PHILIPPE MONGIN

DOES OPTIMIZATION IMPLY RATIONALITY?

ABSTRACT. The relations between rationality and optimization have been widely discussed in the wake of Herbert Simon's work, with the common conclusion that the rationality concept does not imply the optimization principle. The paper is partly concerned with adding evidence for this view, but its main, more challenging objective is to question the converse implication from optimization to rationality, which is accepted even by bounded rationality theorists. We discuss three topics in succession: (1) rationally defensible cyclical choices, (2) the revealed preference theory of optimization, and (3) the infinite regress of optimization. We conclude that (1) and (2) provide evidence only for the weak thesis that rationality does not imply optimization. But (3) is seen to deliver a significant argument for the strong thesis that optimization does not imply rationality.

1. INTRODUCTION

1.1. *Aims and Strategy of this Paper*

The following model of rationality is pervasive in economics and widespread elsewhere in the social sciences. A rational person's preferences are represented by a real-valued objective function, and his choices correspond to the values of the ("instrument") variable that maximize this function over the set of available options. This model is justified, at least implicitly, by the normative claim that rational individuals always satisfy their preferences within their feasibility constraints, or briefly put, that rationality implies optimization. The claim does not relate at all to the content or value of the person's preferences. It thus takes for granted what is classically called the instrumental sense of rationality. Familiar though it sounds, this claim conflicts with the suggestions of ordinary language. Even restricting attention to the instrumental sense, rationality appears to be a broad and ill-defined concept. A rational individual, it is sometimes said, is one who chooses in a way that is "appropriate" to the conditions of his choice; or, following another suggestion, one who acts "on good reasons". These rather vague definitions suggest that the technical notion of optimization cannot simply follow from the ordinary concept of instrumental rationality.[1]

Synthese **124:** 73–111, 2000.

But the ordinary concept does appear to be at least compatible with a maximizing interpretation. Making the best choice is one way of choosing "appropriately"; the optimality property of a solution is a "good reason" for implementing it. Accordingly, some writers have claimed that optimization provides a modelling of rationality, albeit not the only one. (From now one I will dispense with the adjective "instrumental" before either "rationality" or "rational", but it should be clear that I am concerned all along with instrumental rationality, not with other forms.) In his collection *Models of Bounded Rationality* (1983) Herbert Simon broadly subscribes to this position, which I shall refer to as the *conciliatory account*. Simon documents mechanisms or methods of 'bounded rationality' which, as he understands them, are particular cases of rationality, while the mechanisms or methods of optimization (relabeled by him 'absolute rationality') are another particular case. Although he has occasionally been misunderstood on this, Simon does not a priori discard the optimizing model of choice. In an earlier paper (Mongin 1984) I follow the conciliatory account and emphasize that there are two levels of analysis of the rationality concept: on the one hand, the generic level at which rationality is loosely defined in terms of appropriate choice and, on the other, the specific level at which models (such as those of "bounded" and "absolute" rationality) are introduced and can be compared with each other. Each model aims at expressing the generic notion of rationality better than its rivals, and it is the social scientist's task to assess these conflicting claims by paying attention to the circumstances of choice. Simon's famous contention that an absolute rationality strategy might turn out to be less rational than a 'satisficing' or bounded rationality strategy goes hand in hand with the following methodological stance: which model is the more relevant of the two entirely depends on the cognitive circumstances. Simon's position, and the conciliatory account more generally, are thus highly flexible. This account does have a polemical import nonetheless: it denies the received view among economists that rationality implies optimization. However, it does not deny the significance of optimization as a possible rendering of rationality.

This paper will add further evidence against the economists' received view (in Sections 2.5 and 3 below) but the gist of the argument is to question the conciliatory account itself. I want to take issue with the seemingly unproblematic assertion that if rationality does not imply the optimization condition, at least the converse holds. Accordingly, I have collected here various arguments or constructions of rational choice theory which seem to suggest that under some relevant choice circumstances, to optimize is

not rational. At least, this is the initial suggestion. Whether the analysis will confirm it is very much the object of this paper.

The first topic – cyclical choices – is mostly borrowed from the psychology of decision. After several writers in the field, I argue that cyclical choices are normatively acceptable in certain choice circumstances. But I will add that this argument does *not* have the effect of excluding optimization from the area of rational choice, even for the range of circumstances it applies to (essentially when the agent chooses between multidimensional objects). Hence, the first group of objections remains inconclusive as far as the main project of this paper is concerned. At least, it leads to a conclusion of some sort. After briefly reviewing – and dismissing – the famous *money pump* argument, I claim that those who believe that optimization implies rationality cannot draw much comfort either from the discussion of cyclical choices.

The second topic is borrowed from pure preference theory, where optimization has been redefined in terms of properties of the agent's choices. Revealed preference theory, as it is has been called, is a methodologically contentious part of economics, but I do not aim at reviving the classic objections it has raised. Rather, taking its axioms at their face value, I classify them according to their normative strength and emphasize the cognitive difficulties of the choice operations with which optimization is formally equated. The discussion here turns out to be mostly relevant to the claim that rationality implies optimization. Indeed, the suggestion was strongly made in the revealed preference literature, that it provides a *justification* to the optimization principle. I dismiss this claim. The discussion here does not yet deliver an argument against the rationality of optimization. However, it already points towards its central weakness – i.e., the imbalance between the internal costs of optimization and what it achieves in terms of the initial objective.

How to analyze this imbalance is the third and last topic of the paper. I mention the classic point in practical philosophy, which was revived by Ryle (1949), that any rational decision criterion implies an infinite regression of decisions. I will suggest that this difficulty becomes most acute whenever rational choices are construed as optimizing ones. I have based my discussion on a very simple model of a firm's decision which is adapted from the more abstract framework in Mongin and Walliser (1988). Admittedly, the conclusions are sensitive to the underlying assumptions, but it does seem that at long last, the infinite regression argument substantiates the claim that under certain choice circumstances, it is *positively irrational* to optimize. Remarkably, this critical point follows from pursuing Herbert Simon's initial objection against the absolute rationality model

that the latter does not take into account the internal costs of the decision-maker's optimization. Simon stopped at the conciliatory account, but his pathbreaking analysis had a much stronger potential.[2]

1.2. *Some Background Distinctions*

There are at least two ways in which optimization can be discussed normatively, i.e., as a principle to be applied by the agent and as a modelling principle to be employed by the theorist. Contrary to the more elegant phrasing, "Is it rational to optimize?", the title of this article equivocates between these two meanings. The ambiguity is part of its subject matter, as I hope the comparison between the three topics will make clear. As long as one highlights situations pathological for optimization - such as cyclical choices – or even when one stresses the heavy cognitive requirements it imposes on him – as in the reinterpretation below of revealed preference theory – one lays oneself open to the following defensive move. (It is implicit in much of theoretical economics.) It consists in replacing the *agent-relative* normative interpretation by a *theorist-relative* modelling principle such as: 'to each act that seems intuitively rational, it is possible and desirable to attach an optimizing description'. The strength of the infinite regression objection is that, if properly formulated, it also hits this seemingly cautious reformulation of optimization. More generally, I do not think that a critique of optimization can be successful if it just addresses the normative issues involved in the agent-relative version, and none of the methodological issues involved in the theorist-relative one.

The most basic notion of optimization is agent-relative, and perhaps best stated in terms of the following microeconomics theorem. If an agent's preferences can be represented as a weak ordering, that is, a reflexive, transitive and complete binary relation, and this ordering is continuous in a technically suitable sense, then his preferences can also be represented by a continuous utility function; and conversely (Debreu 1959; for a good treatment see also Malinvaud 1971, 18–20). The theorem leads to the conclusion that on compact sets of alternatives it is possible to maximize the numerical representation and, hence, the agent's preferences.[3]

The result does not indicate that the agent is *effectively* a maximizer. The supplementary statement that he is – a statement which is more informal than the preceding one – is generally taken for granted among economists. When they write that an individual is 'endowed' with a utility function that can be maximized on the domain of choice, they typically also imply that, at the moment of choice, this individual will select one of the maximizing values of the 'instrument' variable. (Economists are often worried about which optimal solution is selected if there are more than

one, but this is again a different issue.) By and large, the only problem for optimization that is recognized by economists concerns the existence of a numerical representation, not its use by the agent. This is unduly restrictive. The statement that the agent is effectively a maximizer when maximization is possible is not an analytical statement. Philosophers of social sciences have struggled with an analogous problem in connection with what they call the rationality principle.[4] Symmetrically, it would be unduly restrictive to assess optimization only in terms of effective maximization, while taking for granted the conditions stated in the existence theorem. There are two sides to the optimization coin. One group of arguments in the paper is clearly on one side of the distinction between the conditions of maximization and effective maximization (since cyclical choices call into question the crucial transitivity condition). But, as will be seen, the other arguments relate to the two sides of the coin at the same time. I could hardly state them properly if I were to specialize the paper one way or the other.

Among the conditions stated in the microeconomics theorem that guarantees the existence of a representation, I will immediately dispose of continuity. At least in the theory of choice under certainty, to which I limit myself here, continuity is usually seen as a mere technical requirement.[5] Evidence for this is provided by the textbook discussions of the lexicographic model of choice, whereby the agent maximizes a discontinuous index. Standard texts in microeconomics (e.g., Malinvaud l971, 18–20) do not claim that it is irrational to maximize a lexicographic ordering instead of a scalar function. The implicit understanding is that consumer theory uses the latter kind of representation for convenience, not for substantial reasons. There are occasional dissenters, however. Harsanyi (1976, 93) – though not in so many words – suggests that ordering partial criteria hierarchically is only one stage towards defining the optimizing rationality model. In his opinion as I reconstruct it, this model should not count as fully formed until it allows for smooth trade-offs between the different dimensions of utility – that is to say, until the agent is endowed with a continuous utility function. This view is intriguing but to discuss it any further would take us away from the main point. So I will endorse the position that continuity is part of the *microeconomic* modelling, but not of the general notion of optimization,[6] and henceforth deal nearly exclusively with the remaining two conditions on the preference relation.

1.3. *A Warning on Method*

The intuitive notions of individual rationality which are to be confronted with more technical conditions are not susceptible to preliminary definition

except for the highly general statements made in Section 1.1. The difficulty here is of a familiar kind, and like others in rational choice theory, I will turn it around without disposing of it entirely. The point is to reach a reflective equilibrium, i.e., to progressively adjust abstract conceptions and examples to each other. This method works in a relatively simple way on the examples of intransitive choices. Here, it is just a question of confronting the, as yet insufficiently determined, notion of rationality with particular circumstances in which the meaning of this word can be more precisely appreciated. When one compares optimization with other formal conditions as in the second and third topics, the reflective equilibrium method interacts with the no less classical strategy of analyzing concepts in terms of necessary and sufficient conditions. When this stage is reached, it is not only a question of balancing partial examples against a vague theory, but also of evaluating the normative plausibility of a choice method in terms of equivalent (or at least necessary) conditions that are somehow more interpretable.

2. CYCLICAL CHOICES AND RATIONALLY DEFENSIBLE INTRANSITIVITIES

The psychology of decision and decision theory literatures abound with examples of cyclical choices. Some of them appear to be rationally defensible, a point well emphasized for instance in Anand's (1987) and Bar-Hillel and Margalit's (1988) surveys. If one accepts the revealed preference view that choices unproblematically reflect preferences, a view which this paper does not purport to challenge, it should follow that in the circumstances spelled out by these examples, the agent's preferences are both rational and incompatible with optimization, since they violate transitivity. From a potentially large collection, I have selected three cases each of which illustrates one facet of the normative assessment of intransivities.

2.1. *The Choice of Spouse Experiment*

I start with the classic study by May (1954) on choosing a spouse. The experimenter classified the objects of choice along three dimensions – intelligence, beauty, fortune – each measured on a distinct qualitative scale. The subjects were faced with successive binary choices. May obtained 62 replies, of which 17 demonstrated a cycle and 21 amounted to applying a lexicographic rule. The other 24 were based on coherent trade-offs between the three dimensions – they were thus the only ones to conform to the microeconomic theory of optimization. One of the first of its kind,

the experiment appears crude by contemporary standards but remains instructive nevertheless. May provided the following plausible interpretation for the 17 cyclical subjects: they had consistently applied the majority rule to the three dimensions and inadvertently stumbled on the Condorcet paradox. May's interpretation suggests justifying cyclical choices by analogy with what can be said for deliberative assemblies. The cycles are arguably the undesirable consequence (inevitable and possibly well understood by the agent himself) of adopting a decision rule which conforms to otherwise impeccable properties such as anonymity, neutrality, etc. The analogy is quite simple to defend in the context of the particular study, because May had imposed from the start the splitting up of the 'spouse' object of choice into intelligence, beauty, and fortune, so that the subjects had to relate to the dimensions in making their choices. However, these somehow predetermined choices may just be an artifact of his questionnaire. There is no evidence in the study that choices made in more natural situations would deliver the same high proportion of cyclical answers.

There is a related problem with many other discussions of cyclical choices in the literature. They involve objects of choice that are redescribed by the observer in terms of vectors of 'characteristics'. Individual choices are then often rationalized as in May, i.e., by analogy with social choice, while characteristics play the role of voters. Their aggregation would have to obey the alledgedly compelling, but logically overdetermined constraints leading to Arrow's impossibility theorem or some related impossibility result. However, the obvious difference with social choice is that voters are observable and relevant in a way characteristics are not. To assert the role of characteristics in explaining choices requires a significant empirical hypothesis, and at the *normative* level too, they raise a problem. If the splitting up of the object of choice, as envisaged by the observer, does not agree with the agent's own understanding of this object, it is unclear why the characteristics must be relevant to the rationality or otherwise of his choices. This important objection will have to be addressed also vis-à-vis the next examples.

2.2. *Arguments About Intransitive Indifference*

Contemporary decision theorists have widely come to recognize that cyclical choices might result from perceptual effects (e.g., Fishburn 1970a). Suppose that the agent only notices temperature differences of 3 degrees; he will then identify 17 °C with 19 °C, 19 °C with 21 °C, but not 17 °C and 21 °C. If he is asked to choose between three baths having temperature 17 °C, 19 °C, and 21 °C, he will express indifference between successive baths, and strict preference among the extreme ones, so that, clearly,

his choices cannot reflect a transitive preference. On the other hand, they cannot be deemed irrational.

Impressed by this simple fact, a number of today's decision theorists regard as being normatively compelling the transitivity of only the *strict* preference relation. A weak preference relation that satisfies the latter property, but not necessarily full transitivity, is called quasi-transitive. This weakening is no doubt a significant departure from the ordinary conception, but it suffices to preserve a notion of optimality; for a 'maximal element' to exist on a finite choice set only the acyclicity of strict preference is needed, and this is an even weaker property than quasi-transitivity.[7]

Luce's semi-ordering (1956), which I do not redefine here, constitutes a less radical departure from orthodoxy than quasi-transitivity. While remaining compatible with intransitive indifference, a semi-ordering has implications that a quasi-transitive ordering lacks. For example, if an individual is indifferent between a and b and strictly prefers b to c, he does not strictly prefer c to a. Starting with Luce's semi-ordering, Tversky (1969) has devised the *lexicographic semi-ordering*, which contrary to the latter, does not even satisfy acyclicity. Essentially, it is a lexicographic preference relation such that the first dimension obeys the semi-ordering axioms, and can thus possibly have an intransitive indifference; the dimensions after the first might conform to the ordering axioms. The following example shows a lexicographic semi-ordering at work. An employer favours intelligence over years of experience, but to measure this intangible quality he has to make do with a scale that permits 3 unit errors. So, the strict preference of the employer will produce cycles like:

$$(115, 7) > (117, 3) > (120, 0) > (115, 7).$$

Tversky's concept is attractive because it leads to cyclical choices by combining two ingredients neither of which, when taken in isolation, would conflict with acyclicity. Since intransitive indifference cannot be said to be irrational, the conclusion that the employer's choices are not irrational follows from the point that lexicographic preferences do not involve any irrationality *per se* – a widely accepted point as I mentioned earlier in discussing continuity.

Like May's, this example is open to the charge that the employer's preferences are assumed to be already structured in terms of the given dimensions. Notice, however, the following difference. May started from empirically observed choices and offered to rationalize them in terms of a normatively defensible non-transitive preference. Tversky started from a normatively defensible non-transitive preference relation, and showed that

it sometimes entailed cyclical choices. Since I am not planning to question revealed preference theory in the paper, I will not emphasize this difference. However, by and large, Tversky's *a priori* reasoning is more clearly relevant than May's possibly disputable *a posteriori* reconstruction. The next example to come is also of the *a priori* type.

2.3. *The Horserace 'Paradox'*

In Blyth's (1972) horserace example, the subject is asked to choose among successive pairs of bets on horses A, B and C, respectively. The finishing position of each horse depends on whether the going is hard, soft or heavy, and this in turn depends on an unknown state of nature. If it rains (= S_1), horse A beats B and C, and B beats C; if it does not rain, but the weather is wet (= S_2), horse B beats C and A, and C beats A; if the weather is dry (= S_3), horse C beats A and B, and A beats B. Accordingly, when S_1, it is better to bet on A than B, and on B than C; when S_2, it is better to bet on B than C, and on C than A; when S_3, it is better to bet on C than A, and on A than B. Blyth claims that the *ex ante* preferences resulting from these data should conform to the rule that:

> the agent prefers to bet on X than Y iff the probability that X beats Y is greater than $\frac{1}{2}$.

But this apparently plausible rule leads to cyclical choices, as is easily verified in the particular case where $p(S_1) = p(S_2) = p(S_3) = 1/3$. The agent then bets on B against C, on C against A, and on A against B.

When probabilities are equal, Blyth's better in effect applies the simple majority rule three times, which takes us back to the choice of spouse experiment. While it shares with May's this curious analogy, Blyth's example is more interesting not only because, like Tversky's, it results from an *a priori* reasoning, but also in the following respect that was lacking in Tversky: the multidimensional structuring of the object is now unproblematic. The distinction between states of nature is 'objective' in a sense that the distinction between psychological dimensions was not. Commenting on Blyth, Bar-Hillel and Margalit (1988, 132) go so far as to write that in this case, "the cyclicity of choice is in the external world". I agree with their essential point although I do object to their particular formulation. (That is, the observer cannot dispense with the psychological assumption that the agent bases his choices on considering the objective states; this assumption is easy enough to accept, but it is not itself grounded in the "external world".)

Are we then at last presented with a non-artificial paradox of optimization?[8] This is what Blyth would like his reader to conclude. In order to assess the paradox, it is crucial to distinguish between two dif-

ferent cases, i.e., that of several two-horse contests on the one hand, and that of a single three-horse contest on the other. When faced with a race between two horses, Blyth's rule of decision says, very plausibly, that the individual should bet on the horse with the higher chance of finishing the first. Each of the three two-horse contests will elicit a different answer from the individual, hence a cycle of bets, but no paradox at all. There is nothing strange in the fact that non-identical choice situations lead to non-identical choices! Now, one should resist the temptation of dealing with choices in successive two-horse races *as if they were pairwise choices in a single three-horse contest*. When the three horses are running together, to bet on *A* rather than *B* means something different than in the other class of situations – to wit, it means, to bet that *A*, rather than *B*, will finish first of *A, B and C*. Although the type of contest is not the same, Blyth's rule recommends again that one should bet on *X* rather than *Y* if *X* has the greater probability of finishing the first of the two. This is absurd. One should now, of course, bet on *X* rather than *Y* if *X* has a greater probability than *Y* of finishing the first of the three. When the three probabilities are equal, this leads to being indifferent between any two bets one is faced with.

A rational individual always applies the same rule, which is to select the horse with the highest probability of winning, where the meaning of 'winning' is fixed by the nature of contest. I suspect that Blyth has insisted on his curious rule because he has overlooked a trivial difference. The seemingly striking horserace counterexample has not stood up to scrutiny.[9]

2.4. *General Assessment of Cyclical Choice*

The failure of Blyth's paradox sheds some light on May's example. The latter reveals rationally defensible cycles in precisely those circumstances – pairwise choices among at least three objects given at a time – in which I just argued that a rational horse-race better cannot exhibit a cycle of strict preferences. What is then the conceptual point underlying this discrepancy? Unlike May, Blyth assumed that there were given numerical exchange rates (probabilities) assigned to the dimensions. So the difference between the two boils down to this: to follow Blyth's rule would be to discard *existing*numerical information on the characteristics, while to choose spouses cyclically reflects the agent's failure to *generate* this information – a failure which may be understandable given the choice circumstances. If I actually believe that beauty, fortune and intelligence approximate the independent constituents of an ideal spouse and are truly incommensurable qualities, what would be irrational for me, after all, would be to renege on my choices just to restore cyclicity. Like a member of a political assembly,

I might well foresee that choices made in a sequence will become cyclical, but there is nothing I can do about it.

The two substantial examples of this section, May's and Tversky's share a common pattern. In either case, the agent is faced with two vectors of characteristics $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$, and he evaluates the amounts x_i, y_i, of each characteristic i in terms of a given numerical scale u_i. Under these assumptions, a decision rule amounts to a particular way of exploiting the information given by all the $u_i(x_i)$, $u_i(y_i)$. Both May and Tversky assume that comparisons are *made first within characteristics and across objects*. That is, the decision rule first computes the differences $u_i(x_i) - u_i(y_i)$ for each i, and then reaches a conclusion by aggregating these differences across the i. Classical decision theory (of which expected utility theory is a particular case) requires that comparisons be *made first within objects and across characteristics*, i.e., it begins by computing $U(u_1(x_1), \ldots, u_n(x_n))$ and $U(u_1(y_1), \ldots, u_n(y_n))$ according to some aggregation function U, and only at the end is a difference taken to decide which of x and y is selected. By construction, the classical decision procedure cannot lead to intransitivities, while it is easy to guess – and May's and Tversky's examples confirm - that the other does.[10]

There is a case to be made in favour of the non-classical method. Firstly, in the 'Paretian' case when one of the two alternatives dominates the other with respect to all characteristics, the method leads to an answer simply by inspecting the $u_i(x_i)$, $u_i(y_i)$ values. Secondly, and relatedly, it starts by comparing quantities that are already commensurable, thus postponing the more problematic step of aggregating those which are initially incommensurable. Intuitively, this is a wise procedure to follow because at least in the 'Paretian' case, the last step becomes dispensable, while in the general case, the way of performing it can be adjusted to the context. Sometimes, it will be possible and not too costly to define a complete system of exchange rates between dimensions. Sometimes however, this is impossible (because dimensions appear to be truly incommensurable), or impractical (typically, when there are too many dimensions). Then, various second-best rules – such as the majority rule or comparisons along only a few prioritized dimensions[11] – come into play. The non-classical procedure is compatible with any of these resolutions, and thus enjoys a flexibility that classical decision theory lacks. Thirdly, and more tentatively, it seems to be the closer of the two to the deliberative mode of decision-making. To deliberate about a course of action is to examine arguments for and against it in turn, and suspend judgment until sufficiently many significant arguments have been considered. Here, the course of action is to choose X rather than Y, and the relevant arguments are the differences in each scale

$u_i(.)$. Judgment is passed after sufficiently many of the differences have been calculated.[12]

The following distinction will perhaps help one to assess the normative strength of intransitive choices. Generally in this paper, I will say that in a given situation, a choice (in particular, a choice involving a failure of optimization) is *strongly* rational if it appears to be entailed by all pretheoretic notions of rationality that come to mind in this situation, while a *weakly* rational one is entailed by at least one, but perhaps no more than one, of those pretheoretic notions. Are the cyclical choices analyzed in this section weakly or strongly rational? The discussion of the last paragraph – about the advantages of the non-classical method of aggregation – suggests that intransitive choice behaviour is at least weakly rational in a wide range of situations. Can one go beyond this conclusion? Certainly not on the basis of the arguments made thus far. To show that certain intransitive behaviour is strongly rational amounts to showing that transitive behaviour would not even be weakly rational under the same circumstances, and that is more than one can hope to establish here.

2.5. *On the 'Money Pump' Argument*

For the sake of comparison, a word might be said about the opposite, more popular strategy of arguing that intransitivities are not even weakly rational in certain relevant circumstances. Best known in this area is the 'money pump' argument, as initiated by Davidson, McKinsey and Suppes (1955). It aims at showing that an agent entertaining cyclical strict preferences incurs the risk of being ruined. Here is the authors' amusing example (notice that it involves a multidimensional object of choice once again). The head of an American university department offers a prospective employee three options: a = a top-level post at \$50,000 a year; b = an intermediate-level post at \$55,000 a year; c = a low-level post at \$60,000 a year. The candidate prefers a to b because more prestige compensates for less money, and b to c for the same reason; but he also prefers c to a because the financial gain compensates for the loss of prestige. In each case the preferences are strict. The head of department, who is not overly scrupulous, propositions the candidate as follows in an attempt to bankrupt him: 'I see you prefer b to c, so I'll let you have the choice between b and c in exchange for \$25'. The candidate quickly hands over the cash. 'I believe I am right in thinking that you would prefer a to b. Never mind, I'll let you have the choice between a and b if you just give me \$25'. No sooner said than done. 'It would appear that you prefer c to a ...' and so on. Hence the term 'money pump'. The scenario is meant to illustrate that an individual having cyclical preferences makes himself a prey to exploitation by oth-

ers. Alledgedly, this conclusion establishes that cyclical preferences are irrational (not even weakly rational, in the above terminology). Notice the *a priori* structure of the argument. Like Tversky's and Blyth's, it goes from assumed preferences to entailed cyclical choices, although, this time, with a view of disqualifying the assumed preferences.

Davidson, McKinsey and Suppes's scenario, when analyzed, is seen to rest on a number of tacit suppositions. The most obvious one is that the candidate places a monetary value on each of the successive choices. This assumption is in the spirit of standard microeconomics, and can be formalized by stating that the agent has a complete and continuous preference relation over a set of alternatives made up of various combinations of professional positions and money amounts. Such a formalization points towards a weakness in the argument for acyclicity – it does depend on accepting further decision-theoretic axioms in the first place. Even a defender of the argument has to concede that it does not apply universally; it applies only to those circumstances in which a price for swapping choice sets can plausibly be assumed.

There are other, more subtle assumptions on the choice process at each stage, and this leads to identifying a further weakness in the argument. Note first that at each stage, the choice set must be binary. If, at some point, the choice were between a, b and c, there would be a strong reason for the candidate to hand over nothing at all - since in this case each option he can choose is ranked below another one which is simultaneously available, and this is a strong reason for not choosing anything from the set.[13] Now, even granting a sequence of binary choice sets $\{b, c\}$, $\{a, b\}$, $\{c, a\}$, it is not clear why the candidate would want to pay anything in order to be given the choice between a and b. Specifically, suppose that the candidate anticipates one step (and only one step) ahead of his current position. Following one line of argument, he would foresee that he would next drop b to get a which he actually prefers to b, so he should plan to wait and make only one payment, namely the second. But through the same anticipatory reasoning, when the candidate is actually faced with $\{a, b\}$, the second payment becomes worthless to him. Similarly with the third payment when he is actually confronted with $\{c, a\}$. He ends up paying out nothing at all. This is not a waterproof line of reasoning, however. Assuming that the candidate's initial knowledge includes all of the three preference statements, his decision to wait until the second stage in order to pay 25$ and get option a is open to the following objection. The candidate is planning to choose a, which he knows is ranked below c, although he knows c can be made available now for the very same 25$! Underlying this objection is the general principle that one should not pay anything for

an option that one knows to be worse than another that one knows can be got for the same price – a dominance principle, really. An even simpler application of this principle will block the candidate's tentative choice of c. So the conclusion is the same as before, though it is reached by a more roundabout argument. *The candidate will hand over nothing at all*. To sum up a sequence of binary choice sets with anticipation one step ahead brings about the same state of affairs as does the single choice set $\{a, b, c\}$, i.e., abstention.

The foregoing counterargument is based on a perhaps little plausible hypothesis – i.e., one-stage myopic anticipation – but nothing in the initial wording of the money pump argument excludes this possibility. Commentators – notably Schwartz (1986), who was among the first to write in detail about this issue – have usually claimed that the argument does not hold unless one assumes that the agent has no anticipation whatsoever about the future. This assumption is scarcely plausible to begin with, and, as cycles are reproduced, it becomes even less so. One might perhaps concede that the candidate would lose a few $25 bills but not that he would bankrupt himself!

A number of objections, some of them related to the preceding argument, have been raised against the money pump argument. Most of the commentators end up claiming that is inconclusive.[14] It is safe to say that other things being equal, acyclical preferences would save the candidate from bankruptcy, that is to say, glossing over the logical slide, from irrationality. But of course, the argument did not aim at establishing this trite point. It aimed at showing that acyclical preferences are the only ones which preserve him from irrationality. In my previous terminology, the argument aimed at showing that in the circumstances, acyclical preferences are not only weakly, but strongly rational. The failure to establish this conclusion exactly parallels the failure to establish that under other circumstances, cyclical preferences are strongly rational.[15]

3. OPTIMIZATION FROM THE PERSPECTIVE OF REVEALED PREFERENCE THEORY

Revealed preference theory analyzes the relationships between (a) the properties of individual choices when these choices bear on subsets of the full set of alternatives, and (b) the existence and properties of a binary preference relation on the full set. In Samuelson's (1938) original version, the theory applied only to the neo-classical consumer, which imposes a special structure on the set of alternatives and on the collection of choice subsets. In order to meet the requirements of social choice theory, a vari-

ation that is both more elementary and conceptually more general was devised after Arrow (1959). This version does not restrict alternatives in any way and describes choices just in terms of set-theoretic operations. It is the only one needed here. I will rely more particularly on Richter's (1966, 1971) and Sen's (1970, 1971) results which I have somehow combined. I will thus make it clear how the idea of optimization can be broken down axiomatically into conditions stated on the choice function.

This discussion will serve two purposes at once. First, formal equivalences have been invoked to justify optimization in terms of alledgedly 'natural' equivalent properties of choices. This justification strategy underlies much of the abstract literature on revealed preference (though not the initial Samuelsonian version, which was meant to be descriptive). To illustrate, I cite at some length Sen's early work, where it is put to use explicitly.[16] So one aim of this section is to assess a *prima facie* significant argument made for the claim that rationality implies optimization. I will dismiss the argument, a conclusion which connects with the second purpose of this section. It also serves to emphasize the *cognitive costs* of the mental operations by which an agent maximizes a transitive and complete preference relation. The theorems of revealed preference theory are one way of introducing this issue, even if few writers have thus far discussed it from this perspective – Plott (1973) being one leading exception. The theme of cognitive costs is central to the provocative claim that not only does not rationality imply optimization but even the converse implication may not hold.[17]

3.1. *Those Famous Conditions Alpha and Beta*

Formally, we are given a non-empty set X of unspecified objects of choice, a non-empty family Σ of subsets of X, not including $\emptyset$, and finally a function h: $\Sigma \to 2^x \setminus \{\emptyset\}$ satisfying $h(S) \subseteq S$, for all $S \in \Sigma$. Thus h picks out a subset of each set of options S. For this reason it is called a *choice function*. The condition $h(S) \neq \emptyset$ is universally accepted in the theory.[18] The technical question is, what conditions on h are equivalent to the property that 'h arises from optimization', or, more explicitly, that 'there exists a binary relation that the agent maximizes in each choice situation S'. This property will be defined formally as follows: there exists a transitive and complete binary relation, R, such that:

$$\forall S \in \Sigma,\ h(S) = \{x \in S | \forall y \in S,\ xRy\}$$
(= the set of the best elements in S according to R).

Other formal definitions of 'h arises from optimization' would be conceivable.[19] However, this particular one agrees best with the prelim-

inaries of Section 1 and my general emphasis, throughout this paper, on the two conditions of transitivity and completeness.

Consider first the particular case where the choice function is complete and single-valued, that is, $\Sigma = 2^X \setminus \{\emptyset\}$ and $\forall S \in \Sigma$, $h(S)$ is a singleton. Then it can be shown that h arises from optimization if and only if h satisfies property α, which is defined as:

$$\forall S, S' \in \Sigma (S \subseteq S' \text{ and } x \in h(S') \cap S) \Rightarrow x \in h(S).$$

'If the world champion in a particular discipline is a Pakistani, he must also be a Pakistani champion' (Sen 1970, 17). This condition has often been stated as a minimum rationality requirement in the social choice or game theory literature. Consider now the more general case where the choice function is still complete but not necessarily single-valued, i.e., we just require that $\Sigma = 2^X \setminus \{\emptyset\}$. Then, it can be shown that h arises from optimization if and only if it satisfies properties α and β. Property β is defined thus:

$$\begin{aligned}&\forall S, S' \in \Sigma \ (S \subseteq S' \text{ and } x \in h(S') \cap h(S)\\ &\Rightarrow (y \in h(S) \Rightarrow y \in h(S')).\end{aligned}$$

'If a Pakistani is world champion, then all the Pakistani champions are world champions' (*ibid*.). The conjunction of α and β is readily seen to be equivalent to:

$$\begin{aligned}&\forall S, S' \in \Sigma \ (S \subseteq S' \text{ and } h(S') \cap S \neq \emptyset)\\ &\Rightarrow (h(S') \cap S = h(S)),\end{aligned}$$

a condition introduced by Arrow (1959). It is sometimes called the strong axiom of preference.

Conditions α and β are most famous among those stated by revealed preference theory. Being apparently such weak rationality conditions, they provide interesting support to the orthodox economist's contention that rationality implies optimization. A first objection to this move is that they do not have the same normative force. It is customary to interpret α and β as referring to consistency conditions in a notional *sequence* of choices, and I will not depart from this tradition here. Then, α says: 'If one first discards an alternative when choosing from some subset, it will not be retained later when choosing from a more comprehensive subset than the former'. Condition β says: 'If one first selects two alternatives from some subset, and one later selects the first from a more comprehensive subset than the former, one will then retain the second option along with the first'. Thus, α prescribes that the selection process in which larger and larger choice

subsets are considered be not creative, while β prescribes that it be not destructive. The former requirement appears to be more basic than the latter. Suppose that $X = \{a, b, c\}$ and the choice function h is defined as follows: $h(\{a, b\}) = \{a, b\}$, $h(\{a, c\}) = \{a, c\}$, $h(\{b, c\}) = \{b, c\}$, $h(\{a, b, c\}) = \{a\}$. Let us interpret $h(\{a, b, c\})$ as the final choice on X, whereas the other values of h represent preliminary attempts at selection. For instance, X is a set of jobs the agent is considering applying to, and h serves as a coarse formalization of his deliberation. There seems to be no normative inconsistency involved in a deliberation of this sort; still, h satisfies α but blatantly violates β. Conversely, I cannot think of a choice function and an interpretation for it such that β might be satisfied, α violated, and no normative inconsistency is involved. I submit that the persuasive force of β, if any, does not come from rationality considerations but rather from accompanying *equity* connotations. Think of a sport championship in which three players compete with each other in pairs, and from these results a winning set is declared at the end; and now reinterpret h as representing a particular championship having taken place in this way. It would be unfair to oust a player at the expense of an equal, so that h now describes a normatively objectionable selection process. The persuasive force of Sen's commentary about β might well come from the ambiguity of his 'champion' example.

Note in passing that the following point has emerged again: individual choice theory does not allow for exactly the same formal considerations as does social choice theory. But in the context of intransitive choices, the claim that the two theories are disanalogous (since characteristics are not individuals) had the effect of weakening a criticism leveled against optimization; whereas, this time, the claim that they are disanalogous (since alternatives are not individuals) runs against a possible defense of optimization.

A second line of argument involves distinguishing α from α and β taken together in terms of the memory and computation requirements implicit in either axiomatization. If $X = \{a, b, c\}$ and the choice function h is defined on the pairs as follows: $h(\{a, b\}) = \{a, b\}$, $h(\{a, c\}) = \{c\}$, $h(\{b, c\}) = \{b, c\}$, α is compatible with three solutions for $h(\{a, b, c\})$, i.e., $\{b\}$, $\{b, c\}$ and $\{c\}$. By contrast, α and β taken together entail the unique solution $\{b, c\}$. Consider now the following simple sequential choice procedure: the agent takes an arbitrary pair, retains only one best element in the choice made from this pair, forms another pair with the remaining element, and again retains only one best element. This is an algorithm for computing $h(\{a, b, c\})$, and it is cost-effective. It saves memory space since from the pair, say, $\{a, b\}$, only a or b is retained as a result of the choice made on this

pair. The algorithm also economizes on calculations: at the next stage after $\{a, b\}$, the agent will compare c and a, or c and b, but not both. Obviously, the procedure satisfies α, but violates β, and thus the optimization model as it has just been axiomatized. If the procedure is now modified to satisfy β, the memory space will have to be expanded, and more calculations to be performed. For larger sets than the three-element one mentioned here, the difference in total operating costs can be considerable. These negative considerations have to do with rationality in the generic sense. Hence, they would have to be balanced against any normative argument that can possibly be made for β, with the likely conclusion that β cannot be as strong a rationality requirement as α.

3.2. *Completeness*

The previous discussion was confined to those choice functions which are defined on all non-trivial subsets, i.e., when $\Sigma = 2^X \backslash \{\emptyset\}$. To stop at this case is tantamount to taking for granted one of the two axioms of optimization that are at stake in this paper, i.e., completeness. Does revealed preference theory provide a normative argument for optimization when completeness is not taken for granted? I claim not. To argue for this point, I will consider incompletely defined choice functions and restate optimization in terms of transitivity only. I will then claim that the available equivalence result for this case does *not* provide a justification.

Let us say that h defined on an arbitrary set Σ of nonempty subsets is *binary* if some binary relation R exists, such that:

$$\forall S \in \Sigma, h(S) = \{x \in S | \forall y \in S, xRy\},$$

and that 'h arises from optimization' if it is binary and the underlying relation R is transitive. (In accordance with the purpose of this subsection, I am leaving completeness out of my previous definition.) Then, a first result of interest tells us that h is binary if and only if it satisfies property α^+:

$$\forall x \in X, \forall S \in \Sigma\ (x \in S \text{ and } [\forall u \in S, \exists S' \in \Sigma\colon u \in S', \text{and } x \in h(S')]) \Rightarrow x \in h(S).$$

This property is logically stronger than α, and not comparable with α and β together. It is often formulated in terms of the so-called *revealed preference relations*. Define: xVy ('x is directly revealed to be preferred to y') if there is S such that $x \in h(S)$ and $y \in S$. Then α^+ becomes:

$$\forall x \in X, \forall S \in \Sigma (x \in S \text{ and } [\forall u \in S, xVu]) \Rightarrow x \in h(S).$$

In words, x is directly revealed to be preferred to y if there is a choice subset (possibly different from the initial subset) from which x is chosen when both x and y can be. Condition α^+ thus imposes a form of coherence on the individual's multiple choices, and as such, has significant normative force. When he compares x and y in S, this comparison must implicitly take into account that the same elements might have already been compared in some other subset S'. The trouble is that α^+ still does *not* ensure that h arises from optimization.

There is a need for a stronger notion of revealed preference than V. The early developments of revealed preference theory testify to the fact that it is not obvious how to formulate it. Define: xWy ('x is indirectly revealed to be preferred to y') if $x^0, x^1, \ldots x^n$ exist such that

$$x^0 = x, x^n = y \text{ and } x^0 V x^1 V \ldots x^{n-1} V x^n.$$

In words, x is indirectly revealed to be preferred to y if there exists a sequence of sets $S_1, \ldots, S_{n-1}$, and of alternatives $x^1, \ldots, x^{n-1}$ chosen from these sets, such that x is directly revealed to be preferred to x^1, x^{n-1} directly revealed to be preferred to y, and each x^i directly revealed to be preferred to x^{i+1}. The conclusion now is that h arises from optimization if and only if h satisfies *property* κ (called *congruence* by Richter 1966):

$$\forall x, y \in x, \forall S \in \Sigma(x \in h(S), y \in S \text{ and } yWx) \Rightarrow y \in h(S).$$

(This result is essentially the set-theoretic trivialization of the theorem in consumer theory which Samuelson groped for, and which was finally proved in the 50's.) By comparison with α^+, which it implies, κ strongly reinforces the constraint imposed on multiple choices. The choice between any two elements of S, x and y, is now influenced by what earlier choices revealed *not only directly, but also indirectly* about x and y. Mathematically, W is the transitive closure of V.

On one reading, κ is just another coherence condition imposed on the decision-maker. I do not think that that this interpretation is very plausible. Consider again $X = \{a, b, c\}$ and assume now that:

$$h(\{a, b\}) = \{a, b\}, h(\{a, c\}) = \text{not defined}, h(\{b, c\}) = \{b, c\}.$$

The only solution conforming to κ for $h(X)$ is:

$$h(X) = \{x \in X | \forall y \in X, xWy\} = \{a, b, c\}.$$

Here, the missing information on $h(\{a, c\})$ has been replaced by a calculation. The comparison between a and c is made in terms of the statements:

$aVbVc \equiv aWc$ and $cVbVa \equiv cWa$. Emphatically, it is the decision theorist who makes the calculation of the W relation and draws the consequences. By assumption, there is no choice on the agent's part to parallel the theorist's comparison between a and c. Should the agent nonetheless repeat the theorist's mental process, which in the particular instance, leads to treat a, b and c completely symmetrically? There is at least one typical situation which calls for a negative answer – this is when the agent is *truly unable* to reach a conclusion concerning a and c. When considering an incompletely defined h, one certainly does not want to exclude this important possibility. For concreteness, think of the alternatives in the money pump example: a top-level job with low pay (a), an intermediate-level job with average pay (b), and a low-level job with high pay (c). Imagine a candidate expressing choices according to h. There seems to be nothing irrational, and there is even some crude plausibility, in the candidate's failure to reach a conclusion concerning the extreme options a and c.

I conclude that the transitive closure W is not automatically invested with a meaning in terms of the agent's activities or proclivities, and that on a natural interpretation of incompleteness, it is not. I do emphasize the distinction between V and W, and between the conditions α^+ and κ. A rational agent is in some sense committed by the earlier choices he made, but I cannot see the sense in which he is committed by the inferences *the observer draws* from these choices.

I will not expand on the cognitive costs implied by κ, precisely because I cannot see how to interpret this condition from the agent's point of view. The criticism just presented is consistent with the claim sometimes made even among classical theorists that the completeness axiom does not enjoy the same normative status as transitivity.[20]

3.3. *A Counterargument*

At this juncture, a defender of optimization can conceivably resort to a curious defence of optimization which was suggested – perhaps in passing – by Sen (1971, in 1982, 48–49). It consists in arguing that one of the initial special cases – that of a complete choice function – is, the appearances notwithstanding, the relevant one to consider. The gist of this argument is to justify optimization not in terms of κ, but of the milder conditions α and β. Why, Sen asks, should one restrict attention to a particular family Σ of subsets? The nature of the choice problem – for instance, in consumer theory – might entail fixing Σ – in that example, Σ is the set of all 'budget triangles'. But at the level of a general argument about rationality, no selection criterion presents itself. In the absence of a reason for selecting

any particular subset, it is appropriate, Sen claims, to adopt the set of *all* subsets of x.[21]

A curious feature of this argument is that it appears to reproduce at the metalevel of the theorist a mistake that decision theory is normally wary of making at the level of the agent, i.e., the confounding of non-comparability with indifference.[22] Revealed preference theory formally captures the agent's indifference by allowing for multi-valued $h(S)$, and non-comparability by allowing that h be not defined for all S. Now, consider a revealed preference theorist who does not have any reason to restrict the subsets relevant to the agent's choice. By recommending that this theorist should reason on the set of *all* subsets, Sen in effect analyzes the theorist's indeterminacy as a case of *indifference* between all subsets. I am following here the analogy provided by revealed preference theory: if the theorist's 'choice function' contains all possible subsets, this means that he is 'indifferent' between all subsets. But rather than being indifferent between the subsets, the theorist has failed to establish relevance comparisons between them. It is a case of non-comparability, not of indifference. If this is so, it seems right to investigate the properties of choice functions not for the single maximal domain that Sen recommends, but for all conceivable domains Σ. I do not want to suggest that Sen himself would lay much emphasis on this little piece of argument I tried to disentangle.

3.4. *Optimization and Path-Independence*

Another attempt to justify optimization can be made by appealing to an alternative decomposition in terms of path independence. Informally, this property says that the choice finally made from the whole set x should not depend on the path taken through the set Σ of choice subsets. Its first technical formulations are pre-war in origin, when microeconomists – following Pareto's lead once again – were investigating the integrability of demand functions. Abstracting from the rich framework of consumer theory, we get set-theoretic definitions of path independence, such as Plott's (1973):

(IP) $\quad \forall S, S' \in \Sigma, h(h(S) \cup h(S')) = h(S \cup S').$

Plott's axiom ensures that the choice made in any set T coincide with the result of the two-step choice defined by first choosing from within two subsets S and S', which together make up T, and second, choosing between the alternatives thus selected. One-step choices should coincide with two-step choices whatever the splitting of T into S and S', so that the axiom may be taken to state independence of path. It is easy to see that if h arises from optimization, it satisfies (IP), but an example in Plott

(1973, 1081) shows that the converse does not hold. At least, (IP) implies α (this follows from another observation of Plott's 1973, 1087), and this implication is strict. It thus follows that in terms of logical implication, path independence lies between optimization and condition α. So logically, the alternative justification of optimization still fails.

Condition (IP) seems to be an attractive stopping place to formally define a rationality concept implied by, but not implying, the optimization condition. This move would be a very clear instantiation of what I have called the conciliatory account in Section 1. If (IP) or similar conditions were not met, the individual's final choice could change even though the objective choice circumstances remain the same. Economists have long been emphatic that rationality entails that choices should remain invariant with respect to the external circumstances. The "rational consumer", for instance, reacts to the relative prices and his available budget, and to no other piece of information. In this view of rationality, the external circumstances being fixed, the outcome of a deliberation should not depend on the way it is conducted. However, there is another, quite opposite intuition about path independence and rationality. A rational choice, I mentioned at the beginning, is a choice made for good reasons. A well-conducted deliberation is by itself a good reason for the choice it results in. But crucially, the concept of a well-conducted deliberation does *not* involve that of a uniquely determined conclusion; that is, the external circumstances being fixed, another well-conducted deliberation could possibly lead to a different conclusion. It is the properties of the *procedure*, not of its *outcome*, that are referred to by the adjective 'well-conducted'. Simon's later work (e.g., 1976) usefully contrasts the 'procedural' view of rationality, against the economist's 'substantive' view, which is outcome-oriented.[23] The two conceptions depart from each other in the way they deal with the agent's *internal* choice circumstances, such as his memory or his computation abilities. Typically, the 'substantial' theorist either does not take internal circumstances into account, or somehow forces them into the description of the available means or other external circumstances. By contrast, a 'procedural' theorist regards internal circumstances as being distinctive factors of the choice. He can even accept the principle above, to the effect that rational choices should remain invariant with respect to the *objective* choice circumstances; it is enough for him to argue that internal choice circumstances can be objective. These are important distinctions to make, but I do not want to expand on them here (more on them, in Mongin 1986). I just wished to clarify the conflicting rationality intuitions surrounding Plott's (IP). The property this axiom formalizes is no more than weakly rational since not all of the *prima facie* intuitions of rationality

warrant path independence. The initial suggestion notwithstanding, one cannot make much of condition (IP).

I summarize this idiosyncratic review by stressing that it does not warrant the claim that an inference from rationality to optimization can be read into the theorems of revealed preference theory. Notice carefully that once again, the converse, more problematic inference from optimization to rationality has not really been addressed. However, the argument has brought with it a significant by-product. It shows that the issue of internal circumstances of choice must urgently be addressed. Using naive principle of cost-effectiveness – i.e., that memory demands strictly increase with the number of previously selected alternatives, and calculation costs with the number of operations performed – I have tried to discriminate between α and β. Interestingly, Plott (1973) did not proceed differently when at one point of his paper, he argued for some weaker forms of path-independence than (IP), and thus implicitly against optimization.[24] The simple point about cognitive costs is this. Any concept of rational choice implies making good use of the available means; this is part of the notion of an 'appropriate' choice. Accordingly, the agent's incurring cognitive costs as a consequence of complying with an alleged rationality condition must always count negatively in the overall rationality assessment of this condition. In the next section, I sharpen this simple point to argue that optimization can sometimes be not even weakly rational.

4. THE INFINITE REGRESS OF OPTIMIZATION

All decision principles lead to an infinite regress. Applying a principle is just another decision to make, and the question arises of whether or not this decision satisfies the given principle. As a particular application of this general problem, we have just seen that to optimize is, implicitly, to *decide* to implement a choice function of a certain kind. Supposing now that the cost of each kind of choice functions is known, one needs to ask whether it was optimal to implement a choice function of that particular kind. If one now assumes that to answer this question requires a choice function of higher order, which itself has an implementation cost, a new question arises, and so on *ad infinitum*. To investigate the problem exemplified by this reasoning I will now drop the reference to the choice functions of revealed preference theory. They have helped to motivate the discussion, but are just an example. It should be clear that any account of optimization is *prima facie* open to an infinite regress objection. A conveniently general formulation is to say that to optimize requires one to run a costly *algorithm*, that to optimally select the latter requires one to run another

costly algorithm, etc. In this section, I will explain, and then illustrate, how this problem can be turned into a significant objection against the optimizing model of rationality.

4.1. *Not Every Infinite Regress is 'Vicious'*

I first need to make a distinction between two kinds of infinite regress. In order to introduce it, I discuss (and rebut) Ryle's statement of the infinite regress of decision in *The Concept of Mind* (1949). In a well-known passage, he raised the following objection against those conceptions of action which he calls *intellectualist*:

> To put it quite generally, the absurd assumption made by the intellectualist legend is this, that a performance of any sort inherits all its title to intelligence from some anterior internal operation of planning what to do. Now very often we do go through such a process of planning what to do, and, if we are silly, our planning is silly, if shrewd, our planning is shrewd. It is also notoriously possible for us to plan shrewdly and perform stupidly, i.e., to flout our precepts in our practice. By the original argument, therefore, our intellectual planning process must inherit its title to shrewdness from yet another interior process of planning to plan, and this process yet another interior process of planning to plan, and this process could in its turn be silly or shrewd. The regress is infinite, and this reduces to absurdity the theory that for an operation to be intelligent it must be steered by a prior intellectual operation. (Ryle 1949, 31–2)

Ryle's critique of 'intellectualism' has attracted considerable attention from contemporary philosophers of mind. Certainly, this is a striking passage, but what exactly is its polemical target? All of the existing theories of rational choice describe the performance of choice as being steered by some "anterior internal operation of planning", and they assess the rationality of choice in terms of the rationality of this underlying process. So all these theories are 'intellectualist' and hence open to Ryle's sweeping objection. It would hit not only the optimization model but also any of its tentative competitors, such as Simon's 'satisficing'. But is there an objection after all? I think not – at least, not for the reasons Ryle suggests. He evidently believes that in order to reject a theory, it *suffices* to show that it leads to an infinite regress (cf.: "the regress is infinite, and this reduces to absurdity the theory"). To see that this cannot be the case, I will consider two examples.

Take the standard assumption in game theory that the rules of the game (i.e., the sets of strategies and the pay-offs) are common knowledge. By definition, this means that each player knows the rules of the game, knows that each player knows the rule of the game, and so on. Are we to dismiss the common knowledge assumption as being 'absurd' on the grounds that (for suitably large state spaces) no finite level of mutual knowledge reached between two players can exhaust the content of the assumption? Or, to take

a more explicitly decision-theoretic counterexample, consider this other familiar notion in game theory – rationalizability. Informally, a strategy s_1 is rationalizable for player A if it is a best reply to some strategy t_1 of player B, which is itself a best reply to some strategy s_2 of A, the latter being itself a best reply to some strategy t_2 of B, and so on, possibly *ad infinitum*.[25] To paraphrase Ryle, A's planning to play s_1 inherits its shrewdness from another act of planning (in this instance, by B) to play t_2, and "the regress is infinite". Are we to reject rationalizability on this ground? This would be a ridiculous inference, as in the common knowledge case, and here is a quick argument why it should be resisted: there are restatements of both rationalizability and common knowledge which simply make no reference to infinity.[26]

The situation exemplified by these two concepts is typical. Not every infinite regress is logically vicious, and one way to recognize this point is to see whether the infinite regress apparently entailed by a concept or a thesis manifests itself in all of their equivalent restatements. Generally, infinite regresses can be regarded as infinite sequences for which there exists some appropriate notion of convergence. Then, a regress will be said to be 'vicious' or 'harmless' depending on whether the associated infinite sequence is divergent or convergent. Viewed this way, infinite regress arguments cannot be expected to deliver ready-made refutations against a whole class of theories as they allegedly do in Ryle (e.g., 'intellectualist' theories), and not even against all the instantiations of a single theory. The convergence properties of a sequence are sensitive to the values of its terms, and an instantiation of a theory is typically associated with specific parameter values. The most natural definition of convergence in the present context of decision-making is *stationarity* for some integer n – the level $n + 1$ leading back to the n-level decision, and identically for the levels after $n + 1$. That is, an infinite regress of decisions is 'harmless' if there exists a logical level such that the decision recommended by the given principle at this level coincides with the decision recommended by the same principle at all subsequent levels. Alternatively and less plausibly, it would be conceivable to use the more general notion of *asymptotic convergence* (relative to either the space of level 1 decisions, or, perhaps with different results, to the space of utility values).

I now move to the next issue of how to apply the general method sketched here to optimization.[27]

4.2. *A Simple Example of the Infinite Regress of Optimization*

When applied to optimization, the infinite regress can be destabilizing in two distinct ways. First, an optimal solution in the ordinary sense can be

replaced by a meta-optimal solution which is in turn replaced, and so on, level $n+1$ never 'supporting' level n. Second, the optimizing method *itself* can give way to another, for example, the search for a 'satisficing' value, a method which, in turn, is liable to be evicted, and so on.[28] Another natural distinction to make cuts between two kinds of internal costs incurred by an optimizer, i.e., *calculation* and *information* costs. The former are incurred when the optimization problem is well-defined and the optimizer is trying to solve it. The latter summarize the resources expended in formulating the problem itself, that is, in constructing the optimizer's set of alternatives and objective function. This distinction permeates Simon's (1955) attempt to model bounded rationality as well as his work on automated chess-players (1979).[29]

I will try to illustrate, by means of a very simple model, an infinite regression of the simpler type in terms of the first distinction (i.e., replacing of the optimal decision but not of the optimizing method itself). In terms of the second distinction, the model involves only calculation costs; this is just to make it more transparent because it would be possible also to include information costs. Given E, the set of states of nature, which I assume to be finite with at least two distinct elements, D, the (also finite) set of decisions, and U, the utility function: $E \times D \to R$, the optimizing theory of decision says that the agent finds a rule

$$r_1^* : E \to D$$

satisfying the optimization condition, that is to say, such that for each e:

$$(1) \qquad U(e, r_1^*(e)) = \max_d U(e, d) = \max_{r_1 \in D^E} U(e, r_1(e)).$$

(As usual, D^E denotes the set of all functions from E to D.) Let us now assume (that the theorist attributes to the agent)[30] the following cost function on rules:

$$C_1 : E \times D^E \to R.$$

Then (according to the theorist) the agent must find a rule

$$r_2^* : E \to D^E,$$

which to each e assigns a level 1 rule $r_2^*(e)$ such that:

$$(2) \qquad U(e, r_2^*(e)(e)) - C_1(e, r_2^*(e)) = \max_{r_1 \in D^E} [U(e, r_1(e)) - C_1(e, r_1)].$$

Comparing (2) and the second equation in (1), there is nothing to ensure that there is e satisfying:

$$r_2^*(e) = r_1^*$$

or even such that:

$$r_2^*(e)(e) = r_1^*(e).$$

And one can a fortiori change 'there is e' into 'one has for all e' in the preceding sentence. Briefly, meta-optimization can very well contradict optimization.

Let us take a particular application where the agent is a business, assuming that its production function depends only on the factor labour. The state variable e that is relevant to his decision-making is the wage rate, U is the business's profit function. This gives some flesh to the maximization programme (1). To explicate (2), let us now assume that the business employs programmers with the job of determining the optimum level of production, and they are paid the same rate e. The difference between (2) and (1) arises because the business is trying to 'internalize' its programming expenses, as measured by C_1, which it had not originally taken into account. The result of the second calculation can obviously upset the result of the first.

What has been said for the first two levels apply to all others. Accordingly, I will define:

– an infinite sequence of sets of rules:

$$R_1 = D^E, R_2 = D^{E^E} \approx D^{E\times E}, \ldots$$

– a corresponding sequence of cost functions:

$$C_1 : E \times R_1 \to R, C_2 : E \times R_2 \to R, \ldots$$

The optimization programme of order n generalizes condition (2). As the preceding discussion also illustrates, to calculate the optimum level of output (both physical and intellectual) at level $n-1$ imposes on the business a cost determined by C_{n-1} – a cost which is taken into account only in the programme of order n. No programme automatically takes account of its own cost. The observation that meta-optimization can contradict optimization now applied to programmes of any order n.

Let us say that an infinite regression *converges for all* $e \in E$ (*for some* $e \in E$) if $n \in N$ exists such that for all $e \in E$ (resp. some $e \in E$):

$$r_n(e) = r'_n(e).$$

Various plausible assumptions can be introduced on the programming costs, and some of them quickly lead to the conclusion that infinite regresses typically do not converge. I just mention one particularly easy case.

ASSUMPTION:

$$(\forall n) \quad (\exists r_n, r'_n \in R_n)(\forall e \in E) \\ |\mathrm{Im} r_n| > |\mathrm{Im} r'_n| \Rightarrow C_n(e, r_n) > C_n(e, r'_n).$$

(That is, *ceteris paribus*, a rule costs more to implement the more distinct decisions it includes.)

Now, if the Assumption holds, then either r_1^* is constant, or the infinite regress does not converge for any e. The proof is immediate. In words, if r_1^* is not constant, then it pays at level 2 to associate each e with the constant rule having value $r_1^*(e)$. That is, r_1^* is evicted by r_2^*. By a similar argument, r_2^* is evicted at the next level, since it is not a constant rule. And so on *ad infinitum*. Alternative assumptions that are just as easy to formulate would prevent the convergence of the infinite regress for at least *some* e.

The little model of this section has a special feature which calls for a comment.[31] The variable e influences *all* levels of decision. However, not every meta-optimization problem, even of the simpler sort, has this particular structure. For concreteness one can imagine that the same programmers are called upon at each stage to perform a calculation that they had not performed at the previous stage. To generalize beyond this concrete case, one could hypothesize that each level n defines a new category of programmer especially suited to his task, but that the wage rates e_n of the different categories are fixed proportionally to one another, so that the cost functions C_n can all be re-expressed as functions of the single variable $e_1 = e$. The infinite regress would completely change in character if the choice of e did not influence all the costs C_n at the same time.

4.3. *Some Counterarguments*

From discussing the issue with both economists and philosophers[32] I have found out that a defender of optimization is likely to dismiss the infinite regress objection in one of these three ways: (a) by flatly denying that it should be part of the assessment of decision-making principles; (b) by saying that it is *a priori* relevant, but in effect undecidable, because – allegedly – individuals only reason at finite levels of thought; (c) by claiming that the objection affects all principles of decision in the same way and thus cannot

be part of an argument specifically aimed against optimization. I am going to see whether, and how, the infinite regress objection can be sharpened against these points taken in turn. In this way, I am to make precise the objections not so much in the absolute as dialectically. When I say here 'a defender of optimization' I mean someone who believes that optimization implies rationality.

I need some terminology to organize the discussion. Let us say that a particular application of a principle of decision is *reflectively coherent* if this principle leads to the same conclusions whether it is applied just to the objective circumstances or *both* to the objective circumstances *and* to the way in which it was applied to the objective circumstances. Then, the infinite regress objection amounts to arguing in formal detail that optimization is a reflectively incoherent principle. I reconstruct (a), as being the claim that reflective coherence is an irrelevant notion; (b), that it is an *a priori* relevant but undecidable notion; and finally, (c), that it is both *a priori* relevant and decidable, but that all decision principles turn out to be reflectively incoherent.

Before I proceed, a word should be said in defence of the objective of reflective coherence. It has to do with an utterly general but I think, normatively compelling, metaprinciple of action, which is to check the consistency of one's principle of action with the particular way in which it is employed. In ethics, a metaprinciple of this sort requires one to avoid using means that would apparently be efficient to reach one's moral end but undermine it in a deeper sense; for instance, it condemns the imposition of tolerance by brutal means such as physically eliminating the intolerant members of society. Reflective coherence, as I define it, is but a version of this broad normative metaprinciple that seems to be well suited for decision principles. Why has the reflective coherence of the optimization principle to be addressed at all? This is simply, it seems, because the latter is currently being the object of a normative inquiry. But I can also easily connect the concern for reflective coherence with the generic notion of rationality. If a particular application of a principle is reflectively incoherent, this is certainly a 'good reason' for giving up this principle, at least in the circumstances of this application; and it is possible to discard the choice resulting from this application on the ground that it is 'inappropriate'.

Having said this, I have already disposed of the disqualifying claim a). Actually, it is typically made in contexts where there is a doubt on what kind of inquiry is being conducted. In terms of assessing the predictive value or descriptive accuracy of the economists' optimizing model, reflective coherence may or may not be a pointless consideration. In the context of this paper, I simply cannot see how it can be side-stepped. Position a)

is not only congenial to orthodox economists, but it also appears to be Simon's; at least, it is compatible with his complete silence on the infinite regress argument.[33] I interpret Simon's attitude as being influenced by the primarily positive style of inquiry he is conducting.

Claim (b) stresses that human individuals cannot cope with an infinite number of reasoning stages, as well as perhaps the following more subtle point: beyond a certain reflective level n, they can no longer *define* their choice objects (here, the algorithms) nor, *a fortiori*, the associated costs. This claim is mostly impressive by dint of a hidden assumption – i.e., that the infinite regression actually sets out actual stages to be gone through by the individual. Even granting this literal interpretation, the claim is formidable only if it impossible for a human to cover an infinite number of logical stages in a finite time. But clearly, the time a single stage takes might be very small. The model of converging series or sequences could be invoked here again, just as it served in the early days to eliminate the alleged paradoxes of the impossibility of movement.

Another way of disposing of claim (b) is to object that it cannot be consistently sustained by a defender of optimization. I understand that those who make it do agree that reflective coherence is a *prima facie* relevant consideration. So they implicitly make their defence of optimization dependent on reflective coherence. But then, if the reflective coherence of optimization is ultimately undecidable, so is their own defence. Incidentally, it would have been curious if, for once, a 'naturalist' point like that made in (b) could have rescued optimization.

I have another, more complex argument to offer against those defenders of optimization who excuse themselves on the ground that actual agents are incapable of defining higher-order costs beyond a point. Even if these costs make no sense from the agent's point of view, they can very well make sense from the optimizing theorist's point of view. It is perfectly compatible with the 'naturalist' point contained in (b) that the economist makes an assumption of well-defined higher-order costs. I further argue that he should *of necessity* make this assumption. Or else he will have no analytical means of defending the coherence of his own approach. Suppose that the defender of optimization has claimed that in point of fact, agents are unable to define costs beyond some given level n. On one reading of this claim, agents in his view optimize up to level n, but no farther than that. Hence they might not optimize in the real sense, and the presumption is that they do not. Hence my opponent's theory is not really an optimizing theory; he is incoherent. Luckily for him, there is a more palatable interpretation of the claim that agents optimize up to level n but no farther. This interpretation says that they optimize at levels $n + 1$,

$n + 2, \ldots$, but are unaware of that. But in order to make this point, the defender of optimization will have to make the assumption of well-defined cost functions beyond level n. Willingly or not, he will be involved in the infinite regression argument if he is to eschew the charge of incoherence.

Emphatically, the last paragraph involves a shift from *the agent*'s reflective coherence to *the theorist*'s conceptual coherence. That is, supposing for the sake of the discussion that the initial version of the infinite regress is not damaging for the optimizing agent, I have tried to propose one which is worrying for the theorist. The optional parenthetic statements added here and there in Section 4.2 were intended for this version.[34] It is important to offer it as a rebuttal, given the widespread tendency among economists to interpret the optimizing agent's deliberations non-realistically. They might interpret cost functions non-realistically only from some level on, as I assumed in the discussion of the last paragraph. They might interpret the utility function realistically, and the cost functions non-realistically. They might even interpret both the utility and costs non-realistically – a widespread view in effect. Whatever the exact move, it has the effect of shifting the difficulty from the agent to the economist.

As to claim (c), it has an unsophisticated variant which has already been covered. If one believes that all decision principles lead to an infinite regress (an indisputable point), and one also believes, like Ryle, that an infinite regress *ipso facto* constitutes an objection, the conclusion follows that the objection does not help discriminate between particular principles. Here, the conclusion has exactly the cash value of the second part of the antecedent. However, there is a sophisticated version of (c), which is based on the presumption that under similar circumstances, both optimization and the other principles of decision lead to divergent infinite regresses. The first thing to mention is that this statement does not sound like a satisfactory *defence* of optimization. Rather, it sounds like saying that optimization is as unjustified as any alternative. Second, I do not even think that it is a very plausible statement to make. In the previous model of the firm, think of decision routines such as "optimize up to some level n, and at higher levels n', systematically adopt the constant rule recommending the rule of level $n' - 1$". These routines are very different from optimization, which consists in recalculating the n'-optimizing solution for each level n'. In effect, they are boundedly rational, not optimizing. Trivially, the implied sequences of costs and of higher-rules are stationary, whereas it has been said that the corresponding sequences for optimization are not. (These pseudo-optimizing routines may be grossly inefficient, and thus *very* boundedly rational – but this is a consideration different from con-

vergence.) This is a crude comparison, but it casts doubts on the claim that the infinite regress argument treat all principles of decision alike.

5. CONCLUSIONS

I have tried to question the standard optimizing model of rationality from different angles. I repeat the basic distinction between two critical purposes of the paper. The brief review of the money-pump argument in Section 2, and the more original discussion of revealed preference theory in Section 3 added further evidence against the (still not uncommon) claim that rationality implies optimization. The main critical purpose, however, was to challenge the converse statement, and the conciliatory account which goes along with it. In a similar paradoxical vein, Fishburn sought to establish 'the irrationality of transitivity in social choice' (1970b). Quite obviously, both here and in Fishburn's attempt, the aim was not to conclude that transivity or optimization was always irrational, only that it sometimes was. To avoid crude misunderstandings, I will once again exploit the analogy between individual and social choice, and let the reader transfer what Fishburn said of his problem to mine:

> The examples presented show that social transitivity is untenable as a *general desideratum* for social choice functions. This does not say that [the social preference relation] should be intransitive for *every* profile of preference but only that there are [some profiles of preference] for which transitivity [of social preference] should not be required. (p.122, my italics)

Another important distinction cuts between two ways of discussing optimization – either in terms of its underlying conditions, as in Section 2, or generally, as in Sections 3 and 4. On one reading,[35] optimization is only the act of selecting a best element for the preference relation or for the utility index representing the preference relation, and it should be kept distinct from its preconditions. I suggested in Section 1.1 that this construal would hinder the discussion. This can now be confirmed with the benefit of hindsight. Think of an individual satisfying the κ condition of revealed preference theory. As this agent's choices unfold, he is simultaneously making optimizing choices and demonstrating to the observer that his underlying preference satisfies one formal precondition of optimization, i.e., transitivity. There is no sense here in discussing the act of optimizing separately from its precondition. Even the infinite regress argument against optimization is not limited to the act of optimizing, as it may appear at first sight. In the example I formalized, only calculations costs were involved. But information costs too should in principle be included, so that the firm's

k-level problem will become also one of *finding* a suitable cost function C_k, instead of being solely the problem of maximizing given C_k. This larger deliberation blurs the supposedly sharp distinction between optimization in the narrow sense and its antecedent conditions.

As I showed, the most damaging examples of intransitive choices are still compatible with the claim that optimization is at least *weakly* rational, and the assessment of revealed preference conditions also is. The only significant argument against the claim that optimization implies rationality is the infinite regress of optimization.[36] I emphasize that the latter does not, alas, constitute the matter of an impossibility theorem, only of an *argument*. It is effective if the burden of proof falls upon the optimizing viewpoint. This makes the whole paper a dialectical exercise rather than a demonstration. But on balance, I hope to leave the reader with an asymmetrical impression, i.e., that the view that optimization implies rationality is less plausible than the contrary view.[37]

To establish that optimization sometimes takes the agent away from rationality, or that it takes the social scientist away from a coherent theory, would have serious methodological consequences. In the absence of suitable restrictions, the optimizing model could not take advantage any more of standard construals devised for rationality in general – such as Weber's 'ideal-types', Popper's 'situational logic', or Davidson's 'principle of charity'. But when it comes to methodology, other considerations will of course influence the social scientist's judgement. For instance, the analogies between the principle of least action in physics and the maximization hypothesis in microeconomics have been repeatedly stressed, and they were arguably part of some of the best economists' heuristics (at least, Pareto and Samuelson). Exploiting this sort of analogies, it is possible to say something for optimization independently of any rationality considerations. To wit: maximizing theories derive significant benefits from their generalizability, mathematical simplicity, elegance, and heuristic fruitfulness.[38] Maximizing theories, not only of economists and biologists, but even of some physicists, are thought to be difficult to test, but even this handicap has sometimes be turned into an advantage in the name of the necessary continuity of research programmes. To evaluate these and other related justifications was not part of the normative assessment made here.

ACKNOWLEDGEMENTS

This essay has evolved from a translation of an earlier French paper. I am grateful to Ann Broome and Richard Bradley for having prepared

this translation. I am also grateful to Nick Baigent and the Public Economics group at the University of Graz, John Broome, Dan Hausman, Frédéric Laville, Wlodek Rabinowicz and two anonymous referees for their constructive comments.

NOTES

[1] After Weber (see Weber, 1949) social scientists usually define instrumental rationality in terms of appropriateness, but I will also make use of an alternative definition in terms of reasons that is widely used in philosophy. Note that all these notions, and arguably even the question of this paper, can be traced back to Aristotle's discussion of *proairesis* in the *Nichomachean Ethics*.

[2] Related to the objective of this paper is a book by Michael Slote, *Beyond Optimizing* (1989). This work also attempts to go beyond Simon's conclusion by arguing that in some circumstances, it is irrational to optimize. However. Slote's strategy to establish this point is different from mine; in particular, he does not investigate the infinite regression problem.

[3] I am focusing on this particular equivalence theorem because it is famous among economists and the conditions of transitivity and completeness are universally known. There are, however, alternative formal renderings of optimization that are slightly less demanding (mostly in terms of the acyclicity condition, to be discussed in Section 2.2). I do not think that these variants would call for substantial changes in the overall argument.

[4] In the wake of Popper (1967) there has been a debate on whether the "rationality principle" was analytic or synthetic. Popper's followers (e.g., Watkins 1970) typically claim that it is empty of substance but synthetic nonetheless. This discussion is often phrased in the language of "situational logic" and assumes that the "situation" is given to the individual in the same way, roughly, as the utility function and constraints are given to the economic textbook's agent. In both cases, the theorist needs a *supplementary* claim – i.e., that the agent draws the consequences from the "situation", or that the agent effectively maximizes his utility function under the constraints - in order to make use at all of the principle (respectively: of rationality, of optimization).

[5] Expected utility theory has a continuity axiom which, by contrast, is often construed as being significantly loaded. For an early discussion, see Marschak (1950).

[6] This position is not unlike Becker's (1976), who distinguishes between rationality in general and microeconomic rationality.

[7] By a 'maximal element' I mean one to which no element in the choice set is strictly preferred. A preference relation is said to be acyclic if, when an individual strictly prefers a_1 to $a_2, \ldots, a_{i-1}$ to $a_i, \ldots, a_{n-1}$ to a_n, he does not strictly prefer a_n to a_1, irrespective of the length n of the chain of strict preferences. See, e.g., Sen (1970).

[8] In this paragraph I am indebted to a discussion with John Broome.

[9] For one reason or another, most commentators have resisted the suggestion that the horserace presented a real paradox. See the commentaries at the end of Blyth's article, in particular those by Good and Winckler (Lindley et al., 1972, 375).

[10] Tversky (1969) and others have investigated the formal disanalogies between the two modes of aggregation. They coincide only in particular cases – roughly, when the aggregation function U is separable with respect to the $u_i(x_i)$. In the field of risky choice, regret theory (Loomes and Sugden 1982) is another application of the non-classical method,

leading again to intransitivities. More generally, Fishburn (1982, 1988) has extensively investigated intransitive variants of expected utility theory. Like continuity, the transitivity axiom is seen as less compelling in the field of risky choice than it is in the field of choice under certainty.

[11] Nick Baigent has also pointed out the decision-theoretic analogue of the Pareto-Extension rule, which amounts to treating as being indifferent those options which are not Pareto-ranked. However, if the agent is to make a decision, he has to find a way of breaking the ties.

[12] The first and second points of this paragraph are very much Tversky's (1969, 1972a, b).

[13] My discussion of the money pump argument crucially depends on allowing for the possibility of making no choice at all.

[14] Beside Schwartz (1986, 128–131) see, among others, Anand (1987, 1993), Bar-Hillel and Margalit (1988), Schick (1986) and Sugden (1985). It should be stressed, however, that the argument can become rather sophisticated when pursued in terms of dynamic rationality principles; see McClennen (1990) and Rabinowicz (1995).

[15] There have been other suggestions less famous than the money pump argument to show that transitivity is a *strongly* rational property. The reader will find further references and arguments in Fishburn (1991). He concludes that "reasonable people sometimes violate transitivity and may have good reasons for doing so" (p. 131).

[16] Sen's (1997) recently published work on maximization and choice functions has taken a different direction, and is generally more critical of the standard model. I will not go into it because it does not overlap with the argument of this paper.

[17] Section 3 was discussed most carefully by Nick Baigent, Daniel Eckert, Ben Lane and Hans-Peter Weikard during a seminar in Graz. I am very grateful for their many comments, not all of them, however, I have been able to echo here. In particular, because this section is part of a broader argument, I had to refrain from covering the ingenious axiomatic variants that they pointed out to me.

[18] The value $h(S) = \emptyset$ could have been taken to mean that no choice is made from S. Instead, the theory covers his possibility by assuming that h is not defined on S.

[19] Especially, definitions based on acyclicity – as pointed out by my Graz colleagues.

[20] See for instance Aumann: "of all the axioms of utility theory, that of completeness is perhaps the most debatable" (Aumann 1962, 446). Luce and Raiffa (1957, 25) have also criticized this axiom.

[21] "Why therefore restrict the domain of an axiom to $[\Sigma]$ and not to $[2^X \setminus \{\emptyset\}]$ when (a) the satisfaction of [the axioms] is not possible either in the case of $[\Sigma]$ or in that of $[2^X \setminus \{\emptyset\}]$, and (b) there is no a priori reason to expect that the axiom is valid on $[\Sigma]$ but not on [the complement of Σ]" (Sen 1982, 48). Condition (a) need not retain us here; it refers to the point that there are typically too many subsets for Σ to be observable.

[22] In a different (and actually more critical) piece on revealed preference theory Sen (1982) has usefully discussed this problem. It connects with the philosophical conundrum of Buridan's ass (on which, see Rescher 1982).

[23] Correspondingly, Simon (1976) de-emphasizes the distinctions of his earlier papers between 'bounded' and 'absolute' rationality, or 'satisficing' and 'optimizing'. Mongin (1986) argues that the contrast between procedural and substantive models of rationality is more fundamental than these nonetheless more famous distinctions.

[24] See also Campbell's (1978) definition of 'calculation viability' which follows Plott's intuitions. The theory of algorithmic complexity would clearly be to the point here. Mark Johnson's work on choice functions is a step in this direction.

[25] The sequence is infinite only if there are an infinite number of strategies, which happens in particular when the rationalizability concept is applied to *mixed* strategies.
[26] As to rationalizability, compare for instance Bernheim's (1984) initial definition with a non-iterative restatement in Bernheim (1986). As to common knowledge, compare Aumann's (1976) sequential definition of common knowledge with the unproblematic he also gives in terms of the meet of the agent's partitions, and compare, more generally, the iterative with the circular or fixed-point approach to common knowledge (Lismont and Mongin, 1994).
[27] There have been only few systematic discussions of the infinite regress of optimization. To the best of my knowledge, they are those of Göttinger (1982), Mongin (1984), Mongin and Walliser (1988) and Lipman (1991). Conslick (1996) and Laville (1998) include the issue in their recent surveys of bounded rationality.
[28] More on the latter in Mongin and Walliser (1988). It is, however, a much more difficult issue to tackle than the former.
[29] Compare with Conslick's (1988) discussion of 'optimization cost'.
[30] This, and other parenthetic statements below, are meant to alert the reader to an alternative interpretation of the formalism available *in terms of theoretical steps taken by an ideal observer*. More on this interpretation below.
[31] Jean-Pierre Dupuy once raised this issue during a seminar.
[32] Among the latter, Wlodek Rabinowicz, whose comments were very helpful.
[33] It was not Simon, but Winter (1975, 81–85) who first alluded to the infinite regress problem in the course of discussing bounded rationality. Actually, rather than the infinite regress itself, Winter alluded to reflective coherence (in his words: the 'optimization which takes account of its own cost')
[34] See footnote 29.
[35] Apparently endorsed by Dan Hausman and Wlodek Rabinowicz in earlier discussions.
[36] In its first occurrence in Mongin (1984), the infinite regression argument was meant to question the claim that rationality implies optimization, and thus *support* the conciliatory account. The present paper puts it in its proper perpective.
[37] There is an interesting dialectical precedent in the philosophy of decision theory, which in some sense sets a standard for papers like the present one – McClennen's "Sure-thing doubts" (1983). It confronts technical principles (there, von Neumann independence and the "sure-thing principle") with pretheoretic concepts of rationality. McClennen's discussion is not entirely conclusive but oriented nonetheless, and it leaves the reader on the impression that one side of the argument is stronger than the other.
[38] Interestingly, the decision theorist Schoemaker's (1991) wide survey of optimization appears to eventually favour a defence in terms of these physical and biological analogies.

REFERENCES

Anand, P.: 1987, 'Are the Preferences Axioms Really Rational?', *Theory and Decision* **23**, 189–214.
Anand, P.: 1993, 'The Philosophy of Intransitivite Preference', *Economic Journal* **103**, 337–346.
Aristotle, *The Nichomachean Ethics*, in H. H. Joachim (ed.), Oxford University Press, Oxford, 1951.

Arrow, K.J.: 1959, 'Rational Choice Functions and Orderings', *Economica* N.S., **26**, 121–127.

Aumann, R. J.: 1962, 'Utility Theory Without the Completeness Axiom', *Econometrica* **30**, 445–462.

Aumann, R. J.: 1976, 'Agreeing to Disagree', *Annals of Mathematical Statistics* **4**, 1236–1239.

Bar-Hittel, M. and A. Margalit: 1988, 'How Vicious Are Cycles of Intransitive Choice?', *Theory and Decision* **24**, 119–145.

Becker, G. S.: 1976, *The Economic Approach to Human Behavior*, The University of Chicago Press, Chicago.

Bernheim, D. B.: 1984, 'Rationalizable Strategic Behavior', *Econometrica* **52**, 1007–1028.

Bernheim, D. B.: 1986, 'Axiomatic Characterizations of Rational Choice in Strategic Environments', *Scandinavian Journal of Economics* **88**, 473–488.

Blyth, C. R.: 1972, 'Some Probability Paradoxes in Choice from Among Random Alternatives', *Journal of the American Statistical Association* **67**, 366-373; and 'Rejoinder', 379-381.

Campbell, D.: 1978, 'Realization of Choice Functions', *Econometrica* **46**, 171–180.

Conslick, J.: 1988, 'Optimization Cost', *Journal of Economic Behavior and Organization* **9**, 213–228.

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

Davidson, D., J. C. C. McKinsey, and P. Suppes: 1955, 'Outline of a Formal Theory of Value, I', *Philosophy of Science* **22**, 140–160.

Debreu, G.: 1959, *Theory of Value*, Cowles Foundation Monograph, Yale University Press, New Haven.

Fishburn, P. C.: 1970a, *Utility Theory for Decision Making*, Wiley, New York.

Fishburn, P. C.: 1970b, 'The Irrationality of Transitivity in Social Choice', *Behavioral Science* **15**, 119–123.

Fishburn, P. C.: 1982, 'Nontransitive Measurable Utility', *Journal of Mathematical Pschology* **26**, 31–67.

Fishburn, P. C.: 1988, *Nonlinear Preference and Utility Theory*, Johns Hopkins University Press, Baltimore.

Fishburn, P. C.: 1991, 'Nontransitive Preferences in Decision Theory', *Journal of Risk and Uncertainty* **4**, 113–134.

Göttinger, H. W.: 1982, 'Computational Cost and Bounded Rationality', in W. Stegmüller, W. Balzer and W. Spohn (eds.), *Philosophy of Economics*, Springer, Berlin.

Harsanyi, J. C.: 1976, *Essays on Ethics, Social Behavior and Scientific Explanation*, D. Reidel, Dordrecht.

Laville, F.: 1998, 'Modélisation de la rationalité limitée: de quels outils dispose-t-on?', *Revue économique* **49**, 335–365.

Lindley, D. V., I. J. Good, R. L. Winckler, and J. W. Pratt: 1972, 'Comment', *Journal of the American Statistical Association* **67**, 373-379

Lipman, B.: 1991, 'How to Decide How to Decide How to . . . : Modeling Limited Rationality', *Econometrica*, **59**, 1105–1125.

Lismont, L. and P. Mongin: 1994, 'On the Logic of Common Belief and Common Knowledge', *Theory and Decision*, **37**, 75–106.

Luce, R. D.: 1956, 'Semi-order and a Theory of Utility Discrimination', *Econometrica* **24**, 178–191.

Luce, R. D. and H. Raiffa: 1957, *Games and Decisions*, Wiley New York.

McClennen, E. F.: 1983, 'Sure-Thing Doubts', in B. P. Stigum and F. Wenstop (eds.), *Foundations of Utility and Risk Theory with Applications*, D. Reidel, Dordrecht, p. 117–136.

McClennen, E. F.: 1990, *Rationality and Dynamic Choice*, Cambridge University Press, Cambridge.

Malinvaud, E.: 1971, *Leçons de théorie microéconomique*, Dunod, Paris.

Marschak, J.: 1950, 'Rational Behavior, Uncertain Prospects, and Measurable Utility', *Econometrica* **18**, 111–141.

May, K. O.: 1954, 'Intransitivity, Utility and the Aggregation of Preference Patterns', *Econometrica* **22**, 1–13.

Mongin, P.: 1984, 'Modèle rationnel ou modèle économique de la rationalité?', *Revue économique*, **35**, 9–64.

Mongin, P.: 1986, 'Simon, Stigler et les théories de la rationalité limitée', *Information sur les sciences sociales/Social Science Information* **25**, 555–606.

Mongin, P. and B. Walliser: 1988, 'Infinite Regressions in the Optimizing Theory of Decision', in B. Munier (ed.), *Decision, Risk and Rationality*, D. Reidel, Dordrecht, pp. 435–457.

Plott, C. R.: 1973, 'Path Independence, Rationality, and Social Choice', *Econometrica* **41**, 1075–1091.

Popper, K. R.: 1967, 'La rationalité et le statut du principe de rationalité', in E. M. Claassen (ed.), *Les fondements philosophiques des systèmes économiques*, Payot, Paris, pp. 142–150.

Rabinowicz, W.: 1995, 'To Have One's Cake and Eat It, Too', *Journal of Philosophy* **00**, 586–620.

Rescher, N.: 1982, 'Choice Without Preference: A Study of the History and the Logic of the Problem of "Buridan's Ass" ', in *Essays in Philosophical Analysis*, University Press of America, Lanham, chap. 5.

Ryle, G.: 1949, *The Concept of Mind*, Hutchinson, London.

Samuelson, P. A.: 1938, 'A Note on the Pure Theory of Consumer Behavior', *Economica* **5**, 61–71.

Schick, F.: 1986, 'Dutch Bookies and Money Pumps', *Journal of Philosophy* **83**, 112–119.

Schoemaker, P. J. H.: 1991, 'The Quest for Optimality: A Positive Heuristic of Science?', *Behavioral and Brain Sciences*, **14**, 205–215; followed by comments by other authors, 215–237; and 'Author's Response', 237–240.

Schwartz, T.: 1986, *The Logic of Collective Choice*, New York, Columbia University Press.

Sen, A.: 1970, *Collective Choice and Social Welfare*, Holden Day, San Francisco.

Sen, A.: 1971, 'Choice Functions and Revealed Preference', *Review of Economic Studies* **38**, 307–317; reprinted in A. Sen, 1982, chap. 1.

Sen, A.: 1973, 'Behaviour and the Concept of Preference', *Economica* **40**, 241–259; reprinted in A. Sen, 1982, chap. 2.

Sen, A.: 1982, *Choice, Welfare and Measurement*, Blackwell, Oxford.

Sen, A.: 1997, 'Maximization and the Act of Choice', *Econometrica* **65**, 745–779.

Simon, H. A.: 1955, 'A Behavioral Model of Rational Choice', *Quarterly Journal of Economics* **69**, 99–118.

Simon, H. A.: 1976, 'From Substantive to Procedural Rationality', in S. J. Latsis (ed.), *Method and Appraisal in Economics*, Cambridge University Press, Cambridge, pp. 129-148; reprinted in H. A. Simon, 1983.

Simon, H. A.: 1979, *Models of Thought*, Yale University Press, New Haven.

Simon, H. A.: 1983, *Models of Bounded Rationality*, MIT Press, Cambridge, MA.

Slote, M.: 1989, *Beyond Optimizing. A Study of Rational Choice*, Harvard University Press, Cambridge, MA.
Sugden, R.: 1985, 'Why be consistent?', *Economica* **52**, 167-184.
Tversky, A.: 1969, 'Intransitivity of Preferences', *Psychological Review* **76**, 31–48.
Tversky, A.: 1972a, 'Choice by Elimination', *Journal of Mathematical Psychology* **9**, 341–347.
Tversky, A.: 1972b, 'Elimination by Aspects. A Theory of Choice', *Psychological Review* **79**, 281–299.
Watkins, J. W. N.: 1970, 'Imperfect Rationality', in R. Borger and C. Cioffi (eds.), *Explanation in the Behavioural Sciences*, CambridgeUniversity Press, Cambridge.
Weber, M.: 1949, *Max Weber on the Methodology of the Social Sciences*, translated and edited by E. A. Shils and H. A. Finch, The Free Press of Glencoe, Ill.
Winter, S.: 1975, 'Optimization and Evolution in the Theory of the Firm', in R. H. Day and T. Groves (eds.), *Adaptive Economic Models*, New York, Academic Press, pp. 73–118.

THEMA, Centre National de la Recherche Scientifique &
Université de Cergy-Pontoise,
33 boulevard du Port, F-95000 Cergy
France
E-mail: mongin@u-cergy.fr