-off between parsimony and de-idealisation in descriptive theories of choice, and the role that the rationality assumptions play in it. Last, it contrasted the consequentialist with the deontological assessment of rationality, and discussed the respective problems of applying these two standards in the normative discussion. For each of these topics, important questions remain unsolved, on which the progress of bounded rationality research crucially depends.

Short Biography

Till Grüne-Yanoff's research interests lie at the intersection of philosophy of science, decision theory and political philosophy. He has written papers in these areas for the *Journal of Economic Methodology*, *Analyse&Kritik*, *The Stanford Encyclopaedia of Philosophy*, *The Internet Encyclopaedia of Philosophy*, and two book anthologies published by MIT Press and Routledge, respectively. Current research involves modelling preference change, investigating the role of models in economics, and examining the relevance of behavioural research on public policy. Grüne-Yanoff holds a B.A. in Philosophy and a B.Sc. in Economics from Humboldt Universität, Berlin, and both an M.Sc. in Economics and Philosophy and a Ph.D. in Philosophy from the London School of Economics.

Notes

* Correspondence address: Department of Philosophy and the History of Technology, Royal Institute of Technology, Teknikringen 78B 100 44 Stockholm, Sweden. Email: gryne@infra.kth.se.

[1] For an extensive terminological history, see Klaes and Sent.

[2] Chess also nicely illustrates the difference between costs of computation and the cost for information: information is openly available (it only includes the rules and everything that happens on the board), and hence low in cost. Figuring out the possible positions several moves ahead and making decisions how to react in each case is difficult, hence the computational costs are high.

[3] Of course, there are behavioural theories that bear no or little resemblance to EUT at all, and still use the concept of rationality. In particular, some theories do away with the assumption that the agent has well-defined preferences. Bounded rationality generally does not include such theories, and they will not be discussed further here. For an introduction to such theories of 'expressive rationality' or *Wertrationalität*, see Hargreaves-Heap.

Works Cited

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

Almond, G. A. 'The Political Attitudes of Wealth'. *Journal of Politics* 7.3 (1945): 213–55.

Arkes, H. and K. R. Hammond. *Judgment and Decision Making: An Interdisciplinary Reader.* Cambridge, MA: Cambridge UP, 1986.

Auman, R. 'Rationality and Bounded Rationality'. *Games and Economic Behavior* 21 (1997): 2–14.

Binmore, K. 'Modeling Rational Players: Part I'. *Economics and Philosophy* 3 (1987): 179–214.

Bishop, M. 'In Praise of Epistemic Irresponsibility: How Lazy and Ignorant Can You Be?' *Synthese* 122 (2000): 179–208.

Borges, B., D. G. Goldstein, A. Ortmann, and G. Gigerenzer. 'Can Ignorance Beat the Stockmarket?' *Simple Heuristics that Make Us Smart.* G. Gigerenzer, P. M. Todd, and the ABC Group. Oxford: Oxford UP, 1999. 59–75.

Camerer, C. F. 'Bounded Rationality in Individual Decision Making'. *Experimental Economics* 1 (1998): 163–83.

——. 'Prospect Theory in the Wild: Evidence From the Field'. *Choices, Values, and Frames*. Eds. D. Kahneman and A. Tversky. Cambridge: Cambridge UP, 2001. 288–300.

Cherniak, C. *Minimal Rationality*. Cambridge, MA: MIT Press, 1986.

Cohen, L. J. 'Can Human Irrationality Be. Experimentally Demonstrated'. *Journal of Behavioral and Brain Sciences* 4 (1981): 317–29.

Conlisk, J. 'Why Bounded Rationality?' *Journal of Economic Literature* 34 (1996): 669–700.

Cozic, M. 'Methodological Foundations of Bounded Rationality'. *mimeo*, I-ENS Ulm, Paris, 2006.

Diecidue, E. and P. P. Wakker. 'On the Intuition of Rank-Dependent Utility'. *Journal of Risk and Uncertainty* 23.3 (2001): 281–98.

Edgeworth, F. Y. 'The Element of Chance in Competitive Examinations'. *Journal of the Royal Statistical Society* 53.3 (1890): 460–75.

Friedman, M. 'The Methodology of Positive Economics'. *Essays in Positive Economics*. Chicago, IL: U of Chicago P, 1953. 3–43.

Gigerenzer, G. 'Fast and Frugal Heuristics: The Tools of Bounded Rationality'. *Handbook of Judgment and Decision Making*. Eds. D. Koehler and N. Harvey. Oxford: Blackwell, 2004. 62–88.

——. 'The Adaptive Toolbox'. *Bounded Rationality. The Adaptive Toolbox.* Eds G. Gigerenzer and R. Selten. Cambridge, MA: MIT Press.

——. 'On Cognitive Illusions and Rationality'. *Poznan Studies in the Philosophy of the Sciences and the Humanities* 21 (1991): 225–49.

——. 'On Narrow Norms and Vague Heuristics: A Reply to Kahneman and Tversky'. *Psychological Review* 103.3 (1996): 592–6.

——, P. M. Todd, and the ABC Group. *Simple Heuristics that Make Us Smart.* Oxford UP, 1999.

Grether, David M. and Charles R. Plott. 'Economic Theory of Choice and the Preference Reversal Phenomenon'. *American Economic Review* 69 (1979): 623–38.

Gruene-Yanoff, T. 'Intentional Action Explanations are Not Inherently Normative'. *mimeo* KTH, Stockholm, 2006.

Guala, F. and L. Mittone. 'Experiments in Economics: External Validity and the Robustness of Phenomena'. *Journal of Economic Methodology* 12 (2005): 495–515.

Haerle, W. and Alan P. Kirman. 'Nonclassical Demand: A Model-Free Examination of Price-Quantity Relations in the Marseille Fish Market'. *Journal of Econometrics* 67.1 (1994): 227–57.

Hargreaves-Heap, Shaun P. *Rationality in Economics*. Oxford: Basil Blackwell, 1989.

Harless, David W. and Colin F. Camerer. 'The Predictive Utility of Generalized Expected Utility Theories'. *Econometrica* 62 (1994): 1251–89.

Hausman, D. *The Inexact and Separate Science of Economics*. Cambridge: Cambridge UP, 1992.

Hogarth, R. M. *Judgement and Choice: The Psychology of Decision.* New York, NY: John Wiley & Sons, 1980.

Kahneman, D. 'Maps of Bounded Rationality: Psychology for Behavioral Economics'. *American Economic Review* 93.5 (2003): 1449–75.

Kahneman, D., P. Slovic, and A. Tversky, eds. *Judgment under Uncertainty: Heuristics and Biases*. New York, NY: Cambridge UP, 1982.

Kahneman, D. and A. Tversky. 'On the Psychology of Prediction'. *Psychological Review* 80 (1973): 237–57.

——. 'Prospect Theory: An Analysis of Decision under Risk'. *Econometrica* 47 (1979): 263–91.

Kalai, E. and E. Lehrer. 'Rational Learning Leads to Nash Equilibrium'. *Econometrica* 61.5 (1993): 1019–45.

Klaes, Matthias and Esther-Mirjam Sent. 'A Conceptual History of the Emergence of Bounded Rationality'. *History of Political Economy* 37.1 (2005): 27–59.

Koehler, J. J. 'The Base Rate Fallacy Reconsidered: Descriptive, Normative, and Methodological Challenges'. *Behavioural and Brain Sciences* 19 (1996): 1–53.

Loewenstein, G. and J. Lerner. 'The Role of Affect in Decision Making'. *The Handbook of Affective Science*. Eds. Richard Davidson et al. Oxford: Oxford UP, 2001. 619–42.

Loomes, G. and R. Sugden. 'Regret Theory: An Alternative Theory of Rational Choice under Uncertainty'. *Economic Journal* 92 (1982): 805–24.

Martignon, L. and K. Blackmond Laskey. 'Bayesinan Benchmarks for Fast and Frugal Heuristics'. *Simple Heuristics that Make Us Smart*. G. Gigerenzer, P. M. Todd, and the ABC Group. Oxford: Oxford UP, 1999. 169–89.

von Neumann, J. and O. Morgenstern. *The Theory of Games and Economic Behavior*, Princeton, NJ: Princeton UP, 1947.

Newell, A. and H. A. Simon. *Human Problem Solving*. Englewood Cliffs, NJ: Prentice-Hall, 1972.

Nisbett, R. E. and L. D. Ross. *Human Inference: Strategies and Shortcomings of Social Judgment*. Englewood Cliffs, NJ: Prentice-Hall, 1980.

Odean, T. 'Are Investors Reluctant to Realise their Losses?' *Journal of Finance* 53.3 (1998): 1775–98.

Quiggin. 'A Theory of Anticipated Utility'. *Journal of Economic Behaviour and Organization* 3(4) (1982): 323–43.

Payne, J. W., J. R. Bettman, and E. J. Johnson. *The Adaptive Decision Maker*. Cambridge: Cambridge UP, 1993.

Roth, A. E. 'Introduction to Experimental Economics'. *The Handbook of Experimental Economics*. Eds. John H. Kagel and Alvin E. Roth. Princeton, NJ: Princeton UP, 1995. 1–98.

Rubinstein, Ariel. *Modeling Bounded Rationality*. Cambridge, MA: MIT Press, 1998.

Samuels, R., S. Stich, and M. Bishop. 'Ending the Rationality Wars: How To Make Disputes About Human Rationality Disappear'. *Common Sense, Reasoning and Rationality*. Ed. R. Elio. New York, NY: Oxford UP, 2002. 236–68.

Samuels. R., S. Stich, and L. Faucher. 'Reason and Rationality'. *Handbook of Epistemology*. Eds. I. Niiniluoto, M. Sintonen, and J. Wolenski. Dordrecht: Kluwer, 2004. 131–79.

Sargent, T. S. *Bounded Rationality in Macroeconomics*. Oxford: Oxford UP, 1993.

Schoemaker, P. 'The Expected Utility Model: Its Variants, Purposes, Evidence and Limitations'. *Journal of Economic Literature* 20 (1982): 529–63.

Schwarz, N. F. Strack, D. Hilton, and G. Naderer. 'Base Rates, Representativeness, and the Logic of Conversation: The Contextual Relevance of "Irrelevant" Information'. *Social Cognition* 9 (1991): 67–84.

Selten, R. 'The Chain Store Paradox'. *Theory and Decision* 9.2 (1978): 127–59.

Shleifer, A. *Inefficient Markets: An Introduction to Behavioural Finance*. Clarendon Lectures. Oxford: Oxford UP, 2000.

Simon, Herbert A. 'A Behavioral Model of Rational Choice' *Quarterly Journal of Economics* 69 (1955): 99–118. Reprinted in H. A. Simon *Models of Bounded Rationality*. Cambridge, MA: MIT Press, 1982. 239–58.

——. 'Bounded Rationality in Social Science: Today and Tomorrow'. *Mind & Society* 1.1 (2000): 25–39.

——. 'From Substantive to Procedural Rationality'. *Method and Appraisal in Economics*. Ed. S. J. Latsis. Cambridge: Cambridge UP, 1976. 129–48.

——. 'Invariants of Human Behavior'. *Annual Review of Psychology* 41 (1990): 1–19.

——. *Models of Man*. New York, NY: John Wiley & Sons, 1957.

——. 'Rational Choice and the Structure of the Environment'. *Psychological Review* 63 (1956): 129–38.

Starmer, C. 'Developments in Non-Expected Utility Theory: the Hunt for a Descriptive Theory of Choice under Risk'. *Journal of Economic Literature* 38 (2000): 332–82.

Stein, E. *Without Good Reason*. Oxford: Clarendon Press, 1996.

Sugden, R. 'Experiments as Exhibits and Experiments as Tests'. *Journal of Economic Methodology* 12 (2005): 291–302.

Todd, P. M. and G. Gigerenzer. 'Précis of Simple Heuristics That Make Us Smart'. *Behavioral and Brain Sciences* 23 (2000): 727–41.

Thaler, R. H. *The Winner's Curse: Paradoxes and Anomalies of Economic Life*. Princeton, NJ: Princeton UP, 1992.

Tversky, A. and D. Kahneman. 'Advances in Prospect Theory: Cumulative Representation of Uncertainty'. *Journal of Risk and Uncertainty* 5 (1992): 297–323.

Weirich, P. *Realistic Decision Theory*. New York, NY: Oxford UP, 2004.

Williamson, O. *Economic Organisation: Firms Markets and Policy Control*. Brighton: Wheatsheaf, 1986.

Zermelo, Ernst. 'Über eine Anwendung der Mengenlehre auf die Theorie des Schachspiels'. *Proceedings of the Fifth International Congress on Mathematics*. Cambridge Vol 2 1913. 501.