Synthese (2013) 190:2267–2289
DOI 10.1007/s11229-011-9971-z

The limits of unification for theory appraisal: a case of economics and psychology

Michiru Nagatsu

Received: 23 July 2010 / Accepted: 23 June 2011 / Published online: 3 July 2011

Abstract In this paper I examine Don Ross's application of unificationism as a methodological criterion of theory appraisal in economics and cognitive science. Against Ross's critique that explanations of the *preference reversal* phenomenon by the 'heuristics and biases' programme is ad hoc or 'Ptolemaic', I argue that the compatibility hypothesis, one of the explanations offerd by this programme, is theoretically and empirically well-motivated. A careful examination of this hypothesis suggests several strengths of a procedural approach to modelling cognitive processes underlying individual decision making, compared to a multiple-agent approach which Ross promotes. I argue that the debate between economists and psychologists are both theoretical and empirical, but cannot be resolved by appealing to the ideal of unification.

Keywords Unification · Ad hocness · Economics and psychology · Preference reversals · The compatibility hypothesis · Multiple-agent models · Procedural models

1 Introduction

Philosophical theories of explanation, causation, measurement and so on are often abstracted from historical case studies of past developments of mature science. But conceptual analyses of this sort are also expected to play some normative or regulative role in contemporary science (unless one is committed to the particular philosophical view that the rational reconstruction of past scientific theories has no implication for contemporary scientific practice). This study concerns one of the normative functions of the concept of *unification*, namely to provide us with some principled way

M. Nagatsu (✉)
Department of Philosophy, University of Tartu, Lossi 3, Tartu 51003, Estonia
e-mail: michiru.nagatsu@ut.ee

of appraising competing hypotheses when the relevant scientific communities cannot reach a consensus based on the available empirical evidence. A debate in contemporary economics supplies a good case: in the last forty years or so, *Expected Utility Theory* (EUT), one of the most prominent theories of how individuals make choices under risk and uncertainty, has been tested extensively both in the laboratory and in field experiments. Although the rigorous and systematic tests of EUT have accumulated persistent anomalous observations, there is little consensus among researchers, in particular between economists and psychologists, as to how to explain the data. Do the data refute the theory? And if so, how should we modify it? Or should we rather abandon it and come up with something different altogether? Or is there still room to argue that the data are an experimental artefact and that therefore the theory is untouched? The present study's main focus, Don Ross's monograph *Economic Theory and Cognitive Science: Microexplanation* (2005), proposes an answer that is original in at least two respects. First, Ross does *not* follow the 'explaining away' strategy common in economics, which (i) questions the external or 'ecological' validity of the experimental results and (ii) suggests that people's choice behaviour will conform to the standard EUT model once they make 'real' choices in economic contexts. On the contrary, Ross accepts the experimental results as a serious challenge to EUT. Second, however, he insists that "a separate economic science" (Ross 2005, p. 180) provides a better explanation than those proposed by the psychological 'heuristics and biases' programme and promoted by some behavioural economists. Ross's argument for a separated economic science crucially depends on accepting the methodological requirement that explanation be *unification*.[1] This makes Ross (2005) an interesting case, in which a normative role of the concept developed by historians and philosophers of science is tested in a contemporary scientific debate.[2] Through this case study, I will show that unificationism cannot settle this particular debate. The main reason is that Ross's rival, the 'heuristics and biases' programme (more specifically the compatibility hypothesis, which I shall examine in detail), is not 'Ptolemaic' (non-unificatory) in any sense. Although my case does not constitute a genuine counter-example to unificationism (in which some explanation is better yet less unifying), it conveys a sense of the limited role of unification in decision science, with an emphasis on some intricate theoretical and empirical aspects of the debate.

As will become clear later in the study, I am, roughly speaking, critical of EUT-based explanations and friendly to their rivals. The readers, however, should not interpret my theoretical preference as implying a naïve and unproductive dichotomy between economics and psychology; rather, I hope to suggest a more nuanced picture of two alternative models of cognitive processes underlying individual decision making: the multiple-self model and the procedural model.

The study proceeds as follows: first, I analyze the unificationist account advanced by Ross (2005), drawing on a Lakatosian framework (Sect. 2). As a background,

[1] I should note that unificationism is only one aspect of Ross (2005), which contains other stimulating methodological and substantial theses that are worth analyzing.

[2] Gintis (2009) also argues for unifying the social sciences with, among others, game theory. However Gintis's concept of unification emphasizes the *consistency* of different disciplines and thereby differs from the one that developed in the philosophy of science.

I will introduce the multiple-agent model, an economic model of inter-temporal choice, and show that the appeal to this model in explaining *intra*-temporal choice such as PRs, suggests that Ross (2005) is a Lakatosian unificationist. In the following four sections I will evaluate Ross's application of unificatioinism to behavioural decision research. First, I will assess, at the methodological level, Ross's charge that the 'heuristics and biases' programme is ad hoc, or 'Ptolemaic' (Sect. 3). I will then extend the assessment by looking in detail at a specific hypothesis that Ross criticizes, namely, the compatibility hypothesis (Sect. 4). Based on this assessment, I will argue that the compatibility hypothesis is not ad hoc in any sense; on the contrary, it points towards what I call a 'procedural model', an explanatory psychological model that, both on empirical and theoretical grounds, is at least as well-motivated as the multiple-agent model (Sect. 5). Finally, I will compare these two models, taking into account some recent neuroscientific evidence (Sect. 6). Section 7 summarizes the argument and concludes.

2 The unificationist approach to theory appraisal

Perhaps Kitcher (1981) is the best known account of scientific explanation as *unification*. Roughly put, Kitcher's unificationism is the thesis that scientific explanation should derive descriptions of more phenomena from fewer *patterns of argument*. Without going into the details of Kitcher's formal characterization of the concept of an argument pattern, one can intuitively grasp the gist of this account by example: if EUT allows us to derive the description of people's choices of mates as well as goods and services in terms of expected utility maximization, then it is more unificatory than a theory that allows us to derive only the description of the latter type of choices in terms of some utility maximization, while explaining the former type of choices using some other principle. Lakatos (1970) explicates a similar idea in terms of the concept of *progress*: a scientific hypothesis is progressive if it not only successfully predicts novel facts, but does so while maintaining certain 'core' theoretical features. Lakatos calls such a continuous explanatory enterprise a *research programme* (more on this below). What is common in these two accounts is the intuition that a hypothesis should not be ad hoc, but while Kitcher emphasizes the synchronic, logical aspects of unification, Lakatos looks at its diachronic, historical aspects. Although Lakatos himself does not use the term 'unificationism' to characterize his position, in the following discussion I will mainly draw on Lakatos's diachronic formulation of unificationism. I do this for two reasons: first, Lakatos, but not Kitcher, supplies a well-worked-out account of ad hocness. Second, such an account is necessary for the analysis of Ross (2005) position, which, although he calls it 'Kitcherian' (p. 176), heavily relies on the concept of ad hocness.

Preference theory states that a decision maker's preference ordering should not change over an identical set of options. Although we could say in general that, in violating this presupposition, the decision maker reversed her preferences, the term *preference reversal* (PR) is reserved for a narrower class of reversals in which the decision maker's preference over a pair of options is reversed, *depending on how we elicit her preference*. For example, if Anne says she would price a banana at £1 and

an apple at £2, and yet chooses the banana rather than the apple when both are free, we say that her choice behaviour manifests a preference reversal: the decision maker manifested inconsistent preferences, depending on the way in which they were elicited (pricing or choosing). In the past forty years of behavioural decision research, PR has established its status as an 'anomaly' in relation to EUT. An 'anomaly' in a research programme is, according to Lakatos, "a phenomenon which we regard as something to be explained in terms of the programme" (1970, p. 159, fn.1). In other words, being an anomaly is a relational status *vis-à-vis* particular research programmes. The status of an anomaly therefore may change as these programmes advance: generally, in relation to programme P_1, an anomaly turns into an 'example' when explained within the theoretical framework of P_1; it 'disappears' when independently explained by another programme, P^*; or it becomes a 'counterexample' when explained by P_1's rival programme, P_2 (ibid.). To say that PRs constitute a class of anomalies thus means that at least one research programme is involved in this process. In the present case, there are two programmes involved; one is the psychological 'heuristics and biases' approach advanced by the psychologists Kahneman, Tversky and Slovic, among others. The other is the economic approach characterized by its insistence that human decision making be modelled as utility maximization.[3] PRs are a class of anomalies in relation to the economic programme (P_1), while a rival programme, the psychological programme (P_2) purports to explain it. And yet there is no universal agreement among researchers that the psychological programme satisfactorily explains PRs, thereby leaving the phenomenon (in some sense) anomalous to both programmes. Instead of evaluating the competing hypotheses based on the available evidence, I shall focus on assessing whether methodological considerations regarding unification alone can say something in favour of the economic programme. I shall argue, *contra* Ross (2005), one of the main advocates of the unificationist approach in economics and cognitive science, that it cannot. First, I will describe the contrast Ross makes between 'Ptolemaic' (ad hoc) and 'non-Ptolemaic' (unificatory) science, and then, by illustrating Ross's favourite case, inter-temporal decision making, I will clarify how the debate concerning PRs can be interpreted according to this contrast.

2.1 'Ptolemaic' science and ad hocness

Ross (2005) uses the adjective 'Ptolemaic' to refer to a research programme which relies on an 'ad hoc' explanation in order to accommodate anomalies. But what exactly does 'ad hoc' mean? Lakatos (1970, p. 175, fns. 2, 3) distinguishes three senses of ad hocness: an explanation is ad hoc_1 if it does not predict any novel facts (no excess content); it is ad hoc_2 if it predicts novel facts but fails; it is ad hoc_3 if it predicts novel facts and is corroborated by evidence, but its progress is not led by a general outline of the programme regarding how to accommodate anomalies (*the positive heuristic*). It seems that Ross has in mind ad hoc_3 when he says some programme is 'Ptole-

[3] The distinction between the 'psychological' and the 'economic' is not sharp. In fact, Tversky and Kahneman's Prospect Theory is a model of utility maximization and so, in this specific sense, it is 'economic'. See Sect. 5 for a more nuanced distinction.

maic': a 'Ptolemaic' programme "must sooner or later reach a point of diminishing returns, where the effort required to further improve careful models can no longer be justified by gains in representational parsimony" (Ross 2005, p. 176).[4] To use a Lakatosian term, 'Ptolemaic' programmes are *degenerating* rather than *progressive*. Although Ross recognizes that 'Ptolemaic' phases in science are unavoidable as a process of systematically summarizing data[5] and can be instrumental in future theoretical progress, Ross, just like Lakatos, further requires that a promising programme should have some principled way of unifying existing data. In Ross (2005, p. 175) own words, 'non-Ptolemaic' science ought to be motivated by "wider theoretical considerations" independent of the data it seeks to parsimoniously summarize.

In explaining PRs, however, what constitutes such a unifying principle is not quite clear. As it turns, out, as a 'non-Ptolemaic' strategy Ross has in mind what Lakatos calls a 'creative shift' (Lakatos 1970, p. 137) in the positive heuristic of a research programme. In Ross's construal, the *negative* heuristic of the economic programme—which defines the irrefutable 'hardcore'—requires that an agent's behaviour be modelled as *maximization of utility* defined as indices of consistent and stable preference orderings (preference theory);[6] in addition, the *positive* heuristic tells researchers how to accommodate evidence within the framework of the programme. In the economic programme, the content of preferences is completely unspecified, thereby enabling the programme to be applicable to a large set of behavioural patterns.[7] However, PRs are regarded as a serious anomaly to the economic programme because the phenomenon apparently challenges one of its hardcore assumptions that preferences be consistent in the sense that they conform to the axioms of EUT or its variants. Now, Ross's creative shift is to hypothesize that not only the *content* of preferences but also the *agents*, who act upon preferences, are unspecified. More specifically, he proposes an auxiliary hypothesis that the economic agents characterized with their utility maximizing behaviour are *not* individual human beings but *parts* of individuals. In other words, individuals can consist of more than one economic agent. With this shift, an individual's behaviour exhibiting PRs may be interpreted as resulting from the combination of more than one preference ordering per agent. In this way, the hard core of the preference-based programme may be saved from the refutation and the anomaly turned into an example manifesting the fruitfulness of the programme.

In what follows, I will explain several models of *inter-temporal* choice, including one that motivates Ross (2005) to advocate the multiple-self model in the domain of *intra-temporal* choice, in which PRs emerges as an anomaly.

[4] If non-parsimonious theories are inferior to parsimonious ones in predictive power, then Ross's 'Ptolemaic' programme is ad hoc_2 as well as ad hoc_3.

[5] Lakatos makes the same point by noting that ad hoc_3 explanations are at least empirically corroborated.

[6] Ross identifies this hardcore with Revealed Preference Theory (RPT) advocated by Paul Samuelson. While many economists would agree with this, the interpretation of RPT itself is contested.

[7] There is a common but essentially unfounded worry that the 'thin' interpretation of preferences makes the economic programme tautological or empirically vacuous. For a good discussion see Guala (2006), and his footnotes 22, 39 and 40 for the relevant literature.

2.2 A case of inter-temporal choice: the departure from the standard model of discounted utility

A contemporary version of the idea that an individual consists of more than one 'self' has been developed by the American psychiatrist George Ainslie since the 1970s in the context of inter-temporal choice. In the following, I will first introduce its rival, the received model, whose empirical inadequacy led Ainslie and others to propose alternative models of inter-temporal choice.

The standard model, called the *discounted utility model* (DUM), was originally formulated by Paul Samuelson in 1937.[8] DUM represents people's choices between consumption bundles across different times by flattening all the relevant psychological factors into the single parameter of a *discount rate*. Mathematically, DUM represents the utility at the time t of the consumption profile $(c_t,\ c_{t+1},\ c_{t+2}, \cdots, c_T)$, starting in period t and continuing until period T, as an *inter-temporal utility function* defined as follows:

$$U^t\ (c_t, \ldots, c_T) = \sum [k = 0 \rightarrow T - t] D(k)\ u(c_{t+k})$$

where $D(k) = (1/1 + \rho)^k$

The function $u(c_{t+k})$ can be interpreted as an individual's *instantaneous utility function*, representing her perceived wellbeing in period $t + k$. The other function appearing on the right hand side of the equation, $D(k)$, is her *discount function*, representing the relative weight she attaches in time period t to her wellbeing in period $t + k$. The parameter in this function, ρ, refers to her *discount rate*, representing the rate at which the individual discounts future utilities. For example, Anne's utility from receiving her annual salary of £32,000 for three continuous years may be calculated as follows, if the utility from each year's income is always x and if the discount rate is 10%:

$$\begin{aligned} &U^t(£32{,}000, £32{,}000, £32{,}000) \\ &\quad = x/(1 + 0.1) + x/(1 + 0.1)^2 + x/(1 + 0.1)^3 \fallingdotseq 2.45x \end{aligned}$$

In the present context, it is essential to notice three features of DUM. First, DUM assumes that people use the same discount rate, ρ, over their lifespan (*stationary discounting*). This means, for example, that Anne discounts her £32,000 by 10% at any year t, regardless of whether she is a teenager, middle-aged, or retired. Second, DUM assumes that at any period of time the same discounting by the exponential k is applied to all future periods (*constant discounting*). Third, these two assumptions of stationary and constant discounting ensure that people's preferences do not change over time (*time-consistent preferences*).

Not only is DUM introspectively unrealistic and counter-intuitive, but also it has been shown to be inadequate as a model of actual people's choice behaviour both in

[8] The summary in this section is based on Wilkinson (2008, Chaps. 5 and 6). Wilkinson emphasizes that Samuelson believed neither in the empirical nor in the normative validity of this model.

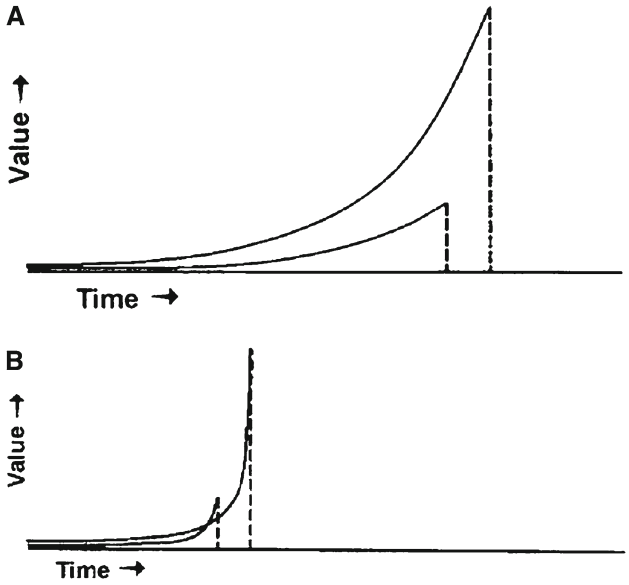

Fig. 1 Exponential discount curves from a smaller-sooner (SS) and a larger-later (LL) reward (**a**) contrasted with hyperbolic discount curves from an SS and an LL reward (**b**) (from Ainslie 2005, p. 636)

the field and in laboratory experiments. For example, ρ is known to decrease as people enter middle age, but increase again as they get older (Read and Read 2004; Harrison et al. 2002). The unavoidable implication of this and other observations is that people's preferences are not consistent over time. This, however, does not mean that modelling inter-temporal choice is impossible; it is still possible to represent time-inconsistent choices by using some other discount function. In fact, various such models have been proposed. The most famous one is called the *hyperbolic* discount function, whose development Ainslie (e.g., 1991) contributed to. Let u_t be the instantaneous utility an agent gets at time t. In a discrete-time form, a hyperbolic, or 'quasi-hyperbolic' discount function is then represented as follows:[9]

$$U^t\ (u_t,\ u_{t+1}, \ldots, u_T) = (\delta)^t u_t + \beta \sum [\tau = t + 1 \rightarrow T](\delta)^\tau u_\tau$$

where δ and β are parameters less than 1, with δ very close to 1. If $\beta = 1$, the hyperbolic utility function reduces to the exponential function. Figure 1 graphically contrasts these two types of discounting.

In Fig. 1b, but not in a, the smaller reward is temporarily preferred just before it becomes available, which is shown by the curve of the smaller reward crossing that of the larger one from below. Although hyperbolic discount functions may be seen as a mere technical adjustment to the exponential function, adopting such models inevitably raises a difficult question of *consciousness* or self-awareness about inconsistencies

[9] I gloss over the mathematical details and use the terms 'hyperbolic' and 'quasi-hyperbolic' interchangeably.

on the part of individual actors. In order to understand this point, imagine the following scenario: Anne purchased two packs of her favourite giant chocolate toffee cookies from Tesco because they were on a 'buy-one-get-one-free' discount. But she knows that eating too much of them can cause health problems that she would like to avoid. So at the time of purchase ($t = 1$) she decided to eat one pack per week, her usual quota. Now the first chance to eat the cookies ($t = 2$) arrives. Quickly finishing one pack, Anne glances at the second one on the shelf, reaches down and eats it all. One of her housemates, Alex, who happens to be a devoted follower of DUM, enters the room and says, "Your preferences are inconsistent, Anne! At time 1 you preferred eating one pack per week to two per week, because of your concern for health. But now, at time 2, you prefer to eat two! If I remember correctly, you bought two packs to get a discount, not to eat them all at once." What would Anne's reply be? I predict that most people would feel uneasy with the following answer: "Sure, Alex, my preferences are indeed time-inconsistent. But what's wrong with them? My discount function is *hyperbolic*, not *exponential* like yours!" The oddity of this answer comes from the fact that hyperbolic functions may describe your inconsistent behaviour but cannot capture the psychological fact that you are aware of your preference at $t = 1$ and that you still see that preference as reasonable at $t = 2$. People are in fact usually aware of their past preferences; in this imaginary case Anne would also see the reasonableness of her past preference. One might object that the individual's self-awareness is irrelevant for the empirical issue of how the observer best models and explains behaviour, but the fact that people often talk about their choices' inter-temporal consistency in this kind of situation reminds us of an important empirical fact that people are capable of foreseeing changes in their preferences and acting accordingly. For example, in an alternative scenario, Anne may, as a means of self-command, choose to buy only one pack, forgoing the chance of discount. Strotz (1955) labels a decision maker who accurately anticipates the future change in her preferences as a 'sophisticated', as opposed to a 'naïve' hyperbolic decision maker, who erroneously believes that her future preferences will be identical to her current ones. It seems that most people lie somewhere in between these two extreme cases. Wherever exactly people are located, it seems necessary to rethink the rationale of the hyperbolic discount model, once such 'internal conflict' is recognized as underlying mechanisms of time-inconsistent choices. Wilkinson (2008, p. 236) thus suggests that the hyperbolic discounting approach lacks a psychological foundation.[10]

In order to explain why people's choice behaviour conforms to hyperbolic discount utility functions, Ainslie and others have developed so called the 'multiple-self model'. In this model, 'multiple-self' refers to the existence within each person of several agents defined by their own interests, or preference orderings. Typically, the short-term 'self' and the long-term 'self' are defined by their distinct preferences, e.g., indulging

[10] This is not necessarily the case, however, once one recognizes the possibility of interpreting the hyperbolic discounting *realistically* (I thank Don Ross for suggesting this). Specifically, the hyperbolic discount model may be interpreted realistically as representing distinctive mechanisms underlying individual inter-temporal decision making, rather than a mere 'curve-fitting' adjustment to the exponential model. Such a literal interpretation of a utility function is unusual not only for the exponential model, but for the economists' general modelling strategy, which has minimal ontological commitment to the functional representation of utilities.

yourself with your favourite sweets vs. maintaining good health, respectively. There are at least three reasons that lend some support to this idea. First, conceptually this is a simple way to make sense of the common observation that we encounter self-control problems: if self-control is a real phenomenon, then there must be at least two 'selves', one controlling and the other being controlled. Singular utility models cannot even make sense of there being an issue of self-control (Wilkinson 2008, p. 232). Second, although contested, a functional magnetic resonance imaging (fMRI) study of brain activities of decision makers suggests that choices between two delayed rewards and choices between immediate and delayed rewards are associated with the activity of distinct brain areas, namely, the lateral pre-frontal cortex and the limbic system, respectively (McClure et al. 2004). This may be interpreted as suggesting that models of conflicting 'selves' are not mere metaphors but may have some physical correspondence at the neuro-physiological level. Finally, some models of multiple-selves provide accurate predictions of time-inconsistent choices. Specifically, Ainslie (2001) models interactions of short- and long-term interests as repeated Prisoner's Dilemma games, accurately predicting a set of various addictive behaviours. Fudenberg and Levine (2006) also apply a dual-self model to predict not failure (such as addiction) but success of self-control, i.e., non-pathological behaviour such as people's strategic limiting of pocket cash to prevent overspending later at a nightclub where their preferences may change under the influence of alcohol and drugs. Ross (2005, p. 341) takes Ainslie's results as "the principal source of [Ainslie's model's] empirical persuasiveness"; further Ross suggests that a wider range of game models (assurance, coordination, inspection games etc.) should be able to explain a wider range of behavioural patterns resulting from interactions of different selves within the "sub-personal marketplace". Here, it is evident that Ross's methodological justification of the 'non-Ptolemaic' model comes not only from its empirical success—which even ad hoc_3 models may achieve—but also from its potential unifying power: with the creative shift of seeing a whole individual as a community of distinct economic agents, it becomes possible to unify models of (both inter- and intra-temporal) individual decision making using preference theory and game theory.

3 The assessment of the 'ptolemaic' critique as a methodological thesis

Does the success of the multiple-self model in the domain of inter-temporal choice lend some support to adopting this model in the domain of intra-temporal decision making? Ross (2005) seems to suggest that it does. In the following I will argue that such unificationist considerations are not conclusive. This section concerns methodological aspects of unificationism, while Sects. 4 and 5 deal with empirical problems with Ross's (2005) unificationist approach.

Ross (2005) characterizes various attempts to model risky decision making as *individuals*' utility maximization (EUT, Prospect Theory, Regret Theory, etc.) as 'Ptolemaic', or ad hoc_3 in our Lakatosian terminology. That is, Ross regards these models as ultimately unable to explain all the relevant phenomena (including PRs) as long as they take individuals as loci of maximization. This position is manifest when Ross, in discussing the preference reversal phenomenon, criticizes *the compatibility*

hypothesis proposed by psychologists and some behavioural economists. One version of this hypothesis states that people, when evaluating options, attach greater weight to information (or input stimuli) that are more compatible with output selection tasks (see Sect. 4 for a more detailed analysis). Ross's target is Tversky and Thaler (1990), who employ this hypothesis to explain PRs. The main pattern of PRs is such that people *price* low-probability high-payoff bets (L bets) higher than high-probability low-payoff bets (H bets), while *choosing* the H bets rather than the L bets, making the apparent 'reversals' of preferences in the two kinds of tasks. This pattern can be explained in terms of compatibility bias because the information regarding price seems to be weighted more in the pricing task, while in the choosing task there is no such bias. Ross criticizes this interpretation for two reasons, one empirical and the other theoretical. The empirical reason is simply that other experimental data (Loomes et al. 1991; Loomes and Taylor 1992) suggest that reversals occur regardless of compatibility biases.[11] Here I focus on the theoretical reasons, since it is these reasons that make Ross's criticism unique. Ross (2005) disregards the compatibility hypothesis because (he thinks) it depends on a classical computational model of the mind in assuming that there are "facts of the matter about whether and how data are matched, as a distinct processing step during computation" (p. 179) of information in the brain of the decision maker. Second, Ross argues that the compatibility hypothesis is unsatisfactory because it is a non-economic explanation in the sense that it refers to evolutionarily formed heuristics that minimize cognitive effort. I will consider the second reason below, and postpone the discussion of the first until Sect. 6.

It might appear that Ross's distinction between economic and evolutionary explanations is not well-grounded in the first place because the explanation based on cognitive effort-minimizing heuristics seems to use the same argument pattern or 'hard core' (optimization) as the economic explanation based on utility maximization.[12] However, there is an important difference between the two types of explanation: the maximization principle in economic explanations is just a mathematical way of representing choice *consistencies*, whilst what is being consistently preferred, or what is being maximized (e.g. profit, pleasure, and the like) is left open to be specified by different auxiliary hypotheses; by contrast, the heuristic explanation is a substantial hypothesis as to what is being maximized, namely, the probability of individuals' survival and reproduction. Margolis (2007) makes explicit this point in discussing 'neglect defaults', another type of evolved heuristics that he proposes. He suggests that what defines such heuristics "is not the economy of using them on particular occasions (which is usually slight), but that the occasions for the default responses are so very common. Without neglecting almost all such occasions by default, a person would be overwhelmed by hesitations" (pp. 88–89), thereby lowering their chance of survival. Since it is normally assumed that in economic choice contexts people are trying to

[11] Although I do not discuss the empirical reason further in this study, I should note that many experimental economists do not regard the compatibility hypothesis as the clearly superior account for empirical reasons. In particular, resurgent interest in stochastic models for binary discrete choice under risk (see e.g., Wilkinson 2008, and his references on p. 200) seems to encourage this scepticism among experimental economists.

[12] I owe this point to Julian Reiss.

maximize money, one needs to specify how these two auxiliary hypotheses (money maximinzing and survival maximizing) are composed to constitute the ultimate drive for choice behaviour. Ross can be interpreted as worrying (understandably) that such an explanation necessarily involves ad hoc assumptions regarding how two causal factors are composed to yield the final effect.

But, is this worry justified? To be sure, from the *economic* unificationist perspective the use of the evolutionary argument in explaining individual decision making is a compromise to "a separate economic science". However, even if one accepts unification as an important ideal, one does not have to accept that the unification should be achieved by the particular economic programme Ross promotes. In fact, I will later suggest that there is a strong rival programme (see Sect. 5). Even if (and this is a big if) Ross could show that his economic programme is more unifying than others under some formal reconstruction of the current theories and evidence, this will not be decisive since such a formulation is sensitive to new empirical evidence and theoretical developments. In sum, Ross's methodological worry about the ad hocness of the compatibility hypothesis (as he understands it) is subjectively understandable as the worry of a defender of a separate economic science, but it is not justifiable.

Moreover, it turns out that the compatibility hypothesis is much more complex and subtle than Ross assumes. Ross interprets the compatibility hypothesis as specifically concerning a mechanism of mental computational processing. This is understandable because Ross refers to Tversky and Thaler (1990), a review article which was written for economists at the time that the compatibility hypothesis was further elaborated by cognitive psychologists. The elaboration includes a discovery of two distinct causal factors involved, and the following reformulation of the two hypotheses, *the strategy compatibility hypothesis*, and *the scale compatibility hypothesis*. I will give a detailed account of this elaboration in the next section, in order to show that Ross (2005) criticism of the hypothesis is not only inaccurate but also neglects its rival programme, which I will discuss in Sect. 5.

4 Decomposing the compatibility hypothesis

In this section, I will review the literature on the compatibility hypothesis in some detail, in order to show that Ross's criticism is largely misguided. In particular, the compatibility hypothesis is decomposed into two distinct models, namely, *the scale compatibility hypothesis* and *the strategy compatibility hypothesis*.

4.1 History

In one of the first studies of PRs, Lichtenstein and Slovic (1973/2006) proposed the hypothesis that "the compatibility or commensurability between a cue dimension and the required response will affect the importance of the cue in determining the response." (pp. 75–76) The idea is based on the input-output model of human perception, information processing and cognition. According to this model, when a subject receives some stimuli, these stimuli (input) are processed within the subject's brain in order to produce appropriate responses (output). Schematically put, this approach thus

focuses on the mechanisms of either the selection of stimuli or the process of stimuli, or both.[13] We will shortly see that Slovic et al.'s studies concern both phases. 'Compatibility' or 'commensurability' is meant to capture the comparability, or similarity, *for the subject* between the initial stimuli and the information used to make decisions, although input (stimuli) and output (responses) are not necessarily comparable in the objective sense (hence 'compatibility'). In other words, compatibility concerns the subjective perception of a relationship between options and tasks.

In the 1970s, parallel to his early work on PRs, Slovic was engaged in a separate line of research investigating the difference between choice and matching responses through the use of two-dimensional stimuli, such as batting averages vs. number of home runs (where the task is to choose between baseball players), speed vs. accuracy (where the task is to choose between typists), and so on. In four experiments, Slovic (1975) asked his subjects first to match different pairs of options (making each pair equal in subjective value), and then to choose between the matched options. He found that the subjects did *not* choose randomly (as was predicted by the equality in subjective value) but tended to select the option that was superior on the more important dimension (in the abovementioned example batting average and typing accuracy, respectively). Slovic's (1975) judgement that a particular dimension is 'more important' than the other is not based solely on the observation of which dimension weighted more heavily in responses, which would make the statement that 'people tend to select the option that is superior on the more important dimension' somewhat tautological. Rather, he hypothesized in advance which dimension would be more important, based on the estimation of how easily one could apply and justify the response; he also confirmed this hypothesis by interviewing the subjects *ex post*, asking for an explanation of their choices.

Later, Tversky saw in this finding "the seeds of a general theory of response-mode effects that had the potential to explain a wide variety of empirical findings, including preference reversals" (Slovic 1995, p. 497), and jointly elaborated and tested this theory in Tversky et al. (1988) and in Slovic et al. (1990). In the former, the authors generalized Slovic's (1995) hypothesis as the *prominence hypothesis*,[14] which states that the more prominent (important) attribute will loom larger in choice than in matching. They further suggested that the prominence hypothesis might be interpreted as an instance of a more general hypothesis, the *principle of compatibility*,[15] which states that the weight of a stimulus (input) attribute is enhanced by its compatibility with the response (output) mode. The rationale of the latter is that the prominence hypothesis indicates that qualitative considerations loom larger in the ordinal procedure of choice than in the cardinal procedure of matching, which may be explained by the principle of compatibility.

[13] The final stage, namely, the production of responses does not seem to be an explicit subject matter of this model. Goldstein and Einhorn (1987) *expression theory* explicitly models the production phase.

[14] Originally, the hypothesis was somewhat clumsily labelled the "more important dimension hypothesis" (Slovic 1975, p. 281).

[15] Decision research psychologists tend to use 'hypothesis' and 'principle' (or 'explanatory principle') interchangeably. See e.g., Tversky et al. (1990).

The principle of compatibility as such, however, tells us little unless substantiated by auxiliary hypotheses regarding what is (and what isn't) compatible with what, so here some homely example may be helpful:[16] a kitchen stove usually has a square array of four burners, with the knobs being linearly arrayed in front of the stove. People often make mistakes regarding which knob to use in order to control, say, the burner in the upper right corner of the square array. This type of mistake can be reduced if the knobs are also squarely arrayed so that each knob is visually matched to the corresponding burner (the upper right knob for the upper right stove, and so on). In the latter arrangement, input stimuli (visual perceptions of the burners) and output responses (controlling of the knobs) are compatible, while in the former they are not, which explains the difference in performance in the two arrangements. This is an example of compatibility in display. Extending from this simple and unproblematic example, one can talk about compatibility in scale, semantic correspondence, and so on. In each case, the hypothesized mechanism is twofold: at the first stage of the input-output scheme, the compatibility between the stimulus attribute and the response mode increases the availability of stimuli by priming or focusing attention on the compatible features of the stimulus; then in the second phase, the compatibility increases (or the lack thereof decreases) the computational ease of processing the stimuli, both phases resulting in the enhanced weight of the compatible stimuli in responses.

In order to elaborate on the compatibility hypothesis, Slovic et al. (1990) designed five experiments, two of which concerned the role of compatibility in prediction (of market value [study 1] and academic performance [study 2]), and three concerned compatibility in preference (over monetary vs. nonmonetary outcomes [study 3], immediate vs. delayed payoffs [study 4] and by matching vs. pricing [study 5]). In particular they were interested in ways in which the compatibility effect creates PRs in the latter set of experiments. The study was partly motivated by their discovery of an asymmetry in the PR data: PR was due mainly (more than 65% of all PRs. See Tversky et al. 1990, p. 210) to the overpricing of low probability, high payoff bets, the so-called $-bets or L-bets. From the compatibility principle, they inferred that the payoffs (which are expressed in dollars) would be weighted more heavily in pricing (which is expressed also in dollars) than in choice. Since the payoffs are much larger in the L bets than in the H bets (high probability, low payoff bets), the compatibility effect seemed to explain the overpricing of L bets, and thus the main cause of PRs. Although the results of studies 3 and 4 of Slovic et al. (1990) were encouraging for this interpretation, they encountered a surprise in study 5, in which they used a *matching* procedure to elicit preferences: that is, they first required the subjects to fill in a missing value so as to equate a pair of options, and then inferred their preference from the value they used. This design made it possible to compare percentages of particular preferences across four different elicitation procedures as shown in Table 1 for $H \succ L$.

In Table 1, the comparison between the results of choice and pricing shows the familiar PR pattern: the subjects choose L bets but priced H bets higher (76 vs. 37% by overall mean). In addition, the comparison between probability matching and payoff

[16] This example is from Slovic et al. (1990, pp. 217–218). The authors are aware of the lack of an independent procedure for assessing the compatibility between stimulus elements and response modes (pp. 218; 228), and resort to an unproblematic example such as the one I'm using here.

Table 1 Percentage of responses favouring the H bet over the L bet for four different elicitation procedures (from Slovic et al. 1990, p. 225)

(Task)/(Bets)	Choice	Probability matching	Payoff matching	Pricing
Small bets (H, L)				
(35/36, $4), (11/36, $16)	80	79	54	29
(29/36, $2), (7/36, $9)	75	62	44	26
(34/36, $3), (18/36, $6.50)	73	76	70	39
(32/36, $4), (4/36, $40)	69	70	26	42
(34/36, $2.50), (14/36, $8.50)	71	80	43	22
(33/36, $2), (18/36, $5)	56	66	69	18
Mean	71	72	50	29
Large bets (H, L)				
(35/36, $100), (11/36, $400)	88	76	69	65
(29/36, $50), (7/36, $225)	83	64	31	55
(34/36, $75), (18/36, $160)	77	79	65	55
(32/36, $100), (4/36, $1,000)	84	68	28	61
(34/36, $65), (14/36, $210)	78	80	36	57
(33/36, $50), (18/36, $125)	68	75	58	46
Mean	80	74	48	56
Overall mean	76	73	49	37

Note that apart from the direct choice task, these percentages are inferred from the probability and payoff matches and stated prices

matching reveals what seems to be the result of the compatibility effect: probability matching favours the H bets, whereas payoff matching favours the L bets (73 vs. 49%). But what surprised the experimenters most was the comparison between choice and matching. They reasoned that, if the compatibility effect was the sole cause, the probability matching would bias the responses in favour of the H bets and payoff matching would bias the responses in favour of the L bets, relative to choice. For the choice procedure was neutral with respect to the compatibility effect. They thus predicted that the percentage of responses favouring the H bets would be ordered as:

$$\%(\text{probability matching}) > \%(\text{choice}) > \%(\text{payoff matching}).$$

In fact, however, they observed:

$$76\%(\text{choice}) \fallingdotseq 73\%(\text{probability matching}) > 49\%(\text{payoff matching}).$$

Slovic et al. (1990) explained this with the prominence hypothesis, according to which the more prominent (important) attribute will loom larger in choice than in matching. Tversky et al. (1988), who extensively investigated the prominence hypothesis, interpreted PRs as caused by the compatibility effect rather than the prominence effect, because they 'saw no a priori reason to hypothesize that probability is more important

than money' (Slovic et al. 1990, p. 226). But given the new result that the subjects favoured the H bets in choice as much as (or sometimes even more than) in probability matching, they reconsidered the possibility that PRs may be caused by the prominence effect rather than (or in addition to) the compatibility effect. Their interpretation that probability is the prominent dimension in risky choice is also supported by the finding that the rating of bets is dominated by probability (see Slovic and Lichtenstein 1968; Goldstein and Einhorn 1987). From this perspective, the result in Table 1 can be understood as the combination of two effects: "a compatibility effect that is responsible for the difference between probability matching and payoff matching (including pricing), and a prominence effect that contributes to the relative attractiveness of H bets in choice" (Slovic et al. 1990, p. 226).

4.2 Separating causes

The interpretation by Slovic et al. (1990) might appear a little confusing, since Tversky et al. (1988) suggested that the prominence hypothesis "may be constructed as an example of a more general principle of compatibility." (p. 513) But how can a special case (the prominence effect) occur in choice tasks separately and independently from its general manifestation (the compatibility effect)? Fischer and Hawkins (1993) clarify this confusion by explicitly distinguishing *scale compatibility* from *strategy compatibility*. The former says that the "weight of any input component is enhanced by its compatibility with the output" (Tversky et al. 1988, p. 513), the mechanism behind this being that scale compatibility makes particular stimuli more accessible and reduces the burden of computation. The latter states that "[q]ualitative preference tasks are more likely than quantitative tasks to evoke a preference for the alternative that is superior with respect to the most important attribute" (Fischer and Hawkins 1993, p. 583). This is presumably caused by the compatibility between the nature[17] of the response task and the nature of the decision strategy it invokes, *not* by the compatibility between the units of payoff dimension of the stimuli and the units of the response scale. The argument of the strategy compatibility hypothesis is as follows (ibid.):

(St. 1) Quantitative response tasks evoke quantitative strategies in which the decision maker makes trade-offs between value attributes.

(St. 2) Qualitative response tasks evoke multi-stage decision processes that involve a mix of quantitative and qualitative strategies, with the latter being used to resolve close decisions.

(St. 3) Qualitative strategies give primary consideration to differences with respect to the prominent attribute.[18]

[17] Fischer and Hawkins use the word 'metaproperty' instead of 'nature' presumably in order to indicate that the two properties in question are not obvious to the experimenter: regarding the property of a response task, what distinguishes quantitative from qualitative response tasks is unknown prior to empirical investigation (in Experiment 2 they address this question; see p. 587); regarding the property of a decision strategy, it is even less obvious which task evokes which strategy.

[18] Fischer and Hawkins note that (St. 3) holds only if we assume the use of particular qualitative strategies, such as lexicographic ordering or elimination by aspects. For example, if a qualitative strategy involved is a

(St. 4) Qualitative tasks are more likely than quantitative tasks to lead to a preference for the alternative that is superior with respect to the prominent attribute (*strategy compatibility hypothesis*).

Thus stated, the strategy compatibility hypothesis turns out to be a generalized version of the prominence hypothesis; it generalizes choice-matching comparison to include any comparison of a qualitative and quantitative preference task. Compare this with the argument of the scale compatibility hypothesis:

(Sc. 1) A response mode primes or focuses attention on the compatible features of the stimulus.
(Sc. 2) Compatibility (non compatibility) between the input and the output scale requires less (more) mental operations, often decreasing (increasing) effort and error.
(Sc. 3) The weight of a stimulus attribute is enhanced by its compatibility with the response mode (*scale compatibility hypothesis*).

This input-output compatibility can be applied not only to scale (e.g., dollar $\Rightarrow$ pricing in dollars), but also to other dimensions, that is, the notion of compatibility can be extended to the nature of the information and the nature of the task (e.g., ordinal info $\Rightarrow$ ordering, common features $\Rightarrow$ similarity judgement) in general. Now it should be clear, however, that strategy compatibility is not generalizable to this hypothesis, since the compatibility of the former concerns different things, i.e., response tasks and decision strategies. The two are thus distinct hypotheses. It is therefore conceivable that they imply opposite predictions. For instance, in choice and matching tasks involving jobs with two dimensions (salary and vacation time), the strategy compatibility hypothesis predicts that the prominent dimension (i.e., salary in this case) looms larger in choice than in matching tasks. On the contrary, the scale compatibility hypothesis predicts that salary will be weighted more heavily in dollar-matching tasks than in choice. Based on this insight Fischer and Hawkins (1993) designed four experiments to detect the strategy compatibility effect and the scale compatibility effect separately (in riskless choice), and observed that the strategy compatibility effect is much larger than the scale compatibility effect.

This is not the end of the story: some anomalies persist, as is often the case in experimental science. For example, Fischer and Hawkins (1993) note a further puzzle, namely that in a study of risky choice by Goldstein and Einhorn (1987), attractiveness rating tasks evoked stronger preference for H (high-probability, low-payoff) bets than choice did, a phenomenon that neither the scale-compatibility hypothesis nor the strategy-compatibility hypothesis (nor a combination of the two) predict. But the point should be clear by now: the compatibility hypothesis is a causal hypothesis, and two distinct causal mechanisms have been identified.

conjunctive rule (in which one eliminates any option that falls below one's aspiration level on any attribute), (St. 3) does not necessarily hold.

5 The procedural approach

The discussion in the previous section makes it clear that the strategy compatibility hypothesis, distinguished from the scale compatibility hypothesis, does *not* presuppose any specific computational model of the human mind. While scale compatibility biases presumably take place in order to minimize computational cost (Sc. 2), strategy compatibility biases are neutral with regard to the presumption of such a mechanism. Although the first premise of the strategic compatibility hypothesis (St. 1) says that quantitative response tasks evoke quantitative strategies, the use of quantitative strategies is not necessarily minimizing computational costs; in fact, quantitative strategies (e.g., maximizing profit) can sometimes require more mental efforts than qualitative ones. The fact that quantitative tasks evoke quantitative strategies may be explained by some other causes such as *framing*. Regarding this point, there is a highly suggestive study by Rubinstein (2006), who conducted a set of questionnaire-based experiments to investigate the effects of economic education on people's decision making. In the experiments the subjects were asked to make a decision, as a CEO of a company, on how many employees they were willing to maintain in order to increase profit. The subjects were presented a table of seven combinations of numbers of workers who would continue to be employed and the resulting profits of the company, and asked to decide upon the number of employees they were willing to continue to employ. The table was constructed in such a way that there was a trade-off between employee protection (reducing a number of layoffs) and profit maximization, as is typically the case in recession phases. Rubinstein found that the group of economics students tended to prioritize profit maximization (45–49% choosing the profit-maximizing number, 100), while other groups sacrificed profit maximization to a varying extent in order to reduce the number of employees who would be fired (only 13–16% of philosophy and mathematics students choosing the profit-maximizing 100). Interestingly, this variation among different groups *disappeared* once the table showing various results was replaced by a formula (profit function) which yields similar values to those presented in the table: in this condition, a similar proportion of subjects (73–77%) regardless of their educational backgrounds chose the profit maximizing value, 100. This result can be interpreted as an example of strategy compatibility biases, where, although the response task is identical (i.e., deciding how many workers to keep), different ways of framing the similar information (table vs. formula) induce different response strategies (choosing from multiple alternatives vs. solving an equation), resulting in different decisions. It seems that the prominent attribute for economics students in this choice is profit, while for non-economics students employee protection also matters. But the majority of non-economics students, who balanced between conflicting goals (employee protection vs. profit-maximization) in the 'table' condition, seemed to have paid less attention to this conflict in the 'formula' condition. Instead, they simply solved the equation to yield the profit-maximizing number of employees. Computational economy does not explain this shift of response strategy because solving the equation is not computationally easier than choosing a value from the table. In the 'table' condition, the subjects only had to identify the maximum profit on the table, and then see the corresponding number of employees. In the 'formula' condition, by contrast, subjects had to compute the profit-maximizing number of employees (x)

based on the function '$2\sqrt{x} - 0.1x - 8$'. Note that in both conditions it was explicitly stated that profits would still be positive even if no workers were laid off.

Nor can the shift be explained by the differences in training, since there was no significant variation among different groups of students. Although an economics education seems to influence what dimension one sees as prominent in choice options, this effect ceases to be significant once the quantitative strategy (i.e., solving the equation) is primed by manipulating the presentation of the relevant information. These findings suggest that strategy compatibility effects can be quite powerful, regardless of considerations of computational costs.

A second thing to note is that the scale compatibility and the strategy compatibility hypotheses are not presumed to be mutually exclusive rival hypotheses, but are expected to capture two different causally relevant mechanisms underlying preference reversals. As in this case, psychological models typically presuppose that a particular phenomenon results from compound causes of heterogeneous natures. This fact refutes Ross's allegation that the compatibility hypothesis is based on the old-fashioned computational model of human minds. On the contrary, it may be argued that psychological models are relatively 'liberal' in allowing for diverse theoretical presuppositions, making a sharp contrast with the standard economic models committed to some form of utility maximization. Criticizing psychological models because of their presumed rigid theoretical commitments is not well grounded.

Contrary to Ross's suggestion, the strategy compatibility hypothesis in fact reflects two different traditions in research on decision making, namely, the reason-based and value-based approaches (see Shafir et al. 1993). Value-based models (such as EUT and Prospect Theory) model individual choice as a maximization of utility which an individual assigns to different objects of choice based on her intrinsic preference. By contrast, reason-based models model individual choice as a result of certain inferential processes. Choosing an option that is more valued with respect to a prominent attribute (e.g., salary as opposed to the number of holidays, when choosing a job) from two equally valued options is an example of a reason-based decision. In the strategy compatibility hypothesis, quantitative strategies correspond to value-based models, while qualitative strategies are better captured by reason-based models. These two approaches are not mutually exclusive, but are meant to capture two different types of mechanism both involved in decision-making processes.

Should such a 'liberal' approach be condemned because it lacks a unifying theoretical framework? Not necessarily: first, the strategy compatibility hypothesis is not a mere conjunction of two types of decision process model, but rather it concerns perceptual or cognitive mechanisms through which different processes (i.e., reason-based and value-based decision strategies) are evoked depending on the nature of the task. In other words, the hypothesis not only identifies two types of mechanism, but also purports to identify the conditions under which these mechanisms are triggered. Second, although the value-based approach is theoretically much more sophisticated, some theorists have started to provide a formal and unifying framework for the reason-based approach. Gold and List (2004) have recently proposed a formal framework that unifies the compatibility effects and framing effects. According to this framework, the violations of both types of invariance—*procedure invariance* associated with compatibility effects and *description invariance* associated with framing effects—take place

because the agent considers a set of implicitly inconsistent propositions (*the logical condition*) along different decision paths that lead to mutually inconsistent decisions (*the empirical condition*). I shall refer to this model as the *procedural* model, but as Gold and List (2004) point out, what the 'procedure' exactly refers to is open to several interpretations of the empirical conditions (e.g., Does an agent consider different propositions in different temporal orders? Or does she weigh those propositions differently? Or are some propositions more focal than others for the agent?).[19] Rubinstein (2003) proposes a similar procedural model in the domain of inter-temporal decision making, claiming that the procedural model is empirically superior to and more intuitive than the hyperbolic discount models (see Sect. 2.2 above for details) that are popular among behavioural economists. Although relatively new in economics, the procedural approach may have the potential to explain both inter- and intra-temporal inconsistencies of choice. The compatibility hypothesis laid the basis of this line of research, by providing detailed psychological mechanisms. One cannot therefore criticize the compatibility hypothesis as ad hoc or 'Ptolemaic,' as Ross (2005) does: the hypothesis does more than "systematically summarizing data", and if taken seriously as an explanation, it points to a rather different theoretical possibility (the procedural model) from the one Ross envisages (the multiple-self model).

6 Can 'wider theoretical considerations' help?

I have defended the compatibility hypothesis against the criticisms that it is based on a misguided theoretical framework and that it is based on no theoretical framework at all. Now I turn to Ross's other claim that the latest neuroscientific research, and the model of the mind supported by it, favour the multiple-agent model. I will show that this is not so, and why.

In criticizing Slovic et al.'s compatibility hypothesis, Ross (2005, pp. 233, 235) makes a suggestive contrast between the "classical model of the mind" and the new model which indicates that human brains are "parallel information processors". The idea seems to be that no centralized process takes place in the brain when the decision maker is considering some choice problem; instead, various parts of the brain process information in various ways without having the Cartesian 'central processer'. This idea seems congenial to Ross's hypothesis that a decision maker consists of several economic agents, each doing its own maximization based on different preferences. The contrast seems to be that in the classical model a computer-like algorithm is performed by the unitary decision maker, while in the model informed by cognitive neuroscience the decision-making processes are 'decentralized'.

Now, assume for the sake of argument that such 'decentralized' model of the mind is supported by neuroscientific or some other wider (non-behavioural) evidence. This would not favour Ross's multiple-self model relative to the procedural model. Note first that the strategy compatibility hypothesis is also consistent with this new picture of the mind. As I suggested above, it is an open question how a 'decision path' should

[19] Starmer (2000, p. 35) defines 'procedural theories' as the theories that try to model the psychological processes that lead to choice. Note that how such 'processes' are interpreted is left open under this definition.

be interpreted; Gold and List (2004, p. 260) define a decision path as "the order in which the agent considers the propositions in a sequential decision process", but suggest that several empirical interpretations are possible: if we interpret a decision path actually taken by the agent as a set of the propositions which are considered weightier, or more focal, than are other propositions, then the hypothesis will be consistent with the fact that different parts of the brain process information in a parallel way. But even if a decision path is literally interpreted as a temporal order, the hypothesis may be consistent with the decentralized processing model, which is compatible with the idea that different parts of the brain are 'taking turns', as in a sequential game between different selves. These considerations suggest that the procedural model and the multiple-agent model are both compatible with the decentralized model of the mind.

Moreover, the multiple-agent model and the procedural model may be mutually compatible (Harrison 2008, p. 337). Recall that the procedural model consists of two conditions, logical and empirical. The former is a requirement that different decision paths lead to different final decisions on the target proposition. Gold and List (2004) explicate this by claiming that an agent's initial dispositions must be *implicitly inconsistent* for this 'path dependence' to happen. An agent's initial dispositions are implicitly inconsistent with respect to a proposition φ in a set X if there exist two logically inconsistent sets of propositions Ψ_1 and Ψ_2 such that the agent has dispositions to accept all propositions in Ψ_1 and all propositions in Ψ_2, but Ψ_1 entails φ and Ψ_2 entails $\neg\varphi$. Implicit inconsistencies can happen in two ways. First, when the agent violates *deductive closure*, i.e., when there exists a logically consistent set of propositions Ψ such that the agent has dispositions to accept all propositions in Ψ, Ψ entails φ, and yet the agent has no disposition to accept φ. For example, suppose that the agent has initial dispositions to accept P and $(P \Rightarrow Q)$, but for some reason she also has a disposition to accept $\neg Q$. The set $\{P, (P \Rightarrow Q)\}$ entails Q but the set $\{\neg Q\}$ (trivially) entails $\neg Q$, meaning that the agent's initial dispositions are implicitly inconsistent. Second, an implicit inconsistency occurs whenever the agent's disposition is *explicitly inconsistent*, i.e., when the agent has dispositions to accept both φ and $\neg\varphi$ simultaneously. Although Gold and List suggest that path dependence occurs mainly because deductive closure is violated, they do not exclude the possibility that the agent is explicitly inconsistent with regard to the decision on the target proposition. In this case, one way to interpret the underlying psychology is to suppose that the individual has mutually contradicting dispositions, one accepting and the other rejecting the target proposition. This interpretation is consistent with the idea that the individual really consists of more than one agent, each characterized by its own distinct set of preferences. In this sense, the procedural model is compatible with the multiple-agent model.[20] The upshot is that, because both are compatible with the decentralized model

[20] Ross has recently argued for the similar but more general view that the personal and sub-personal level models may be neither mutually exclusive (i.e., both can provide genuine explanations) nor reductive (i.e., the former is not necessarily reducible to the latter) because the two models concern "distinct scales of resolution" on the actual causal processes (Ross et al. 2008; Ross 2009). On this view, my following evaluation of the procedural and multiple-agent models regards relative explanatory merits of the two models rather than their absolute proximity to the true explanation.

of the mind that Ross advocates, and because both may be interpreted as compatible with each other in the sense explicated above, neither the decentralized model of the mind, nor neuroscientific studies purported to support it, can lend any support to one over the other model.

There is, however, a respect in which the procedural model is superior to the multiple-agent model. On the one hand, the multiple-agent model presupposes a game-like interaction among different agents in determining the final decision as an equilibrium (or equilibria). On the other, the procedural model presupposes certain cognitive or perceptual mechanisms through which one decision path rather than others is taken, depending on how problems are described or how decisions are elicited. One of the empirical conditions is explicated by the strategy compatibility hypothesis. These mechanisms are presumed to capture ways in which individuals change their responses to extensionally equivalent decision problems depending on descriptions of the problem and procedures of preference elicitation. By contrast, the multiple-agent model alone cannot explain these framing and elicitation effects. This, of course, is not the case if a multiple-agent model is coupled with some model of framing at the whole-person level. If the decision at the whole-person level derives solely from games among the agents within an individual, how could framing at the whole-person level matter? In sum, although the multiple-agent model has some intuitive and empirical appeal in the domain of inter-temporal decision making, the procedural models, in particular the strategy compatibility hypothesis, can better explain anomalies in the domain of intra-temporal decision making.

7 Conclusion

In this study, I have examined a debate on possible explanations of preference reversals (the multiple-agent model vs. the procedural model), and suggested that this particular debate cannot be resolved simply by appealing to a methodological criterion of theory appraisal (such as unification). Specifically, I have criticized Ross's (2005) claim that the compatibility hypothesis proposed by psychologists and behavioural economists is ad hoc. I have first distinguished three cases of ad hocness, based on Lakatos (1970) framework of *research programme*. Second, I have motivated Ross's critique by illustrating how the Lakatosian framework works in the domain of inter-temporal choice, and potentially also in the domain of intra-temporal choice. Third, I have argued that Ross's critique of the compatibility hypothesis as 'Ptolemaic' is unjustified, by showing that the hypothesis consists of two distinct causal hypotheses regarding cognitive processes underlying decision making. I have also suggested that the compatibility hypothesis is based on a well-motivated theoretical framework, i.e., the reason-based, procedural approach. These considerations suggest that the real issue of the debate is not methodological but empirical. Further, I have shown that this empirical debate is not easily resolved by appeal to wider theoretical considerations such as what the true model of the mind is.

Finally, it must be noted that, although I have been more sympathetic towards the procedural model than towards the multiple-self model, the scope of my examination is rather limited; it does not exclude the possibility of a better formulation of

a multiple-self model that explains the data better than any procedural models do. This is an entirely empirical issue. Also, my exercise does not deny the usefulness of unification as a criterion of theory appraisal *in general*; rather it shows its limited applicability in this specific debate. Still I believe this exercise to be useful. For after all, a general, abstract epistemic rule must be used in and evaluated against specific, concrete cases. My attempt in this study has been to present one such concrete case.

Acknowledgments I thank Don Ross, Francesco Guala, Natalie Gold referee for invaluable comments. Joel Smith and Graham Stevens saved me from many mistakes. The study also benefitted from the comments I received in Madrid (International Network for Economic Methodology conference) and Rotterdam (Philosophy PhD seminar). My final thanks go to Luis Mireles Flores who invited me to the latter event and made my stay enjoyable. The usual caveats apply.

References

Ainslie, G. W. (2001). *Breakdown of will*. Cambridge: Cambridge University Press.

Ainslie, G. W. (2005). Précis of breakdown of will. *Behavioral and Brain Sciences, 28*, 635–673.

Fischer, G. W., & Hawkins, S. A. (1993). Strategy compatibility, scale compatibility, and the prominence effect. *Journal of Experimental Psychology: Human Perception and Performance, 19*(3), 580–597.

Fudenberg, D., & Levine, D. K. (2006). A dual self model of impulse control. *American Economic Review, 96*(5), 1449–1476.

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

Gold, N., & List, C. (2004). Framing as path dependence. *Economics and Philosophy, 20*(2), 253–277.

Goldstein, W. M., & Einhorn, H. J. (1987). Expression theory of choice and the preference reversal phenomenon. *Psychological Review, 94*, 236–254.

Guala, F. (2006). Has game theory been refuted?. *The Journal of Philosophy, 103*(5), 239–263.

Harrison, G. W. (2008). Neuroeconomics: A critical reconsideration. *Economics and Philosophy, 24*, 303–344.

Harrison, G. W., Lau, M. I., & Williams, M. B. (2002). Estimating individual discount rates in Denmark: A field experiment. *American Economic Review, 92*(5), 1606–1617.

Kitcher, P. (1981). Explanatory unification. *Philosophy of Science, 48*, 507–531.

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge*. Cambridge: Cambridge University Press.

Lichtenstein, S., & Slovic, P. (1973/2006), Response-induced reversals of preference in gambling: An extended replication in Las Vegas. *Journal of Experimental Psychology* 101: 16–20. (Pages refer to the reprinted version in Lichtenstein, S. & Slovic, P. (Eds.), *The construction of preference*. Cambridge: Cambridge University Press. pp. 69–76.)

Loomes, G., Starmer, C., & Sugden, R. (1991). Observing violations of transitivity by experimental methods. *Econometrica, 59*(2), 425–439.

Loomes, G., & Taylor, C. (1992). Non-transitive preferences over gains and losses. *Economic Journal, 102*(411), 357–365.

Margolis, H. (2007). *Cognition and extended rational choice*. New York: Routledge.

McClure, S. M., Laibson, D. I., Loewenstein, G., & Cohen, J. D. (2004). Separate neural systems value immediate and delayed monetary rewards. *Science, 306*(5695), 503–507.

Read, D., & Read, N. L. (2004). Time discounting over the lifespan. *Organizational Behavior and Human Decision Processes, 94*(1), 22–32.

Ross, D. (2005). *Economic theory and cognitive science: Microexplanation*. Cambridge, MA: MIT Press.

Ross, D. (2009). Integrating the dynamics of multiscale economic agency. In H. Kincaid & D. Ross (Eds.), *The Oxford handbook of philosophy of economics* (pp. 245–279). Oxford: Oxford University Press.

Ross, D., Sharp, C., Vuchinich, R., & Spurrett, D. (2008). *Midbrain mutiny: The picoeconomics and neuroeconomics of disordered gambling*. Cambridge, MA: MIT Press.

Rubinstein, A. (2003). Economics and psychology"? The case of hyperbolic discounting. *International Economic Review, 44*(4), 1207–1216.

Rubinstein, A. (2006). A sceptic's comment on the study of economics. *Economic Journal, 116*(510), C1–C9.

Shafir, E., Simonson, I., & Tversky, A. (1993). Reason-based choice. *Cognition, 49*, 11–36.

Slovic, P. (1975). Choice between equally valued alternatives. *Journal of Experimental Psychology: Human Perception and Performance, 1*, 280–287.

Slovic, P. (1995/2000). The construction of preference. *American Psychologist*, 50: 364–371. (Reprinted in Kahneman, D. & Tversky, A. (Eds.), *Choices, values, and frames.* Cambridge: Cambridge University Press. pp. 489–502).

Slovic, P., & Lichtenstein, S. (1968/2006). Relative importance of probabilities and payoffs in risk taking. *Journal of Experimental Psychology Monograph* 78(3) part 2: 1–18. (Pages refer to the version reprinted in Lichtenstein, S. & Slovic, P. (Eds.), *The construction of preference*. Cambridge: Cambridge University Press. pp. 41–51).

Slovic, P., Griffin, D., & Tversky, A. (1990/2002). Compatibility effects in judgements and choice. In Hogarth, R. (ed.), *Insights in decision making: A tribute to Hillel J. Einhorn (pp. 5–27)*. Chicago: University of Chicago Press. (Pages refer to the version edited and reprinted in Gilovich, T., Griffin D. W., & Kahneman, D. (Eds.), *Heuristics and biases: The psychology of intuitive judgement*. Cambridge: Cambridge University Press. pp. 217–229).

Starmer, C. (2000). Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. *Journal of Economic Literature, 38*, 332–382.

Strotz, R. H. (1955). Myopia and inconsistency in dynamic utility maximization. *Review of Economic Studies, 23*, 165–180.

Tversky, A., Sattath, S., & Slovic, P. (1988/2000). Contingent weighting in judgement and choice. *Psychological Review* 95(3): 371–384. (Pages refer to the reprinted version in Kahneman, D. & Tversky, A. (Eds.), *Choices, values, and frames*. Cambridge: Cambridge University Press. pp. 503–517).

Tversky, A., Slovic, P., & Kahneman, D. (1990). The causes of preference reversal. *American Economic Review, 80*(1), 204–217.

Tversky, A., & Thaler, R. (1990). Anomalies: Preference reversals. *The Journal of Economic Perspectives, 4*(2), 201–211.

Wilcox, N. T. (2008). Stochastic models for binary discrete choice under risk: A critical primer and econometric comparison. *Research in Experimental Economics, 12*, 197–292.

Wilkinson, N. (2008). *An introduction to behavioral economics*. New York: Palgrave Macmillan.