THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

DEPARTMENT OF PHILOSOPHY, LOGIC AND SCIENTIFIC METHOD

PHILOSOPHICAL FOUNDATIONS OF NEUROECONOMICS

ECONOMICS AND THE REVOLUTIONARY CHALLENGE FROM NEUROSCIENCE

Roberto Fumagalli

A dissertation submitted to the Department of Philosophy, Logic and Scientific Method of the London School of Economics for the degree of Doctor of Philosophy, September 2011.

Declaration

I certify that the thesis I have presented for examination for the PhD degree of the London School of Economics is solely my own work other than where I have clearly indicated that it is the work of others.

This authorization does not, to the best of my belief, infringe the rights of any third party.

Abstract

This PhD thesis focuses on the philosophical foundations of Neuroeconomics, an innovative research program which combines findings and modelling tools from economics, psychology and neuroscience to account for human choice behaviour. The proponents of Neuroeconomics often manifest the ambition to foster radical modifications in the accounts of choice behaviour developed by its parent disciplines. This enquiry provides a philosophically informed appraisal of the potential for success and the relevance of neuroeconomic research for economics. My central claim is that neuroeconomists can help other economists to build more predictive and explanatory models, yet are unlikely to foster revolutionary modifications in the economic theory of choice.

The contents are organized as follows. In chapters 1-2, I present neuroeconomists' investigative tools, distinguish the most influential approaches to neuroeconomic research and reconstruct the case in favour of a neural enrichment of economic theory. In chapters 3-7, I combine insights from neuro-psychology, economic methodology and philosophy of science to develop a systematic critique of Neuroeconomics. In particular, I articulate four lines of argument to demonstrate that economists are provisionally justified in retaining a methodologically distinctive approach to the modelling of decision making.

My first argument points to several evidential and epistemological concerns which complicate the interpretation of neural data and cast doubt on the inferences neuroeconomists often make in their studies. My second argument aims to show that the trade-offs between the modelling desiderata that neuroeconomists and other economists respectively value severely constrain the incorporation of neural insights into economic models. My third argument questions neuroeconomists' attempts to develop a unified theory of choice behaviour by identifying some central issues on which they hold contrasting positions. My fourth argument differentiates various senses of the term 'revolution' and illustrates that neuroeconomists are unlikely to provide revolutionary contributions to economic theory in any of these senses.

Acknowledgements

To my parents and close relatives, whose loving support and encouragement helped me to navigate the stormy waters of my academic journey.

To my doctoral supervisors, Jason and Richard, who patiently bore the burden of my intellectual curiosity.

To a number of scholars and colleagues who had occasion to comment on my research at seminars, workshops and conferences.

To the fellow members of the Goodenough College, where I spent the last three years of my doctoral studies: I shall treasure their companionship for the rest of my life.

CONTENTS

POETIC FOREWORD

“Imagine a vast and restless sea, constantly vexed by violent storms.
In the middle of the sea, you and your friends bravely proceed
Strenuously fighting against the hostile winds.

Day after day, the motion of the waves impairs your fragile vessel,
And opens bursts in the hull, and nearly makes the ship shrink:
Till the day comes, when some members of the crew leave.

Yet, the boat is at open sea and can still proceed, though not to proper speed:
Shall you follow your comrades, abandoning the vessel for the unknown?
Or shall you further rest on board, trying to repair the boat?

Before deciding, consider this: the restless marine is the economic realm;
The waves stand for the behavioural, experimental and neuroscience drifts;
The little crew on the open sea is the community of economists”.

INTRODUCTION

Revolutionary scientific change has been the focus of intense philosophical controversies in the second half of the twentieth century. Before those days, scientific progress was usually regarded as a cumulative process whereby novel theories supplant and occasionally reduce earlier ones. In 1962, Kuhn publishes *The Structure of Scientific Revolutions*, which provides an innovative and highly controversial account of intertheoretic transitions. Kuhn's model differs profoundly from the conception of scientific change that was popular among logical positivists, who characterized scientific progress in terms of logical-mathematical derivations between individual theories.

Kuhn focuses not so much on isolated theories, but rather on paradigms, that is "disciplinary matrices" including elements such as taxonomies for the investigated phenomena and scientists' methodological commitments (1970a, p.182). In his view, scientific disciplines undergo sharply delimited phases of normal and revolutionary science. During periods of normal science, scientists take for granted the validity of the received paradigm and regard experimental results that bear against it as anomalies calling for further investigation. If anomalies accumulate, however, scientists come to question the validity of the prevailing paradigm, and the relevant discipline enters a period of revolutionary science.

Phases of revolutionary science culminate in scientific revolutions, that is "non-cumulative developmental episodes in which an older paradigm is replaced […] by an

incompatible new one" (1962, p.86; see also Kuhn, 1981). According to Kuhn, scientific revolutions involve radical modifications in scientists' beliefs and commitments. These modifications are often so profound that the proponents of competing paradigms "live in different worlds" (1970a, p.193).

Kuhn's account of scientific change was soon subject to a number of objections (see e.g. Achinstein, 1968, Lakatos, 1970, and Toulmin, 1972). In response to criticisms, Kuhn (1970a, 1974 and 1981) came to endorse a less extreme view of intertheoretic transitions (see e.g. Sankey, 1993 and 1994). However, his initial model prompted animated discussions among the practitioners and the philosophers of various disciplines. For instance, some economists (e.g. De Vroey, 1975) took several episodes in the history of economic theory to fit with Kuhn's account, while others (e.g. Baumberger, 1977, and Blaug, 1975) argued that a gradualist model of scientific progress provides a more accurate characterization of the development of their discipline.

In the history of science, scientific revolutions have been claimed to occur both in underdeveloped disciplines (e.g. think of the overthrow of the phlogiston theory by Lavoisier's chemical theory) and in mature sciences (e.g. think of the replacement of Newton's theory of universal gravitation by Einstein's general theory of relativity). One such revolution is allegedly under way at the interface between economics, psychology and neuroscience. The story goes as follows. Over the last decade, a growing body of research has come together under the name of Neuroeconomics (henceforth, NE), an emerging discipline which combines findings and modelling tools from economics,

psychology and neuroscience to account for human choice behaviour. Neuroeconomists (henceforth, NEs) aim to integrate findings and modelling tools from NE's parent disciplines. In particular, they often speak of implementing revolutionary modifications in the accounts of decision making provided by those disciplines (see e.g. Camerer, Loewenstein and Prelec, 2005, and Glimcher, 2010).

There are at least two respects in which the emergence of NE promises to constitute an especially significant episode of scientific change. The first one concerns the *scope* of NEs' proposed revolution. Economists, psychologists and neuroscientists have separately achieved significant successes in modelling and explaining choice behaviour. However, they usually employ dissimilar constructs and pursue different explanatory goals (see e.g. Glimcher, 2010, ch.1). The pioneers of NE frequently manifest the ambition to develop a single, unified theory of choice behaviour that spans NE's parent disciplines and "transcends the explanations available to neuroscientists, psychologists, and economists working alone" (Glimcher and Rustichini, 2004, p.452). A second peculiarity of NEs' intended revolution relates to its purported *depth*. The proponents of NE rarely rest content with integrating particular findings from economics, psychology and neuroscience. On the contrary, they often speak of substituting the constructs traditionally employed in these disciplines. For instance, after noting that we are able to "observe the brain better than ever before", Camerer claims that NEs "will eventually […] replace the simple mathematical ideas of economics with more neurally-detailed descriptions" (2005).

The aforementioned assertions point to momentous interdisciplinary rearrangements and raise several questions concerning revolutionary scientific change. To give some examples, are NEs likely to foster a genuine revolution spanning both natural and social disciplines? What obstacles stand in the way of realizing this ambitious project? More generally, do scientific revolutions constitute objective advances in the involved disciplines, or do they call into question the purported progressiveness of science? What implications do scientific revolutions have for the metaphysics and the epistemology of science (see e.g. van Fraassen, 1980, Hacking, 1983, Boyd, 1984, Psillos, 1999, and Stanford, 2010, for a debate between realist and antirealist interpretations of scientific theories)?

The rise of NE provides us with a valuable opportunity to investigate a number of philosophically relevant issues besides these ones. Let us briefly consider some of these issues in turn. The first one concerns the ideal of *interdisciplinary consilience*. The practitioners of economics, psychology, and neuroscience have developed a wide variety of approaches to model human choice behaviour. Regrettably, their models usually have quite a narrow scope and are rarely integrated into unified theories at the intradisciplinary level. Indeed, even when targeting the same explananda, the practitioners of each discipline often rely on dissimilar presuppositions and make use of distinct constructs (e.g. compare the *homo oeconomicus* posited by rational choice theory with the agents figuring in bounded rationality and ecological rationality models). The situation is even more fragmented at the interdisciplinary level, where we presently lack a shared methodology for building a unified account of choice behaviour. In this respect, several questions arise regarding NEs' attempts to integrate evidence and

modelling tools from different behavioural sciences. For example, what findings and constructs are to be employed in developing NE models? To what extent do the methodological divergences between economists, psychologists and neuroscientists affect the prospects of NE?

Secondly, the emergence of NE has important applications to philosophical debates regarding *intertheoretic reduction* (see e.g. Nagel, 1961, and Oppenheim and Putnam, 1958). Interdisciplinary research often prompts intense disputes concerning the nature of intertheoretic relations among the practitioners of the involved disciplines. NEs' proposals proved to be especially controversial for several reasons. To give one example, their attempts to provide a neural microfoundation to economic theory promise to accomplish the first instance of intertheoretic reduction spanning both natural and social scientific disciplines. However, NEs' reductive claims hardly fit with some philosophers' criticisms of intertheoretic reductions (see e.g. Fodor, 1974, Duprè, 1983 and 1993, and Cartwright, 1999). Moreover, they have been questioned on the more pragmatic ground that it is more fruitful to pursue integrative - rather than reductive - approaches between NE's parent disciplines (Craver and Alexandrova, 2008). Do NEs possess the means to develop a reductive unification of economics, psychology and neuroscience? Do they concur on which of these disciplines is best equipped to provide the fundamental constructs for the NE theory of choice? What are the prospects of a non-reductive unification of NE's parent disciplines?

Thirdly, the rise of NE constitutes an especially suitable case study for investigating how notions of *explanation* and criteria of *explanatory relevance* vary across

disciplinary boundaries. In the former respect, the question arises as to how exactly the conceptions of explanation that are respectively presupposed in distinct natural and social sciences differ. What kind of explanations do economists attempt to develop? Are they concerned with providing mechanistic accounts of people's decisions, or do they aim to understand the reasons motivating economic agents (see e.g. Knight, 1935, and Davidson, 1963)? What kinds of methods are best suited for economists' explanatory purposes (see e.g. Mill, 1843, and Weber, 1904)? As to criteria of explanatory relevance, the mere fact that decision making takes place in the brain does not license the conclusion that NE findings are relevant for economics. Still, several NEs presuppose that understanding how decision making is instantiated at the neuro-psychological level is *ipso facto* informative to economists. In doing so, they rely on disputable assumptions concerning the explanatory relevance of neuro-psychological findings for the economic theory of choice (Kuorikoski and Ylikoski, 2010).

A fourth issue of philosophical significance relates to the *pragmatics of modelling* in science. The proponents of NE frequently argue that economists could develop more predictive and explanatory models by incorporating neuro-anatomical and neuro-physiological insights. However, distinct modelling desiderata (e.g. think of tractability and descriptive accuracy) often pull in different directions and make opposing demands on modellers. The trade-offs between distinct desiderata, in turn, impose significant restrictions on the construction of models spanning different disciplines and levels of description. In this respect, one wonders whether - and, if so, on what grounds - economists should include several neural insights into their models of choice. This question has important implications for the potential significance of NE research, as the

trade-offs between the desiderata that NEs and other economists respectively value are unlikely to abate with progress in NEs' observational tools and experimental practices.

A fifth issue concerns the *interpretation* that *models* of decision making are given in different behavioural sciences. The following contrast is especially profound in the current literature at the interface between economics, psychology and neuroscience. On the one hand, economists usually rely on as if models of choice (e.g. think of expected utility theory) which make no assumptions regarding what neuro-psychological processes underlie choice behaviour. On the other hand, many NEs take their models to provide descriptively accurate characterizations of the neuro-psychological substrates of choice behaviour. In particular, some NEs (e.g. Camerer, Loewenstein and Prelec, 2005, p.10, and Glimcher, 2010, p.126 and 133) urge economists to substitute their as if representations with mechanistically informed accounts of decision making. Regrettably, the availability of multiple NE models of choice which posit dissimilar neuro-psychological processes does not fit well with the realistic interpretation many NEs give to those models.

Last but not least, NEs' contributions raise several questions of *normative* significance, especially regarding economic welfare analyses and policy evaluations. The proponents of NE often manifest the ambition to evaluate people's decisions according to a normative perspective. In particular, they aim to ascertain not just what the best way to achieve a given objective is, but also what objectives agents should pursue in specific situations. In what circumstances, if any, do neuro-psychological findings legitimize NEs to influence or interfere with people's decisions? What sort of paternalistic

interventions might be advocated in designing and modifying particular choice architectures? To be sure, more accurate knowledge of the neural substrates of choice behaviour may enable NEs to develop more informative and reliable indicators of well-being. Nonetheless, various factors constrain the relevance of NE findings for economists' normative analyses. To give one example, neuro-psychological empirical evidence does not *per se* provide compelling indications as to what agents ought to choose in specific decision settings. Furthermore, profound differences remain between NEs' and other economists' conceptions of well-being. More specifically, standard economic theory does not take a position as to what agents' objective well-being consists in. For their part, many NEs relate agents' objective well-being to their hedonic states or the activation patterns of particular neural areas.

In the following chapters, I explore these and other philosophically relevant issues with the aim to assess the prospects and the relevance of NE for one of its parent disciplines, namely economics. In doing so, I shall mention in passing what impact NE findings may have on neuro-psychological research and society at large (e.g. think of futuristic forms of neural marketing or the therapeutic benefits derivable from neurally informed accounts of addictive behaviour). My focus on economic theory is motivated both by the significance that NEs' contributions allegedly have for the economic account of decision making and by the lively debates that some NEs' assertions have fostered among economists. Let me expand on this point.

The pioneers of NE often argue that economists can considerably improve their models of choice by incorporating neuro-psychological variables. Some NEs go as far as to

advocate the replacement of various fundamental constructs (e.g. constrained utility functions, optimization tools) of standard economic theory. NEs' calls for a neural enrichment of economic theory have prompted heterogeneous reactions among economic modellers and methodologists. To a first approximation, three prototypical positions can be distinguished in the economists' camp. On the one hand, the *sceptics* doubt (e.g. Harrison, 2008a and 2008b, and Rubinstein, 2008) or even deny (e.g. Gul and Pesendorfer, 2008) the relevance of NEs' contributions for the economic theory of choice. On the other hand, the *enthusiasts* (e.g. Rustichini, 2005) contend that incorporating neuro-physiological insights into economic models will have significant, and arguably revolutionary, implications. In this highly simplified picture, a halfway position is advocated by the *moderates* (e.g. Smith, 2007, ch.14), who cautiously note that it is too soon to judge NEs' achievements and that the extent to which NE will inform mainstream economic theory remains an open empirical question.

When it comes to assessing the potential for success in NE research, many authors refrain from judgement by alleging that NE is a relatively young discipline whose prospects depend on empirical findings that are still to come. *Prima facie*, this moderate stance may seem preferable to the other two positions, as adopting a 'wait and see' attitude is less risky than pontificating about the future of economics, psychology and neuroscience. However, prudently postponing judgement does not appear to be the best way to evaluate the prospects of the NE enterprise. After all, the fact that the case for NE is "mostly based on promise" (Camerer 2008a, p.62) does not prevent one from examining the grounds on which such promise rests. Indeed, one may argue that precisely because the advancement of NE depends on somewhat speculative

assumptions, it is especially important to discriminate between fruitful research avenues and misleadingly attractive dead ends.

The central claim of this thesis is that NEs' contributions help economists build more predictive and explanatory models of choice, yet are unlikely to foster revolutionary modifications in economic theory[1]. As I shall illustrate below, the reasons for my scepticism run deep in the methodological foundations of NE's parent disciplines and involve a number of interrelated philosophical issues. To be clear, I am aware that methodological debates occasionally degenerate into self-referential speculative exercises, and I share the reluctance of many NE practitioners to engage in hair-splitting which might appear to be of little help to the profession. Still, the NE literature is growing very rapidly, with profound dissimilarities in the way different authors conceptualize and develop their research. In such a context, the opportunity - and arguably, the need - arises for a scrupulous methodological appraisal, which enables economists, psychologists and neuroscientists to more accurately assess the merits of NEs' proposals.

My investigation can be broadly divided into two parts. In *chapters one* and *two*, I place the emergence of NE into dialectical context, present the main investigative tools of NEs and reconstruct their case in favour of a neural enrichment of economic theory. In *chapters three* to *seven*, I combine recent neuroscientific findings with considerations from economic methodology and philosophy of science to develop a systematic critique of NE. In such a context, I examine both what more accurate knowledge of the human

[1] In this enquiry, I shall employ expressions such as "economic theory of choice", "traditional theory of choice", "economic account of decision making", etc. to refer to both decision theory and game theory.

neural architecture may add to our understanding of economic behaviour and whether such knowledge justifies a significant import of neural data into economic models of choice. In particular, I articulate and defend several lines of argument (see below) which aim to demonstrate that economists are provisionally justified in retaining a methodologically distinctive approach to the modelling of decision making. Let me anticipate briefly the contents of each chapter in turn.

In *chapter one*, I provide a general framework for understanding and assessing NE research. After comparing the main definitions that NE has been given in the literature, I identify three major respects in which NEs' contributions can be differentiated. I then relate the emergence of NE to previous research at the interface between economics and psychology. Finally, I present the brain-imaging and brain-stimulation instruments that NEs frequently employ in their studies. In *chapter two*, I identify several respects in which incorporating neural insights can help economists to improve their models. In doing so, I reconstruct the main arguments that have been provided in support of a neural enrichment of the economic theory of choice. Moreover, I illustrate how NEs' arguments can be combined in a cumulative case for the neural enrichment of economic theory.

In *chapter three*, I discuss various evidential and epistemological concerns which arise in relation to the collection and the interpretation of neural data. Furthermore, I critically examine the inferences made in many brain-imaging and brain-stimulation studies. In particular, I distinguish between some problems that are likely to be resolved thanks to advances in scanner technology and others that are unlikely to abate with

scientific progress. In *chapter four*, I attempt to demonstrate that NEs overestimate the extent to which their contributions improve economic models of choice. Moreover, I identify some trade-offs between the modelling desiderata that NEs and other economists respectively value and argue that these trade-offs severely constrain the incorporation of neural insights into economic models of choice.

Chapter five calls into question NEs' attempts to develop a unified interdisciplinary theoretical framework by pointing to the profound dissimilarities (e.g. in terms of employed constructs and pursued explanatory aims) between the economic, psychological and neuroscientific accounts of decision making. In such a context, I cast doubt on the possibility of combining NEs' contributions in a cumulative case in favour of NE by identifying some central respects (e.g. how NE is supposed to inform standard economic theory) in which NEs themselves hold contrasting positions. In *chapter six*, I provide a case study which aims to illustrate how the conceptual differences between NE's parent disciplines constrain the relevance of neuro-psychological findings for the economic theory of choice. More specifically, I distinguish three notions of utility - namely decision utility, experienced utility and neural utility - that are frequently mentioned in debates over decision theory. I subsequently examine some critical issues regarding their definition and measurability. In doing so, I critique NEs' calls to replace decision utility with experienced and neural utility as a central concept of decision theory.

The *final chapter* relates the ongoing debate between NEs and other economists to the philosophical literature on revolutionary scientific change. In particular, I differentiate

various senses in which NE has been claimed to revolutionize economic theory and argue that NEs are unlikely to prompt revolutionary contributions in any of these senses. In the *conclusion*, I evaluate the prospects of NE in light of Lakatos' distinction between progressive and degenerating research programs. I then summarize the main problems impeding the advancement of NE and provide some brief remarks regarding the future of NE research.

NEs' contributions are making inroads into both natural and social scientific research, attracting increasing attention and financial resources. However, NE presently constitutes a highly fragmented discipline, whose relation to economics, psychology and neuroscience is hard to characterize precisely. Furthermore, a number of conceptual and empirical issues still wait to be sorted out and explored in the NE literature. This enquiry aims to provide one of the first philosophically informed methodological appraisals of the *relevance* of NE for economic theory and the potential for *success* in NE research. My overall goal is not just to critique the proponents of NE, but also to prompt them to build their case in favour of 'mindful economics' on more solid empirical and conceptual foundations.

CHAPTER ONE

WHAT IS NEUROECONOMICS?

What is NE? How exactly is it related to the economic theory of choice? At first approximation, NE can be characterized as an interdisciplinary enterprise which combines findings and modelling tools from economics, psychology and neuroscience to account for human choice behaviour. At closer examination, however, profound dissimilarities can be found between NEs' approaches. In this chapter, I provide some conceptual and terminological distinctions that will later help us to assess the relevance of NE for its parent disciplines. More specifically, section 1.A examines the most influential definitions that NE has been given in the literature and differentiates various approaches to NE research. In section 1.B, I reconstruct the dialectical context in which NE emerged, devoting particular attention to previous developments in research at the interface between economics and psychology. In section 1.C, I present the brain-imaging and brain-stimulation tools that NEs employ in their investigations, highlighting the main strengths and limitations of each instrument.

1.A DEFINITIONS AND APPROACHES

Over the last few decades, economists have integrated a number of insights from other behavioural sciences into their models of choice. The development of interdisciplinary research programs such as behavioural and experimental economics represented a significant advancement in the economic account of decision making[2]. In recent years, a growing body of research has come together under the name of NE. In spite of its relatively recent origin, NE has already been characterized in remarkably different ways both by NEs and by other researchers. The following list illustrates the diversity of the definitions formulated by the pioneers of the discipline.

i) Some authors speak of NE in distinctively *interdisciplinary* terms. McCabe (2003a), for example, depicts it as "an interdisciplinary research program with the goal of building a biological model of decision making" (see also McCabe, 2003b and 2008). In a similar vein, Glimcher and Rustichini (2004, p.447) characterize NE as the attempt to combine economics, psychology and neuroscience "into a single, unified discipline with the ultimate aim of providing a single, general theory of human behaviour" (see also Glimcher, 2010, p.393, and Rustichini, 2005, p.203-4).

ii) Other times, NE is presented as a specific *application* of economic theory to the modelling of the human neural architecture. For instance, McCabe (2008, p.346) notes that economists' optimization techniques offer neuroscientists a useful way to

[2] The expression "decision making" is often used to denote both observed choice behaviour and the underlying cognitive and computational processes. In what follows, I employ such an expression to refer to observed choice behaviour, without taking a position as to whether economists *qua* economists should be concerned with those processes.

characterize the workings of the brain and maintains that NE represents "an increasingly important route for the export of economic ideas". For their part, Glimcher, Dorris and Bayer (2005, p.253) argue that utility theory provides "the ultimate set of tools" for modelling neural areas' activation patterns. Indeed, they go as far as to claim that while economists typically assume that "it is *as if* expected utility was computed by the brain", neuroscience "suggests an alternative, and more literal, interpretation", according to which "the neural architecture actually does compute desirability for each available course of action" (ibid., p.220; see also Platt and Glimcher, 1999).

iii) Some NEs characterize their discipline as an *extension* of distinct economic research programs. For example, Camerer (2003) defines NE both as a "branch" of behavioural economics, which "expands behavioral economics by using facts about brain activity", and as "a new kind" of experimental economics, which "expands experimental economics by measuring biological and neural processes to understand how people choose, bargain and trade" (see also Camerer, 2007, C26, and Camerer, 2008a, p.44). Zak (2004, p.1737), instead, argues that NE is a "natural extension" of both behavioural economics and the bioeconomic research program[3].

iv) Again differently, NE is occasionally regarded as an application of *neuroscientific* techniques and methods to the economic account of decision making. For instance, the economist Rustichini (2005, p.201) speaks of NE as "a set of papers that apply the concepts, methods, and technical tools of neuroscience to economic analysis".

[3] The idea is that bioeconomics primarily investigates how past processes of natural selection influence contemporary humans' choice behaviour, whereas NE studies the current neural underpinnings of decision making (Vromen, 2007, p.145-6; for further details on the bioeconomic research program, see Landa and Ghiselin, 1999).

Similarly, Zak (2004, p.1737) depicts NE as "an emerging transdisciplinary field that uses neuroscientific measurement techniques to identify the neural substrates associated with economic decisions" (see also Sugrue et al., 2005, p.363).

The claims reported above provide dissimilar characterizations of the theoretical presuppositions and the explanatory aims of NE. To see this, let us compare the contentions presented at points *ii* and *iv* above. On the one hand, Glimcher, Dorris and Bayer (2005) and McCabe (2008) emphasize the suitability of economic modelling tools for representing neural events and processes. On the other hand, Rustichini (2005) and Zak (2004) advocate the modification of the economic theory of choice in light of neuroscientific insights. A proponent of NE might endorse both of these positions without *ipso facto* incurring inconsistencies. For instance, she may argue - in line with Glimcher, Dorris and Bayer (2005, p.215) - that insights from economic theory effectively "guide neurobiological experiments which can, in turn, yield new economic theories". Even so, the point remains that considerable differences can be found in the way distinct NEs conceptualize their own research.

Indeed, not just NEs but also other economists depict the subject matter and the methodology of NE in heterogeneous terms. For example, Payzan and Bourgeois-Gironde (2005, p.2) view NE as a "joint experimental production between neural sciences and experimental economics". Gul and Pesendorfer, instead, characterize it as a research program based on the following two tenets:

> "Assertion I: Psychological and physiological evidence […] can be used to support or reject economic models or even economic methodology. Assertion II: What makes individuals happy ('true utility') differs from what they choose. Economic welfare analysis should use true utility rather than the utilities governing choice" (2008, p.3).

The situation appears to be even more intricate when one considers what *kinds* of NE research have been differentiated in the literature. To render this point more vivid, let us examine the following three examples. In a recent article, Montague (2007a, p.219) argues that there are "two natural [NE]", one which investigates "the way that neural tissue is built, sustains itself through time, and processes information efficiently", the other which primarily examines "the behavioural algorithms running on such a neural tissue". Craver and Alexandrova (2008, p.381-2), instead, differentiate between *neuroeconomics proper*, whose goal is "to explain economic behaviour by revealing how brain mechanisms work", and *economic neural modelling*, which exports economic concepts "in models of brain processes or in the analysis of data delivered by neuroscientific techniques". For his part, Ross (2008a, p.473) distinguishes between *behavioral economics in the scanner*, which uses neuroimaging data to foster the replacement of "standard aspects of microeconomic theory by facts and conjectures about human psychology", and *neurocellular economics*, which relies on economists' constrained maximization and equilibrium analyses "to model relatively encapsulated functional parts of brains" (see also Harrison and Ross, 2010).

I am not concerned here with comparing or assessing the above categorizations. For the purpose of this enquiry, it suffices to note that different researchers - and, at times, the

same author in different papers - propose quite dissimilar definitions of NE and employ such a term to refer to distinct bodies of research[4]. In this respect, one might well claim that we should not overemphasize the importance of drawing sharp disciplinary boundaries (Montague, 2007b, p.407). This, however, implies neither that the existing characterizations of NE are equally appropriate nor that any contribution at the interface between economics, psychology and neuroscience can be plausibly regarded as an advance in NE. In the remainder of this section, I identify three major respects in which NEs' studies can be differentiated. More specifically, I shall distinguish NEs' positions regarding: (1) the *explananda* they target; (2) the *kind of accounts* they aim to develop; and (3) how their accounts *relate* to the economic theory of choice.

1) What are the *explananda* targeted by NEs?

A first criterion in light of which distinct approaches to NE research can be differentiated concerns the explanatory targets of NEs. As anticipated above, most NEs are concerned with investigating the neural substrates of choice behaviour. This, however, falls short of implying that they pursue the same explanatory goals. In this respect, it is useful to distinguish between the *proximate* and the *ultimate* explanatory targets of NEs. The former expression refers to the neural evidence with which most NEs are directly concerned (e.g. think of the activation patterns of particular neural areas). The expression "ultimate explananda", instead, relates to the phenomena that NEs aim to explain by means of such neural evidence. As I argue below, two main positions can be differentiated with regard to the ultimate explananda of NE studies.

[4] Compare, for instance, the characterizations of NE that Camerer puts forward in his articles (e.g. 2005, 2008a and 2008b). Despite invariably speaking of "neuroeconomics", he employs this term in quite different senses.

On the one hand, NEs frequently manifest the ambition to accurately identify the algorithms implemented by the human neural architecture and the inner workings of specific neural populations (*neural ultimate explananda*). For instance, Glimcher, Dorris and Bayer (2005, p.215) argue that NE experiments "can be much more than efforts to locate a brain region associated with some hypothetical faculty" and "will reveal the nature of the economic computations brains perform". On the other hand, NEs often employ neuro-physiological evidence in order to provide more adequate explanations of observed decisions (*behavioural ultimate explananda*). The idea (e.g. McCabe, 2003a and 2003b) is that more accurate knowledge of the human neural architecture helps us to better account for the heterogeneity of human choice behaviour. To be sure, several NEs target both neural and behavioural ultimate explananda in their investigations. Still, these two sets of explananda are not coextensional, and various authors refer to the divide between them in demarcating separate approaches to NE research (see e.g. Craver and Alexandrova, 2008, on "neuroeconomics proper" and "economic neural modelling", and Ross, 2008a, on "behavioral economics in the scanner" and "neurocellular economics").

2) What *kind of accounts* do NEs aim to develop?

My second question asks what kind of accounts of choice behaviour NEs attempt to provide in their studies. In this respect, an instructive distinction can be drawn between NEs' *short term* and *long term* goals. Regarding short term goals, most NEs rest content with developing neurally enriched models of specific behavioural patterns, ranging from

trust and reciprocity (e.g. Zak et al., 2004, 2005 and 2007) to addictive gambling (e.g. Ross et al., 2010, ch.1 and 5). When it comes to long term goals, several NEs manifest the more ambitious aspiration to provide a unified theoretical framework for modelling decision making. For example, Glimcher and Rustichini (2004, p.447) characterize NE as the combination of economics, psychology and neuroscience "into a single, unified discipline with the ultimate aim of providing a single, general theory of human behaviour". In a similar vein, Rustichini (2005, p.203-4) maintains that NE attempts "to complete the research program that the early classics (in particular Hume and Smith) set out in the first place: to provide a unified theory of human behaviour". Indeed, some NEs aim to provide - not merely descriptive, but also - prescriptive insights concerning people's decisions. For instance, Glimcher, Dorris and Bayer (2005, p.214) conjecture that by combining economic and neuroscientific approaches NEs will develop "a methodology for reconciling prescriptive and descriptive economics"[5]. The idea is that NE findings cast light not just on the causal underpinnings of observed decisions, but also on what people ought to choose in specific situations.

3) How do NEs' accounts *relate* to the economic theory of choice?

The pioneers of NE often speak of building an interdisciplinary theoretical framework spanning NE's parent disciplines. The idea is to provide "a mechanistic, behavioral, and mathematical explanation of choice that transcends the explanations available to neuroscientists, psychologists, and economists working alone" (Glimcher and

[5] See alsoVromen (2010a, p.30-1) on the possibility of combining neuro-psychological research, which identifies "the evolutionary problems that our brains evolved to solve", and standard economic theory, which offers a "normative benchmark" by specifying the optimal solution to those problems.

Rustichini, 2004, p.452). However, NEs advocate heterogeneous views concerning the interdisciplinary relationship that purportedly holds between NE and its parent disciplines. In particular, two main positions can be contrasted with regard to how NE supposedly relates to the economic theory of choice. Let me expand on this divide.

In a 1998 article, the economist Rabin distinguishes two ways in which psychological findings can inform the economic account of decision making. On the one hand, he takes some of these findings to suggest *partial* modifications to rational choice models without challenging the way in which those models are typically constructed, i.e. maximization of a utility function under variously definable constraints. On the other hand, he contends that the difficulties people encounter in evaluating their own preferences and experienced well-being point towards "a more *radical* critique" of economic theory, which casts doubt on economists' use of "coherent" and "stable" utility functions (Rabin, 1998, p.12, italics mine; see also Rabin, 2002).

In their 2005 manifesto, Camerer, Loewenstein and Prelec (p.10, italics mine) propose a similar distinction concerning how neuroscientific findings can inform the economic theory of choice:

> "In the *incremental* approach, neuroscience adds variables to conventional accounts of decision making or suggests specific functional forms to replace 'as if' assumptions that have never been well supported empirically […] The *radical* approach involves turning back the hands of time and asking how economics might have evolved differently if it had been informed from the start by insights and findings now available from neuroscience".

These assertions point to what is commonly regarded as a fundamental divide between distinct approaches to NE research. However, the above passage can be interpreted in a variety of ways, and not all of these readings are equally persuasive. In particular, a literal interpretation does not appear to be particularly plausible. To see this, let us consider the aforementioned characterizations of the incremental and the radical approach in turn.

i) Concerning *incremental* NE, it remains obscure what exactly Camerer, Loewenstein and Prelec mean when they prefigure (2005, p.10) the replacement of "as if assumptions that have never been well supported empirically". To be sure, one may well complain about the purported ad hocness or non-falsifiability of some economists' as if defences of rational choice theory. Moreover, the mere fact that virtually any decision can be rationalized in terms of some as if representation falls short of implying that all economists' as if models plausibly account for observed choices. Even so, there are at least two reasons to doubt NEs' calls (e.g. Camerer, 2008a, p.47, and Rustichini, 2005, p.203) to evaluate economic models of choice in terms of their neuro-computational and neuro-cognitive plausibility. Firstly, economic models of choice are not meant to accurately characterize the neuro-computational and neuro-cognitive substrates of people's decisions. And secondly, there are several criteria besides neuro-computational and neuro-cognitive plausibility in terms of which a model can be assessed (e.g. think of tractability). In this respect, it remains unclear why economists should adopt the evaluative standards employed by modellers whose methodological presuppositions and explanatory aims sharply differ from their own.

ii) Even more puzzling is the counterfactual scenario that Camerer, Loewenstein and Prelec (2005) depict in presenting the *radical* approach. More specifically, their characterization of radical NE appears to be vulnerable to at least three criticisms. Firstly, one wonders whether it makes sense to ask how economics might have evolved if current neuroscience had influenced it from "the start". What does "the start" of economic theory stand for? How are we supposed to identify it? Secondly, it is hard to see how we could reliably ascertain how economics might have evolved, had it been informed by the insights now available from neuroscience. Maybe economists would have developed quasi-infallible neurally informed models with tremendous predictive credentials. Or perhaps they would have fallen prey of irredeemable confusion due to pan-explanatory *hubris*. In short, the range of possibilities is so wide that favouring one particular counterfactual scenario would appear to be quite arbitrary. Finally, it remains obscure how exactly speculating about counterfactual developments of economic theory is supposed to inform the current debate over the relevance and the prospects of NE. After all, the point is that we now have some powerful neuroscientific tools at our disposal, and it is the current availability of these instruments which raises issues of methodological significance.

At this stage, one may wonder whether a more plausible characterization of the incremental/radical divide can be provided. As I argued elsewhere (Fumagalli, 2010, Sec.I), incremental NEs typically rest content with enriching specific economic *models* in light of neuro-physiological findings, whereas radical NEs aim to implement substantial changes in economic *theory*. On this account, what the difference between

incremental and radical NE amounts to partly depends on which conception of scientific theories and models one endorses. I shall expand on this issue in *chapter two*. For now, we can explicate the distinction between incremental and radical NE as follows. On the one hand, incremental NEs work on the assumption that traditional economic theory and its axiomatic apparatus offer a suitable basis for modelling people's decisions. On the other hand, radical NEs urge economists to adopt a mechanistic approach to the modelling of choice behaviour and speak of complementing or even replacing economists' constructs such as preference relations and standard equilibrium concepts[6].

One might rebut that the above characterization misrepresents radical NE as an implausibly ambitious project. Nonetheless, the advocates of NE frequently put forward enthusiastic comments in relation to such a far-reaching enterprise. For instance, Camerer, Loewenstein and Prelec (2005, p.10 and 15) boldly assert that neuroscience "points to an entirely new set of constructs to underlie economic decision making" and that NEs will "substitute familiar distinctions between categories of economic behavior […] with new ones grounded in neural detail" (see also Camerer, 2005). Similarly, Rustichini (2003) optimistically speaks of NE as a "revolution" which will soon provide "a theory of how people decide in economic and strategic situations". In this perspective, it is not surprising that most economists - while cautiously welcoming the proposals of incremental NEs - oppose the contributions of radical NEs. For radical NEs

[6] The incremental/radical divide cuts across other informative distinctions regarding NE research. To see this, suppose that you wanted to classify NEs' proposals in terms of their *relevance* for the economic theory of choice. On the one hand, an incremental modification may have a considerable significance (e.g. think of a neuro-physiological variable whose incorporation enabled NEs to more accurately predict people's decisions in heterogeneous choice settings). On the other hand, a radical contribution might have limited relevance for the economic theory of choice (e.g. think of an innovative model which can be applied only to highly controlled experimental settings).

attempt to alter or even replace some fundamental tenets and constructs of economic theory.

To recapitulate, the proponents of NE advocate dissimilar conceptions of the theoretical presuppositions and the explanatory aims of their research. In particular, their use of the term NE is suggestive of a degree of unification and commonality of purpose that is not present in the current NE literature. In this respect, one might insist that some approaches to NE research can be consistently endorsed. However, NEs' accounts are exceedingly heterogeneous to be plausibly considered as expression of *one* and *the same* approach. Indeed, it appears that NE is currently best characterized as - not so much a single, unified discipline, but - a composite research program consisting of a cluster of approaches. In what follows, I shall use the term NE to refer to this cluster of approaches unless stated otherwise. This use of the term is sufficiently general to encompass most of the existing characterizations of NE. Moreover, it can be made sufficiently precise to enable us to assess the significance of specific contributions in NE research.

1.B GENESIS

The first publications explicitly devoted to NE appeared just a few years ago, following some pioneering studies of the neural correlates of choices (e.g. Platt and Glimcher, 1998 and 1999)[7]. The rise of NE has been favoured - and, arguably, made possible - by a series of advances in its parent disciplines. In this section, I place the emergence of NE in dialectical context, relating it to earlier developments in research at the boundary between economics and psychology. After presenting standard utility theory, which provides economists with the basic mathematical tools for modelling decisions in risky and uncertain situations, I examine some attempts to incorporate psychological insights into traditional economic models. In the next section, I shall consider some recent advances in brain-imaging and brain-stimulation technology which enabled NEs to investigate the neural substrates of choice behaviour to an unprecedented level of detail.

Expected Utility Theory (EUT) was introduced by Von Neumann and Morgenstern (1944) following previous work by Ramsey (1931) as a tractable and formally rigorous framework for modelling choices in conditions of risk (see Stigler, 1950, for a review). EUT was generalized by Savage (1954) into subjective expected utility theory, which applies to conditions of uncertainty[8]. Standard EUT rests on a representation theorem according to which, if an agent's preferences satisfy three intuitively appealing axioms, then there exists a unique probability function and a utility function U(.) unique up to

[7] See also TenHouten (1991, p.390) for an early mention of NE as "the study of the neural substrates, and associated mental phenomena, of productive and consumptive economic and socioeconomic behaviour".

[8] An agent faces a situation of *risk* when she ignores which state of affairs will occur but knows both all the potential consequences of her actions and the probability that each consequence has of occurring. An agent faces a situation of *uncertainty* when she ignores not just which state of affairs will occur, but also the probabilities that some consequences have of occurring.

positive linear transformations, such that for any two acts x and y, $x \geq y$ iff $U(x) \geq U(y)$. In this way, choice behaviour can be modelled as if the agent exhibiting it was maximizing the expected utility of her actions.

The axioms at the basis of standard EUT are the ordering axiom, the continuity axiom and the independence axiom. More specifically, the *ordering* axiom requires that agents' preferences satisfy the properties of completeness and transitivity. An agent's preferences are complete if and only if the agent is always able to express definite preferences regarding the options she faces, i.e. for any two acts x and y, $x \geq y$ V $y \geq x$. An agent has transitive preferences if and only if, for any options x, y and z, $(x \geq y \wedge y \geq z) \rightarrow x \geq z$. The *continuity* axiom demands that, if an act x is preferred to another act y but is not preferred to a third act z, then there exists a compound lottery over y and z which is indifferent to x. Formally, if $z \geq x \geq y$, then there exists $\alpha \in [0, 1]$ such that $x \sim [\alpha y; (1-\alpha)z]$. Finally, the *independence* axiom requires that adding a common component to each side of a choice relation does not change the agent's preferences. In other words, if an act x is preferred to another act y, then the compound lottery over x (with probability α) and z is preferred to the compound lottery over y (with probability α) and z, i.e. if $x \geq y$, then $[\alpha x; (1-\alpha)z] \geq [\alpha y; (1-\alpha)z] \forall \alpha \in [0,1]$.

EUT was soon accepted as part of mainstream economic theory. During the Fifties, however, its descriptive validity was called into question by a number of findings. Let us examine three clusters of anomalies in turn. A first group of findings cast doubt on the descriptive validity of the EUT axioms. For instance, contrary to the ordering axiom, agents are rarely able to express complete preferences and often exhibit intransitive

preferences (Loomes and Sugden, 1983, and Tversky, 1969). Moreover, people's preferences occasionally violate the continuity axiom and the independence axiom (see e.g. Allais, 1953, and Tversky and Thaler, 1990). These findings prompted some authors to question also the normative tenability of specific tenets of EUT. For instance, Levi (1986) argues that there is nothing inherently irrational about having incomplete preferences when one faces situations of uncertainty. For their part, Allais and Hagen (1979) provide a normative critique of independence, and Sugden (1991, sec.IV) doubts that the transitivity axiom is a necessary requirement for rationality.

A second cluster of anomalies document that - contrary to EUT - people's preferences are frequently sensitive to several factors besides the expected utility associated with the examined options. For example, agents' choices can considerably vary depending on the way preferences are elicited (see e.g. Diamond and Hausman, 1994, and Lichtenstein and Slovic, 1971) and how the description of the available options is framed (see e.g. Simonson and Tversky, 1992, and Tversky and Kahneman, 1981 and 1986).

A third series of findings point to the mistakes and biases affecting agents' probabilistic estimates. This evidence casts doubt on the EUT tenet that agents act as if they were computing the expected value of each action by impeccably applying the rules of probability calculus and Bayesian updating[9]. For instance, many people overestimate the representativeness of the sample on the basis of which they formulate their

[9] According to Bayes' Theorem, when an agent receives new information *I* concerning an event *E*, she updates her probabilistic estimates in accordance with the formula $P(E|I) = P(E)\,P(I|E) / P(I)$, where $P(E)$, $P(I|E)$ and $P(E|I)$ respectively represent the prior probability, the verisimilitude and the posterior probability of *E*.

probabilistic evaluations and arrive at different probabilistic estimates depending on the order with which they receive information (Tversky and Kahneman, 1974).

When the violations of EUT were first presented, most economists reacted by casting doubt on the reliability and the robustness of the collected evidence. The thought was that the documented anomalies were more likely to reflect peculiar features of the examined choice settings than widespread tendencies of human choice behaviour. EUT, however, came under increasing pressure when the widespread and replicable character of its failures was documented. For instance, some authors (e.g. Kahneman, Slovic and Tversky, 1982, and MacCrimmon and Larsson, 1979) illustrated the robustness of various violations of EUT to modifications in the experimental setting and in the structure of agents' incentives. Others (e.g. Bone, Hey and Suckling, 1999, Griffin and Tversky, 1992, and Wilson and LaFleur, 1995) demonstrated the persistence of several anomalies in presence of experienced agents. These findings prompted many authors to question the descriptive validity of EUT. In the words of Tversky and Kahneman, "the deviations of actual behavior from the normative model are too widespread to be ignored, too systematic to be dismissed as random error, and too fundamental to be accommodated by relaxing the normative system" (1987, p.68).

Three main responses have been developed by mainstream economists. Some attempted to show that learning and incentives tend to reduce or even eradicate the reported violations (see e.g. Chu and Chu 1990, Cox and Grether, 1996, and Smith, 1991). Others insisted that EUT offers accurate predictions of decisions made under specific conditions. For instance, the proponents of the discovered preference hypothesis (e.g.

Binmore, 1999, and Plott, 1996) alleged that standard economic theory captures only the main causal factors influencing economic behaviour and interpreted its violations as resulting from some of the omitted factors. Still others refined the axiomatic apparatus of EUT so as to reconcile such a theory with the collected findings. For instance, some authors (e.g. Aumann, 1962) relinquished completeness, others weakened (e.g. Chew, 1983, and Gul, 1991) or abandoned (e.g. Machina, 1982 and 1987) independence, still others (e.g. McClennen, 1990) relinquished both completeness and independence. Over the last decades, various modified versions of EUT have been developed along similar lines (see Starmer, 2000, for a detailed review).

The proponents of NE are frequently sceptical about the defences and the theoretical refinements put forward by mainstream economists. In particular, they criticize most modified versions of EUT for being *ad hoc* or underconstrained. Before examining NEs' instruments and studies, let us consider three lines of research at the interface between economics and other behavioural sciences which prelude NEs' contributions.

A first series of models, developed since the late Fifties, relate to the concept of bounded rationality. The idea (e.g. Simon, 1955 and 1957) is to construct more predictive economic models by taking into account psychological findings about decision making processes. The rationale in favour of this approach can be explicated as follows. Human individuals lack the cognitive and computational capacities required to make optimal decisions. In particular, they approximate the predictions of standard EUT only when their cognitive and computational limitations have a negligible impact on their behaviour. Hence, if economists are to build predictive models of choice, they

have to take into account the cognitive and computational limitations of real life economic agents. To give one example, the satisficing approach represents agents who - lacking the cognitive resources to compute optimal solutions - rest content with decisions that are sufficiently good for their purposes.

A second series of models relinquish the axiomatic foundations of EUT and define agents' utility functions in light of specific behavioural and psychological findings. By way of illustration, consider *prospect theory* (Kahneman and Tversky, 1979; see also Tversky and Kahneman, 1992, on cumulative prospect theory). In prospect theory, agents' utility depends not only on their absolute levels of consumption, but also on how close these levels are to some specified reference point (see e.g. Markowitz, 1952). More specifically, Kahneman and Tversky propose a value function which: is concave in the domain of gains and convex in the domain of losses to reflect agents' decreasing sensitivity to gains and losses; is steeper in the domain of losses to capture agents' aversion to losses; and has a flex point in the origin to reflect the fact that agents tend to be risk seeking when a prospect's outcomes are all positive and risk averse when those outcomes are all negative.

A third, more radical departure from traditional economic theory is represented by the *heuristics and biases* approach. The proponents of this approach relinquish the idea that decision making results from a unique optimizing strategy and model people's choices as the result of many context-dependent decision rules (Kahneman, Slovic and Tversky, 1982, and Tversky and Kahneman, 1974). More recently, the so-called *ecological rationality* approach has been developed, which postulates that individuals routinely

make decisions on the basis of fast and frugal heuristics (Gigerenzer et al., 1999). In this context, decisions are regarded as rational to the extent that they are well-adapted to the opportunities and risks presented by the choice situation faced by the agent (Goldstein and Gigerenzer, 2002). This view sharply differs from economists' traditional conception of rationality as conformity to context-independent rules of logic and probability theory.

Over the last decade, NEs have urged other economists to incorporate not just psychological, but also neuro-biological insights into their models of choice. The proponents of NE share with behavioural economists the ambition to broaden the range of variables included in economists' models. In particular, both NE and earlier research at the boundary between economics and psychology can be regarded as manifestations of a long-lasting trend to build more predictive models by extending the evidential base of economic theory (Payzan and Bourgeois-Gironde, 2005, p.7; see also Rustichini, 2005, p.201-4).

Having said that, it would be overly simplistic to regard NE as the mere continuation of behavioural and experimental economics with technologically more sophisticated instruments. For NEs rely on findings and methods from cognitive and computational neuroscience which transcend the reach of earlier contributions at the interface between economics and psychology (Montague, 2007a, p.219). Indeed, NEs frequently critique those contributions. For instance, Glimcher, Dorris and Bayer (2005, p.213-4) allege that bounded rationality models "have little or no predictive power outside of their bounded domains". Similarly, Glimcher (2010, p.114 and 120) contends that prospect

theory “has too many interacting parameters [for being] a truly falsifiable theory” and criticizes the heuristics and biases approach for postulating hypothetical heuristics *ad nauseam*.

To recapitulate, standard EUT enables economists to represent agents’ decisions across many choice settings in highly tractable terms. Such a theory, however, has various descriptive shortcomings, which prompted economists and other researchers to modify it in several ways. Some of these refinements (e.g. think of generalized expected utility theory) testify the flexibility of economists’ mathematical tools. Others point to more radical modifications of standard economic theory, which draw on evidence and modelling tools from other behavioural sciences. In this perspective, the anomalies and paradoxes faced by traditional economic theory can be regarded as both “signs of fundamental weaknesses” and a “fertile source of theoretical progress” (Sugden, 1992, p.x).

1.C INSTRUMENTS

Scientific revolutions often result from the introduction of innovative technologies in a particular field of research. By way of illustration, think of the impact that Galileo's telescope studies had on the controversy between the Copernican and the Ptolemaic systems (see e.g. McMullin, 2005). More recently, the development of high-energy machines has disclosed unprecedented possibilities of investigation in theoretical physics, giving new impetus to cosmological speculations and to the search for fundamental particles. The introduction of novel instruments fostered revolutionary modifications not just in physics or astronomy, but also in other natural and social disciplines. Recent developments in brain-imaging and brain-stimulation technology provide us with a striking example of technologically driven revolution in neuroscience. Let me expand on this issue.

The human brain is one of the most complex systems that have been hitherto targeted by scientists. Its staggering complexity stems from several features of the human neural architecture, ranging from the remarkable number and variety of neurons (e.g. in terms of size and shape) to the heterogeneous functional and anatomical interconnections between brain regions. By way of illustration, the cerebral cortex of an adult contains approximately 10^{11} neurons, each having up to 10^4 synaptic connections, with average neural density reaching 10^5 neurons per mm^3 in several areas (Braitenberg and Schuez, 1998, and Rockel et al., 1980).

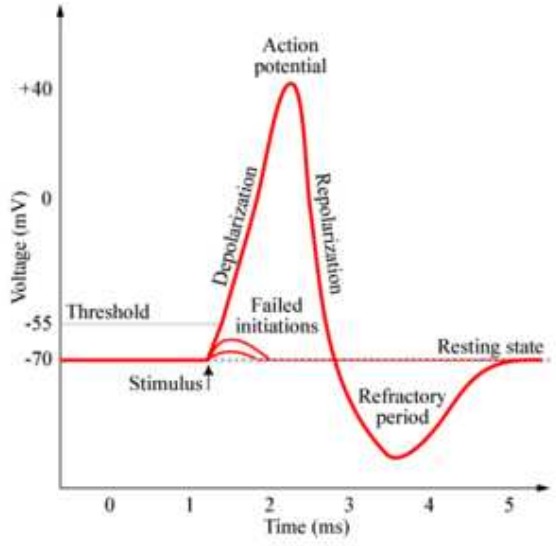
A neuron (see the figure below) typically comprises a cell body (also called soma) and an axon, i.e. a long protoplasmic fiber, covered by myelin sheath, by means of which the neuron transmits electrochemical signals to other neurons. Around the cell body, a branching dendritic tree receives signals from other neurons, while the axons' terminals release neurotransmitters towards the dendrites of the surrounding neurons. Neurotransmission involves a number of neurochemicals (such as serotonin and dopamine) which regulate the activity of various neural populations. The junctions between the axon sender of one neuron and the receiver dendrite (or the cell body) of another neuron are called synapses[10].

Source: Carlson (1992, p.36).

The process by means of which the brain collects information from the external environment can be divided into three steps, namely: transduction, which converts incoming stimuli into membrane voltages and action potentials; encoding and initial processing; and cortical processing (Glimcher, 2010, p.144). The activation process of

[10] Most synapses involve no direct contact between distinct neurons (chemical synapses). In some cases, however, there is some cytoplasmic continuity between adjacent neurons (electrical synapses). The existence of electrical synapses led some to doubt that neurons are more appropriately characterized as physically separate units (neuron doctrine) rather than as parts of a physically continuous network (reticular theory). For a discussion, see Mundale, 2001.

most neurons, which is mediated by various types of ion channels situated on their surface membrane, can be characterized as follows (see the figure below; see also Hodgkin and Huxley, 1952, for an early model). When a neuron does not send any signal, the voltage difference between the outside and the inside of its membrane (*resting potential*) is approximately -70 mV (intracellular more negative). When a neuron sends impulses through its axon, a depolarizing current occurs and the voltage difference moves towards 0 mV. When such a difference is about -55 mV, an *action potential* (also called "spike") is originated, with the voltage difference reaching up to 40 mV. The threshold at which action potentials take place is approximately the same across many neural populations, while their duration varies from 0.1 to a few milliseconds. After a spike, a neuron typically enters a period of mechanical recovery (*refractory period*), with the voltage difference between the outside and the inside of the neuron returning back towards its initial value (repolarization). At this stage, one frequently observes a hyperpolarization, i.e. a state where the voltage difference goes beyond -70 mV (Harada and Takahashi, 1983).



Source: http://soe.ucdavis.edu/ss0708/eghbalis/Notes/U12Notes.html

Image attribution: Adriana.Lico, available from: http://www.wikilectures.eu/images/thumb/7/73/Imagem1.png/608px-Imagem1.png

Before the introduction of brain-imaging and brain-stimulation instruments, neuroscientists investigated the human brain's functional organization by examining subjects who had undergone brain damaging accidents (e.g. Damasio, 1994), invasive surgical interventions (e.g. Freeman and Watts, 1942, Gazzaniga and LeDoux, 1978), and inborn or degenerative diseases (e.g. Baron-Cohen, 1995). In such a context, *lesion studies* constituted the main source of evidence until the Seventies. In a typical lesion study, one associates the cognitive or computational impairment suffered by a patient with a specific brain lesion so as to identify what contribution the damaged region provides to normal brain functioning (Glymour, 1994). Regrettably, lesion studies do not enable one to develop systematic theories of brain functioning (see section 3.C). Thanks to brain-imaging and brain-stimulation tools, neuroscientists elaborated increasingly detailed anatomical maps and distinguished various neural populations according to functional criteria (Bechtel, 2002a, and Mundale, 1998).

In the remainder of this section, I examine the brain-imaging and brain-stimulation instruments that NEs often use in their investigations. For the purpose of this enquiry, I gloss over some of the physical and chemical processes underlying the generation of the signals targeted by NEs' instruments. My characterization, however, is sufficiently accurate to enable us to assess the evidential and epistemological concerns related to the collection and the interpretation of neuro-imaging data. I shall describe the principles underlying functional magnetic resonance (fMRI) and positron emission tomography (PET) in greater detail, as NEs frequently employ these tools in studying the neural substrates of choice behaviour.

i) The *electroencephalogram* (EEG) offers a graphic representation of the electrical impulses and currents which primarily originate from the pyramidal cells in the cortex (Kutas and Dale, 1997). In a typical EEG measurement, a number of electrodes (ranging from 16 to 256) are attached to the surface of a subject's scalp. The electrodes are connected to a differential amplifier, which detects patterns of variation in voltage over time. By placing microelectrodes close to the soma or axon of a neuron, one can measure the overall synaptic activity in a targeted neural population (local field potentials, see e.g. Logothetis, 2002 and 2003). Moreover, so-called evoked response potentials (ERPs) can be obtained by presenting subjects with specific stimuli and by averaging the obtained responses over numerous EEG trials (Rugg and Coles, 1995, ch.1-2). EEG studies can provide detailed information about the timing of neural processes and require less demanding statistical analyses than fMRI or PET for interpretation (Bechtel and Stufflebeam, 2001). Unfortunately, EEG technology enables one neither to accurately locate the spatial origin of the registered signals nor to infer the functional roles played by the examined neural areas. For these reasons, EEGs are employed most fruitfully in combination with other brain-imaging and brain-stimulation instruments[11].

ii) The *positron emission tomography* (PET) enables one to estimate neural areas' activation patterns by measuring the quantity of blood entering them (regional cerebral blood flow). Before undergoing a PET investigation, the subject is injected or inhalates a limited amount of radiolabeled compound (e.g. labeled glucose) having a short half-

[11] Similar remarks apply to magnetoencephalography (MEG), a non-invasive technique which measures the magnetic fields generated by small intracellular electrical currents in the brain (Cohen and Halgren, 2004, and Hansen et al., 2010).

life. The radioactive atoms in the compound quickly decay, releasing positrons. Each of these positrons collides with a nearby electron and annihilates with it. As a result, two photons are emitted, which travel off the annihilation point in approximately opposite directions along a line with random orientation (Vardi, Shepp and Kaufman, 1985, p.8; see also Ter-Pogossian et al., 1980).

As shown by the figure below, only some of those photons are detected by the detector ring, a cylindrical volume of detectors surrounding the head of the investigated subject. For instance, while the annihilation at point *x* is detected (locations D3 and D67), the one at point *y* goes unnoticed because the photons' trajectory does not intersect the detector ring. On the basis of the observed detections, researchers estimate the concentration of the injected compound in different neural areas. This estimate, in turn, is interpreted in light of sophisticated computational models to yield graphic representations of neuro-physiological activity (Ollinger, 1994, and Ollinger and Fessler, 1997).

Source: Vardi, Shepp and Kaufman (1985, p.9)

iii) The *functional magnetic resonance* (fMRI) is an application of the magnetic resonance (MR) technology previously used in clinical research. In order to generate

MR images, one needs both a contrast mechanism for estimating brain regions' activations and a reliable criterion for interpreting the observed signals. MR can measure neural activation patterns via blood-volume changes (e.g. Belliveau et al., 1991), changes in blood oxygenation (e.g. Ogawa and Lee, 1990), tissue perfusion (e.g. Detre et al., 1992) and even water diffusion (e.g. Le Bihan, 2007). Most fMRI studies, however, focus on variations in blood oxygenation levels. These haemodynamic variations are the source of the so-called BOLD (blood oxygenation level dependent) signal (Ogawa et al., 1990, and Toga and Mazziotta, 2002, ch.13).

fMRI has higher temporal and spatial resolution than PET (Dobbs, 2005). Its functioning can be characterized as follows (Buxton, 2002, and Jezzard et al., 2002). The researcher applies a strong magnetic field to the patient's skull to make the nuclei of atoms having odd atomic weight align the axes of their spin. A brief pulse of radiowaves is then used to perturb this alignment. When the pulse ends, the nuclei tend to go back to their aligned state, releasing radio waves whose frequency reflects the features of the targeted atoms (Bechtel, forthcoming)[12].

So far, fMRI investigations of decision making processes have predominantly focused on individual choices and two-agent interchanges, with very few authors investigating multi-agent interactions. In recent years, however, valuable insights concerning the neural underpinnings of collective decisions and social interactions have been acquired (see e.g. Berns et al., 2005). In this respect, it is worth mentioning the so-called

[12] MR images can be acquired after either multiple excitation pulses or a single pulse (echo planar imaging). The quality of MR images is affected both by the time between the excitation pulse and the recovered signal (echo time) and by the time between successive pulse sequences (repetition time). I do not expand on these details for the purpose of this enquiry.

hyperscanning, which allows one to simultaneously monitor neural activations in multiple subjects, each in a separate MRI scanner (see Montague et al., 2002, for the first application; see King-Casas et al., 2005, and Tomlin et al., 2006, for some hyperscan studies of repeated trust games).

iv) The *transcranial magnetic stimulation* (TMS) consists in stimulating or deactivating specific cerebral regions by means of impulses of magnetic energy. More specifically, one applies a time-varying magnetic field and examines the modifications that are subsequently induced in the targeted neural areas (Hanks, Ditterich, and Shadlen, 2006). TMS can provide informative evidence about the functional role played by specific neural regions by selectively and reversibly perturbing their activation patterns (Bechtel and Richardson, 2010, p.256-7). TMS investigations can effectively complement fMRI studies, which often report the activation of areas whose operations are incidental to the execution of the examined tasks. Nonetheless, the reliability of TMS findings is constrained by our limited understanding of how TMS affects overall brain activity. For example, disrupting activation in the areas engaged by a particular task may stimulate activation in other regions whose operations interfere with those of the targeted areas in ways that elude experimental controls[13].

v) *Single neuron measurements* are performed by inserting microscopic electrodes directly in the brain and measure variations in individual neurons' voltage. Despite having a commendably high spatial and temporal resolution, single neuron

[13] The activation patterns of specific neural areas can be altered also by means of *electric stimulations* (see Olds and Milner, 1954, and Mendelson, 1967, for early applications). Yet, the electric fields generated through electric stimulation are less focal and more sensitive to skull conductivity than the ones generated via TMS (Saypol et al., 1991).

measurements often disrupt activity in the investigated areas and are generally applied to non-human species due to ethical considerations. This significantly constrains their usefulness for the study of higher cognitive and computational functions. Moreover, single neuron measurements enable one to monitor the activity of only few neurons at once (Ludvig et al., 2001). Given that performing a task usually activates several functionally interconnected areas (see section 3.C), monitoring single neurons rarely provides informative insights regarding the functional organization of wider brain regions.

To recapitulate, each brain-imaging and brain-stimulation instrument is characterized by specific strengths and limitations. By combining different instruments, NEs can acquire *complementary* information about the examined phenomena and provide *independent* evidential support to the findings obtained with a specific technique (Camerer, 2008a, and Logothetis, 2008a). To be sure, some brain-imaging and brain-stimulation instruments are prone to common biases and experimental confounds (see section 3.B). Still, obtaining convergent results by means of techniques involving independent auxiliary assumptions increases our confidence in each technique, as it is unlikely that those techniques produce similar findings by chance (Bechtel, 2002a, p.S48)[14].

[14] A related debate has taken place in the philosophy of science regarding the so-called derivational robustness of models, i.e. the independence of a model's implications from the assumptions used to derive them (Woodward, 2006; see also Levins, 1968). The following contrast is prominent in the literature. On the one hand, some authors (e.g. Kuorikoski, Lehtinen and Marchionni, 2010) argue that robustness analysis can be employed as a method for confirmation of specific claims about causal mechanisms. On the other hand, others (e.g. Weisberg, 2006, and Odenbaugh and Alexandrova, 2011) allege that robustness analysis is best regarded as a method of discovery rather than confirmation of hypotheses.

Besides combining different brain-imaging and brain-stimulation instruments, NEs often attempt to corroborate neural findings by *triangulating* neural, psychological and choice data (see e.g. Houser et al., 2007, and Wilkinson and Halligan, 2004; see also Rubinstein, 2007 and 2008, on response times). In particular, some NEs employ pharmaceuticals and hormones to investigate the causal underpinnings of choice behaviour (see e.g. Bielsky et al., 2005, on the influence of vasopressin on social recognition). The idea is to identify how people's behaviour varies in response to diverse perturbations of the neuro-physiological processes associated with decision making (see e.g. Baumgartner et al., 2008, Kosfeld et al., 2005, and Zak et al., 2004, 2005 and 2007, on the role oxytocin plays in promoting interpersonal trust).

CHAPTER TWO

THE CASE IN FAVOUR OF NEUROECONOMICS

The proponents of NE often employ the following two-stage argument in advocating the neural enrichment of economic theory. In the first place, they cast doubt on the descriptive and normative validity of the traditional economic account of decision making. For instance, it is often argued - in line with some economists (e.g. Rustichini, 2005, p.202, and Schotter, 2008, p.71-2) and philosophers (e.g. Hausman, 2008a, p.130-9, and Sugden, 1991, sec.I-IV) - that rational choice theory faces frequent, statistically significant and robust descriptive failures; that axiomatic approaches typically fail to ground an informative account of economic behaviour; and that an exclusive reliance on observed choice data would constitute a severe limitation for the economic theory of choice. In the second place, NEs highlight various respects in which neuro-psychological findings may inform the economic account of decision making.

In this chapter, I identify several respects in which neural data have been claimed to be importable fruitfully into economic models of choice. I shall focus in turn on descriptive accuracy, predictive power, model selection, explanatory insightfulness, and applicability to welfare analyses. For each of these desiderata, I reconstruct an argument in favour of a neural enrichment of economic models which builds on the assertions that NEs put forward in the literature[15]. In doing so, I provide various considerations concerning the pragmatics of economic modelling and examine some of the findings that NEs deem to be of great significance for other economists. Moreover,

[15] These arguments are rarely given explicit formalization by NEs. Still, I take my reconstructions to provide a faithful characterization of NEs' positions.

I illustrate how the NEs' assertions can be combined in a cumulative case for the neural enrichment of economic theory. Before proceeding, let me provide three preliminary caveats regarding the contents of this chapter.

Firstly, scientific models are constructed and applied for a variety of purposes, which range from making precise predictions to providing explanatory insights regarding the investigated phenomena. The list of attributes I examine does not exhaust the set of modelling desiderata valued by economists, yet provides us with an informative basis for assessing the merits of NEs' calls to include neural insights into economic models. To be sure, each of these desiderata has been defined in dissimilar ways in the literature. As I aim to illustrate below, however, we can provide instructive insights into model selection and model evaluation in economics by adopting a sufficiently general and uncontroversial characterization of each desideratum[16].

My second caveat relates to the notion of model. A variety of things - ranging from physical objects (e.g. Black, 1962) to descriptions (e.g. Achinstein, 1968) and set theoretic structures (e.g. Suppes, 1960) - can serve as models. In what follows, I focus on mathematical models of decision making unless stated otherwise, without expanding on further distinctions between conceptions of economic models[17]. Also, I gloss over

[16] Significant interrelations exist between distinct modelling desiderata. For instance, descriptive accuracy is occasionally valued because of the predictive and explanatory gains it yields to modellers (see e.g. Camerer, 2007, C28). I shall expand on the interrelations between different modelling desiderata in *chapter four*.

[17] See e.g. Lucas (1980) and Sugden (2000) on economic models as counterfactual or artificial worlds that modellers envision to investigate the properties of their target systems; McCloskey (1990) and Morgan (2001 and 2002) on models as a combination of an uninterpreted mathematical structure and a corresponding narrative interpretation; Gibbard and Varian (1978) and McCloskey (1983) on models as metaphors or caricatures intended to exaggerate or isolate specific features of the examined phenomena.

the various ways in which the relationship between models and their target systems has been characterized in the literature on scientific modelling (see e.g. van Fraassen, 1980, on isomorphism, and da Costa and French, 2003, on partial isomorphism; see also Giere, 1988, on different resemblance relations).

My third caveat concerns the relationship between economic models and economic theory. The proponents of NE advocate the integration of neural insights in relation to both economic models and economic theory. The notions of model and theory have been characterized in several ways in the economic literature (see e.g. Mäki, 1993 and 1996, Guala, 1998 and 2002, and Morrison and Morgan, 1999, ch.1-3), with significant distinctions being made between them. The following divide is particularly prominent in the literature on scientific modelling. On the one hand, the syntactic view (see e.g. Carnap, 1938, Braithwaite, 1953, and Nagel, 1961) defines theories as sets of sentences in an axiomatized system, with models providing an interpretation which relates the formal theory to the objects under investigation. On the other hand, the semantic view (see e.g. Suppes, 1967, Suppe, 1977 and 1989, and Giere, 1988) regards theories as collections of models rather than sets of axiomatic sentences. On this account, a theory is defined in terms of "the class of its models directly, without paying any attention to the questions of axiomatizability, in any special language" (van Fraassen, 2000, p.179)[18].

[18] In the words of van Fraassen (1980, p.44): "The syntactic picture of a theory identifies it with a body of theorems, stated in one particular language chosen for the expression of that theory. [In the semantic] approach the language used to express the theory is neither basic nor unique; the same class of structures could well be described in radically different ways".

In what follows, I speak of a neural enrichment of economic models and economic theory interchangeably, as the cogency of my considerations does not hinge on the difference between these two notions. *Prima facie*, my doing so might seem to obscure a distinction I proposed in section 1.A, according to which incremental and radical NEs respectively aim to modify economic models and economic theory. This, however, is not necessarily the case, as incremental NEs are concerned with *specific* models rather than entire *classes* of models (or theories). More generally, the point remains that incremental and radical NEs respectively attempt to modify standard economic theory to a significantly different extent (see section 1.A).

2.A DESCRIPTIVE ACCURACY

Economic modellers often aim to provide descriptively accurate representations of the features or the behaviour of their target systems. A model's descriptive accuracy can be evaluated along different dimensions (see e.g. Nagel, 1963, and Musgrave, 1981; see also Mäki, 1992, p.329, for a distinction between various respects in which the realisticness of economists' modelling assumptions can be appraised). To render this point more vivid, let us distinguish the two following senses of "descriptive accuracy". On the one hand, there is the question whether a model includes all the relevant properties or traits of the phenomena of interest. On the other hand, one can assess how accurate the model's characterization of each of those properties is (see Weisberg, 2007a, for a similar distinction). Now, a model can be regarded as more or less descriptively accurate depending on which of these two senses of "descriptive accuracy" is considered. For instance, one model may capture most of the properties of its target system but fail to characterize these properties accurately. Another model, instead, may include just a small subset of the properties of the examined phenomena yet offer an accurate characterization of those properties[19].

NEs refer to both of these senses of "descriptive accuracy" when they advocate the incorporation of neural insights into economic models. More specifically, their *argument from descriptive accuracy* can be characterized as follows:

[19] Various authors speak of "realism" instead of "descriptive accuracy". For his part, Mäki (1988; see also his 1992, 2001, and 2007) advocates distinguishing the terms "realism" - which relates to ontological and semantic doctrines in philosophy - and "realisticness" - which designates specific attributes of scientific representations.

P.1 Standard economic theory fails to provide accurate characterizations of the neuro-psychological substrates of people's decisions.

P.2 By incorporating NE insights, economists could provide more accurate characterizations of the neuro-psychological substrates of people's decisions.

P.3 Standard economic theory posits agents having implausible cognitive and computational abilities.

P.4 By incorporating NE insights, economists could construct models which posit agents having more plausible cognitive and computational abilities.

C Economists should incorporate NE insights into their models of choice[20].

Let us examine the various steps of this argument in turn. *Premise 1* asserts that the economic theory of choice fails to provide an accurate characterization of the neuro-psychological substrates of people's decisions. For instance, some NEs allege that while economists consider agents' choices as primitives, these choices result from more fundamental neuro-psychological processes (see e.g. Glimcher, 2003 and 2010). Others contend that although economists regard people's decisions as the product of conscious deliberation alone, human behaviour is shaped by the "fluid interaction between controlled and automatic processes, and between cognitive and affective systems" (Camerer, Loewenstein and Prelec, 2005, p.11). The idea is that standard economic theory focuses exclusively on people's rational motives, thereby failing to reflect the

[20] The conclusion of this argument contains the prescriptive term "should" even though it is derived from descriptive premises. This is because I suppressed for expository simplicity both an intermediate conclusion stating that "economists have some reasons to incorporate NE insights into their models of choice" and an implicit premise claiming that "these reasons license the claim that economists should incorporate NE insights into their models of choice". Analogous remarks apply to the arguments in the sections below.

complexity of decision making processes (Payzan and Bourgeois-Gironde, 2005, p.12; see also Camerer, 2007).

Premise 2 states that incorporating neural data enables economists to provide more descriptively accurate characterizations of the neuro-psychological processes underlying observed decisions. The thought is that brain-imaging and brain-stimulation techniques allow us to disclose the workings of the human neural architecture and measure variables that economists previously regarded as unobservable. In the words of Camerer, NE "is not in opposition to rational choice theory, but sees potential in extending its scope by observing variables that are considered inherently unobservable in [it]" (2008a, p.45). To be sure, many NEs acknowledge that economists usually have to rely on abstractions, isolations and idealizations in order to construct informative models[21]. At the same time, they question the legitimacy of severe simplifications in economic modelling and doubt that economists can remain agnostic regarding the neuro-psychological substrates of choice. Indeed, some (e.g. Camerer, 1998, p.177, and Glimcher, 2010, p.126 and 133) explicitly urge economists to complement or relinquish their as if representations in favour of mechanistically informed models (I shall expand on this issue in sections 4.A and 7.D).

Premise 3, in turn, asserts that economic models posit agents having implausibly sophisticated cognitive and computational abilities. The idea is that the perfect calculators with complete preferences and prodigious reasoning skills figuring in the

[21] The idea is that "because of the high premium economics places on the […] quantification of evidence, attending to all facets of human nature is neither feasible nor desirable" (Rabin, 1998, p.13). See e.g. Cartwright, 1994, and Jones, 2005, on abstractions; Mäki, 1992 and 2009, and Sugden, 2000, on isolations; Cartwright, 1983, McMullin, 1985, and Weisberg, 2007a, on idealizations).

economists' models do not remotely resemble real life human individuals. This lack of descriptive accuracy, in turn, is said to negatively affect both the explanatory and the predictive performance of economic models. To give one example, rational choice models often assume that a greater availability of options is better for agents. This, however, does not fit well with the evidence collected in many studies (see e.g. Sarver, 2008, on how people often take decisions they later regret when more options are available).

Finally, *premise 4* alleges that NEs' investigations provide economists with informative insights regarding agents' cognitive and computational abilities. The thought is that since agents' choices result from underlying neuro-psychological mechanisms, acquiring a better understanding of those mechanisms constrains economists' conjectures regarding agents' cognitive and computational endowment. On the basis of these premises, the argument concludes that economists should include neuro-psychological variables and findings into their models of decision making. In the words of Camerer (2007, C35), "the largest payoff from [NE] will not come from finding rational-choice processes in the brain for complex economic decisions […] The largest innovation may come from pointing to biological variables which have a large influence on behaviour and are underweighted or ignored in standard theory" (see also Camerer, Loewenstein and Prelec, 2005, p.10, and Park and Zak, 2007, p.54).

In rebuttal to the aforementioned claims, an economist may allege that economic models of choice are meant to describe neither the neural substrates of decisions nor the cognitive abilities of real life individuals. Indeed, some authors question the very idea

that traditional economic theory requires us to identify the modelled agents with individual people. As Ross puts it, “from the empirical point of view” the agents modelled by economists are “merely sites of consumption; there is no reason at all to assume they're people, rather than firms or countries or pension funds. [They] are 'representative' optimizers whose ontological status is indeterminate” (2008b, p.130; see also Kacelnik, 2006, and Ross, 2002). To be sure, Ross (2008c, p.738) concedes that economic modellers are typically concerned with “individual optimizers”. Still, he insists that the idea that the paradigmatic model of an economic agent is an individual human being is “in no way part of or implied by the mathematics” of standard economic theory (Ross et al., 2008, p.viii; see also Ross, 2005).

Now, it is true that many economic models can be applied without taking a position regarding what entities the posited agents map onto. Yet, the point remains that most economic models are meant to target the choice problems faced by real life individuals. Furthermore, the vast majority of economists implicitly assume that real life individuals constitute the paradigm case of agents. Indeed, it is hard to see how economists could assess the merits of their own models if they remained agnostic about what sort of entities those models are meant to represent. That is to say, abandoning the idea that “the paradigmatic economic agent is a whole adult person” may suffice to block the NE critique of economic theory (Ross, 2010, p.639). Still, few economists would be willing to relinquish this assumption. For such a presupposition plays a fundamental role in both model construction and model evaluation in economics.

2.B PREDICTIVE POWER

The predictive performance of an economic model can be evaluated in several respects. By way of illustration, let us distinguish between predictive *accuracy*, which refers to the exactness of a model's observable implications regarding the examined phenomena, and predictive *robustness*, which relates to the stability of a model's predictive performance across distinct choice contexts. These two notions point to different aspects of a model's predictive performance. For instance, some models may enable one to formulate very accurate predictions, but only in specific decision settings. Other models, instead, may allow one to make just approximate predictions across a wide range of choice situations. Moreover, both predictive accuracy and predictive robustness can be evaluated along different dimensions. For example, one may deem a model's predictions to be more or less accurate, depending on whether she considers the magnitude of the effects anticipated by such model, the timing at which those effects are predicted to occur, and so on (Moscati, 2006).

The proponents of NE often criticize economic models of choice by emphasizing their predictive shortcomings (e.g. Loewenstein et al., 2008, p.651, and Rustichini, 2005, p.202). Moreover, they argue that economists could obtain substantial predictive gains by incorporating neuro-anatomical and neuro-physiological insights. For instance, Camerer alleges that NEs' investigations identify neuro-psychological variables that predict variations in people's choices (2008a, p.46; see also Camerer, 2008b, p.370). Similarly, McCabe asserts that "by moving the study of decision making a step down, to the brain, neuroeconomics will […] make economic theory more predictive" (2008,

p.348). For their part, Glimcher, Dorris and Bayer (2005, p.221) go as far as to maintain that by revealing the neuro-biological mechanisms underlying people's decisions, NEs will build "a mechanistically accurate economic theory which is by necessity predictive".

NEs have identified several respects in which a neural enrichment of economic models purportedly brings predictive benefits to economists. I am not concerned here with providing a comprehensive review of NEs' appeals to predictive considerations. For the purpose of this section, I shall focus on the following *argument from predictive gains*:

P.1 Neuro-physiological insights enable economists to predict variations in agents' behaviour.

P.2 Neuro-physiological insights enable economists to improve their out of sample predictions.

P.3 Neuro-physiological insights enable economists to identify missing economic variables.

C Economists should incorporate neuro-physiological insights into their models of choice.

Let us consider the various premises of this argument in turn. According to *premise 1*, NEs' contributions help economists to better account for the heterogeneity in agents' choice behaviour. In the words of Camerer, NE research enables economists to "ground economic theory in detailed neural mechanisms which are expressed mathematically and make behavioural predictions" (2007, C26; see also Bernheim, 2009, and Rustichini, 2009).

To be sure, if NEs are to substantiate their calls for a neural enrichment of economic theory, their rationale for incorporating neural insights must go beyond the econometric platitude that when one increases the set of explanatory variables, the correlation between them and the dependent variables is almost bounded to increase[22]. Fortunately, various authors have provided persuasive success stories showing the predictive gains yielded by a neural enrichment of specific models of choice. For example, some (e.g. Hsu et al., 2005, and Kuhnen and Knutson, 2005) show that differential activation in some neural areas can be used to predict agents' behaviour in conditions of risk and ambiguity. Others (e.g. Knutson et al., 2007) illustrate that the activation patterns of particular areas can constitute a better predictor of choices than people's expressed preferences (see also the pioneering studies of Libet, 1965, and Libet et al., 1979). Still others (e.g. McCabe et al., 2001) predict subjects' propensity to trust other players in game theoretic settings by monitoring the activations of specific areas.

The reasoning underlying *premise 2* can be explicated as follows. Economists often need to predict choice behaviour in novel decision contexts. Now, one might succeed in providing *accurate* predictions even without having detailed knowledge of the mechanistic underpinnings of the modelled phenomena (Woodward, 2003, p.232-3). Yet, models which do not take into account the mechanistic underpinnings of their target systems often fail to provide *reliable* and *robust* out of sample predictions (Bechtel and Richardson, 1993, and Craver, 2007; for some definitions of mechanisms,

[22] To give one example, the fact that morning sunshine is significantly correlated with stock returns in several countries (see e.g. Hirshleifer and Shumway, 2003) falls short of implying that economists should import complex meteorological insights in their models of the financial market.

see Glennan, 2005, p.445, and Machamer et al., 2000, p.3). This remark is especially relevant for models examining choice contexts that are subject to rapid and profound modifications. In this respect, it would be of little import to contend that economists could formulate accurate out of sample predictions without importing neuro-physiological insights. For these insights can help economists to ascertain not just *whether* observed choices are consistent with their theories, but also *why* this is the case so as to employ their models to different decision problems (Schotter, 2008, p.79).

NEs' calls to include neuro-physiological insights into economic models can be seen as an instance of a more general position, according to which economic models should capture the influence of the causal factors underlying the modelled phenomena (Mill, 1844, ch.1; see also Cartwright, 1998, 1999 and 2007). The idea is that a model's predictions are reliable to the extent that the causal factors it posits resemble those operating in the examined target systems (Pemberton, 2005). Analogous claims have been put forward by various proponents of NEs. For instance, Fehr and Camerer (2007) hold that identifying the neural substrates of observed choices in specific experiments helps economists to better predict people's decisions in other experimental settings. Similarly, Rustichini asserts that investigating the algorithmic underpinnings of observed choices enables NEs to make more accurate predictions both across different decision settings and for more extended timescales (2009, p.50).

Premise 3 points to a third respect in which neuro-physiological findings are said to yield predictive benefits to economists, namely the identification of missing economic variables (see e.g. Bernheim, 2009). To render this point more vivid, consider the

following example. People's risk preferences cannot always be inferred univocally from their observed choices. Now, suppose that the activation patterns of some neural areas provide informative insights regarding agents' risk preferences and that these activation patterns are accurately observable. Economists could employ observed neural activations as proxies for agents' risk preferences when these preferences cannot be inferred precisely from the available choice data. Neural data could arguably be used to predict also hypothetical and counterfactual choices. To be sure, these predictions are complicated by the fact that several areas (e.g. those related to imaginative faculties) are likely to exhibit differential activations depending on whether agents envision actual, hypothetical or counterfactual decisions. Yet, in spite of these complications, NEs might succeed in building more predictive models by observing the activation patterns that specific areas exhibit while people contemplate hypothetical and counterfactual choice problems (Bernheim, 2009, p.14).

On the basis of the previous premises, the argument from predictive gains *concludes* that economists should incorporate neuro-physiological insights into their models of choice. To better appreciate the rationale for this claim, let us consider one illustration of how including neural insights can increase the predictive power of a standard economic model. Economic agents frequently exhibit a significant degree of prosocial behaviour in game theoretic settings. Consider, for example, one-shot ultimatum games. In a typical ultimatum game, player 1 chooses how to split a given amount M of benefits between herself (b_1) and player 2 (b_2). Player 2, in turn, can either accept or reject the offer of player 1. If she rejects it, the benefits are lost to both players. Game theoretic reasoning predicts that player 2 should accept any positive transfer and that

player 1, anticipating this, should offer the smallest positive amount to her. Yet, subjects typically transfer fairly high amounts of benefits (see e.g. Guth et al., 1982).

In a recent article, Vercoe and Zak (2010) propose a neurally informed model to account for why many agents offer resources to strangers in one-shot ultimatum games where such behaviour does not serve to build their own reputation as altruists. In this model, player 1 derives utility not just from keeping some benefits b_1, but also from offering benefits b_2 to player 2, provided that $b_2 > b^*_2$ (i.e. her offer to player 2 is sufficiently high to be accepted) and $\alpha(\tau) > 0$. The term $\alpha(\tau)$ represents an empathy function, with τ measuring the distress of player 2 perceived by player 1, and constitutes an innovation with respect to standard models of ultimatum games. To see this, let us focus on player 1. Her decision problem can be formalized as follows:

$$\text{Max}_{b1,b2}\ U(b_1) + \alpha(\tau)\, U(b_2)$$

$$\text{s.t.}\quad b_1 + b_2 = M, \quad b_2 > b^*_2$$

While moderate levels of observed distress increase empathy and assistance to others, high levels of perceived distress cause avoidance and inhibit oxytocin release (see e.g. Barraza and Zak, 2009). Moreover, various NE studies (e.g. Zak et al., 2004, and Kosfeld et al., 2005) associated increases in oxytocin with higher levels of empathy and trust (amount of resources an agent transfers in particular games). In light of these findings, Vercoe and Zak (2010) take $\alpha(\tau)$ to capture the effects of oxytocin release, and characterize it as a continuous hyperbolic function with domain and range [0,1] such that $\alpha(0) > 0$, $\alpha(1) = 0$ and $\alpha(\tau^*) > \alpha(0)$, with $\tau^* = \text{argmax}\ \alpha(\tau)$. Their model

predicts that as $\alpha < \alpha(\tau^*)$ rises, the benefits that player 1 offers to player 2 increase. By physiologically manipulating empathy (infusions of oxytocin and other hormones), Vercoe and Zak provide support to the predictions of their model in various experimental settings.

2.C MODEL SELECTION

Economic modellers frequently value the possibility of deriving testable implications from their models. However, the observable implications of economic models of choice are typically conditional upon a variety of *ceteris paribus* qualifications and other auxiliary assumptions. Indeed, the very act of deriving testable implications from economic models often requires one to introduce subsidiary hypotheses about test conditions (Eichner, 1983). For these reasons, empirical and experimental findings contrary to a model's implications can rarely be regarded as direct evidence against the model itself as opposed to some of its auxiliary assumptions (Hausman, 1992, p.207; see also Duhem, 1906, and Quine, 1953, for analogous remarks regarding scientific theories). As Machlup (1955, p.19) puts it: "When the economist's prediction is […] based upon specified conditions, but where it is not possible to check the fulfilment of all the conditions stipulated, the underlying theory cannot be disconfirmed whatever the outcome observed […] our tests cannot be convincing enough to compel acceptance" (see e.g. Blaug, 1992, Hands, 1985a, Hutchison, 1978, and McCloskey, 1983, for a debate).

The proponents of NE often criticize economists for relying on "a plethora of competing models that are either not tested, or if tested often explain the data equally well" (Vercoe and Zak, 2010, p.133). Moreover, they urge economists to select and assess their models in light of neuro-cognitive and neuro-computational evidence. Their *argument from model selection* can be reconstructed as follows:

P.1 Observed choice behaviour can be accounted for in terms of several competing economic models.

P.2 Economists are often unable to discriminate effectively between competing economic models of choice.

P.3 Neuro-physiological evidence enables economists to better discriminate between competing economic models of choice.

C Economists should rely on neuro-physiological evidence to discriminate between their models of choice[23].

The proponents of NE frequently argue along these lines in advocating the use of neuro-physiological data for model selection purposes in economics. For example, after contending that economists often provide various axiomatic systems consistent with observed choices, Camerer (2008a, p.47) alleges that "neural tests could winnow a crowded field of possible theories down to the more plausible theories". Similarly, Rustichini (2009, p.58) complains that economists lack effective strategies for model selection and asserts that a "fundamental role" of NE research consists in "pruning the multiplicity of models, and to make them closer to the hard experimental test" (see also Glimcher, 2010, p.396)[24].

[23] Discriminating between distinct models of the same target system is most plausibly conceptualized as not so much an all-or-nothing affair, but rather a matter of nuanced evaluation. I am not concerned here with providing a precise quantitative measure of the effectiveness with which we can discriminate between competing models of choice. An intuitive measure relates to a modeller's ability to identify a narrow set of plausible models out of many available candidates.

[24] Neural insights have been claimed to inform model selection and theory choice in other disciplines besides economics. For instance, Greene et al. (2001; see also Greene, 2007) take some brain-imaging findings about the neural correlates of moral deliberation to discriminate between deontological and consequentialist moral theories (see Berker, 2009, for criticisms). Similar debates took place in the literature at the interface between psychology and neuroscience. For example, some authors (e.g. Henson, 2005) allege that functional

To render these remarks more vivid, let us consider one example of how NE findings can inform model selection in economics. As we have seen in section 1.B, prospect theory accounts for various violations of traditional EUT by assuming that agents value options with respect to specific reference points. Now, reference points can be determined in terms of several variables, including the level of consumption to which one got used in the past, one's expectations concerning her future levels of wealth, how one's resources compare with those of other agents, etc. Regrettably, prospect theory does not specify how reference points are to be determined, how they vary across choice settings, and what reasons we have to believe that agents value options in reference-based terms.

Recent NE research provides some intriguing insights in this latter respect. For example, Glimcher (2010, p.274-8) argues that a reference-based encoding of external stimuli and choice options is metabolically more efficient (e.g. in terms of required energy consumption) than objective encoding. In his view, current neurobiology speaks against the reference-point independent valuation systems implicitly posited by standard EUT. These remarks arguably provide economists with an additional reason besides predictive considerations to favour prospect theory over traditional EUT models.

A defender of standard economic theory might oppose the use of neural data for model selection purposes in economics on the ground that economic theory does not make explicit assumptions about the inner workings of the human neural architecture. After

---

neuroimaging data enable us to discriminate between alternative psychological theories. Others insist that neuroimaging data do not (e.g. Coltheart, 2004 and 2005, and Harley, 2004) or even cannot (e.g. van Orden and Paap, 1997, and Uttal, 2001) provide decisive evidence concerning the merits of psychological theories.

all - the thought would be - why should economic models be evaluated in light of findings coming from disciplines having dissimilar methodological presuppositions and explanatory aims than economics? In rebuttal to this claim, various authors (e.g. Quartz, 2008) doubt that economists can remain entirely agnostic concerning the neuro-physiological underpinnings of choice behaviour. The idea is that even though economic models make no explicit assumptions regarding the inner workings of the human neural architecture, they do make predictions that can be either confirmed or disconfirmed by neuro-physiological evidence. Hence - the reasoning goes - neuro-physiological data should be used to constrain and discriminate between alternative economic models.

To be sure, NE models of choice are constrained not just by bottom-up neuroscientific findings, but also by top-down behavioural evidence. By way of illustration, consider the pioneering studies of Dorris and Glimcher (2004) on the neural substrates of reward evaluation. After observing that the firing rate of monkeys' LIP neurons correlates with the relative expected desirability of saccadic eye movements, Dorris and Glimcher allege that "the average firing rates of these neurons may also encode the subjective desirability of actions in humans" (2004, p.376). Let us suppose that this was actually the case. As noted by Glimcher (2010, p.234-6), it would be mistaken to infer from it that the human brain encodes *only* relative expected subjective values. For any organism whose valuation system stored only these values would make intransitive choices much more often than humans are observed to do. In this respect, observed behavioural patterns constrain NE models by suggesting that distinct kinds of desirability signals are presumably encoded in the human brain.

2.D EXPLANATORY INSIGHTFULNESS

Economic modellers often aim to not just describe or predict people's decisions, but also explain them. Now, the mere fact that some economic models are predictive or descriptively accurate does not imply that they are also explanatory for economists. In the philosophy of science, dissimilar accounts of explanation have been advocated. For instance, some authors (e.g. Hempel, 1962 and 1965) conceive of explanation as the deductive derivation of a sentence describing the explanandum from a set of premises containing at least one natural law. Others (e.g. Salmon, 1971 and 1984) characterize explanation as the identification of statistical relevance relationships between the explanans and the explanandum. Still others (e.g. Craver, 2006 and 2007) relate explanation to the uncovering of the mechanistic underpinnings of the investigated target systems[25].

The proponents of NE frequently contend that, despite being at a relatively early stage of development, their discipline provides more *explanatorily informative* accounts of choice behaviour than standard economic theory. For instance, Zak (2004, p.1738) boldly asserts that NE research "will allow economists to answer fundamental questions they are unable to address". Similarly, Brocas and Carrillo allege that NE provides "new reliable theories capable of explaining […] individual behaviour and strategic choices" (2010). In spite of NEs' contentions, many economists doubt the explanatory relevance of neuro-physiological findings for the economic theory of choice. As Rustichini puts it:

[25] Each of these accounts of explanation faces objections. For instance, consider Hempel's deductive nomological model. *Pace* Hempel, many generalizations appear to be explanatory even though they fail to satisfy his model's conditions, and many derivations that are intuitively non-explanatory meet the conditions of the model (see e.g. Kitcher, 1981).

"a common criticism that is raised to neuroeconomics is the following. With this method we now know, for example, the specific regions in the brain that are active when some behavior is observed. This information may be very interesting for a neuroscientist […] but what does it add to the understanding of economic behavior?" (2005, p.201).

NEs claim that their neuro-physiological findings provide economists with explanatorily informative insights in several respects. Their assertions can be combined into the following *argument from explanatory insightfulness*:

P.1 NEs' insights help economists to explain the variability in human choice behaviour.

P.2 NEs' insights help economists to provide singular explanations of economic phenomena.

P.3 NEs' insights help economists to account for some anomalies of standard economic theory.

C Economists should incorporate NEs' insights into their models.

Let us consider the various steps of this argument in turn. *Premise 1* asserts that NEs' findings enable economists to better account for how people behave in different choice situations. This claim fits well with the unificationist model of explanation (e.g. Feigl, 1970, Friedman, 1974, and Kitcher, 1981 and 1989), according to which scientific progress consists in disclosing connections between facts and phenomena previously

regarded as unrelated[26]. In the NE literature, various authors maintain that more accurate knowledge of the human neural architecture helps economists to better account for the interpersonal and intrapersonal variability of people's decisions (McCabe, 2003b, p.294; see also Craver and Alexandrova, 2008, p.396). Two kinds of NE contributions can be distinguished in this respect. On the one hand, some authors (e.g. van't Wout et al., 2005 and 2006) illustrate how agents' decisions change depending on the activation patterns of specific neural areas. On the other hand, others (e.g. McClure et al., 2004a, 2004b and 2007) document how the activations of particular areas vary depending on the features of the available options (e.g. the amount, variance and temporal distribution of rewards)[27].

An economist may object that in spite of its descriptive failures, standard economic theory offers a highly general account of choice behaviour, which enables economists to model a wide variety of decisions within a common mathematical framework (e.g. Dixit, 1990). One might even argue that economic theory has a degree of generality comparable to that of some theoretical frameworks in the natural sciences such as Newtonian mechanics and the theory of natural selection (see e.g. Rosenberg, 1992, p.231-2). Regrettably, these assertions do not provide economists with a cogent rebuttal to the NEs' argument from explanatory insightfulness. To see this, let us distinguish between the notions of generality and explanatory reach.

[26] See e.g. Churchland and Churchland, 1996, on how Newton's laws of motion unified previous accounts of terrestrial and celestial motion. For a critical appraisal of the unificationist account of explanation, see e.g. Halonen and Hintikka, 1999.

[27] In NE game theoretic studies, neural activation patterns have been shown to vary depending on: players' degree of cooperation in past moves (e.g. Rilling et al., 2002, and Sanfey et al., 2003); players' perceived fairness (e.g. Rilling et al., 2004a, and Singer et al., 2006); the theory of mind possessed by the players (e.g. Rilling et al., 2004b); and players' reputation (e.g. King-Casas et al., 2005, and Delgado et al., 2005).

In everyday scientific discourse, models and theories are said to be generalizable when they can be applied to a vast range of actual and hypothetical situations (see e.g. Gabaix and Laibson, 2008)[28]. Now, NEs may acknowledge that economists' mathematical tools can be employed to model a wide variety of choices and target systems. This, however, does not imply that economists' models of choice have a great explanatory reach, i.e. provide explanatory insights regarding a vast range of phenomena. Indeed, various NEs complain that economists, psychologists and neuroscientists can offer only "local explanations" of choice behaviour, which allow one to make "only very limited predictions about how [people] will behave in the future" (Glimcher, 2010, p.14; see also Padoa-Schioppa, 2008, p.455, for similar remarks).

Premise 2 states that neural findings can help economists to develop more adequate singular explanations of economic phenomena. The thought is that NEs can better account for "why a certain fact occurred in a certain way", providing us with detailed insights regarding "the causal and structural relations that produced [it]" (Aydinonat, 2010, p.159). By way of illustration, let us consider some NE studies which manipulate people's choice behaviour by means of neurochemicals. These studies aim to "identify the reasons why different conditions produce different behaviors [by] using drugs to cause changes in brain activity" (Vercoe and Zak, 2010, p.143). For example, Kosfeld et al. (2005) show how inhaling oxytocin makes subjects more likely to invest more, yet without altering their risk preferences (see also Zak et al., 2004).

[28] See also Matthewson and Weisberg (2009, p.182) and Weisberg (2004, p.1076), on the distinction between *a-generality* - which concerns how many actual phenomena a model applies to - and *p-generality* - which relates to how many logically, nomologically or physically possible systems are targeted by a model.

The significance of these contributions can be appreciated in light of the interventionist account of explanation, according to which explanation consists in displaying patterns of counterfactual dependence between the investigated target variables and particular sets of intervening variables (Woodward, 2003, and Woodward and Hitchcock, 2003a and 2003b). The idea is that NE findings help economists provide robust generalizations regarding how the examined target variables (e.g. specific agents' decisions) vary under experimentally regimented interventions on specific neuro-physiological variables.

Premise 3 asserts that NEs' insights explain some of the anomalies faced by standard economic theory. For instance, Rustichini maintains that neuro-psychological findings enable NEs to account for some inconsistent decisions by explaining them as the result of an optimal adjustment of people's learning mechanisms to the choice situations they face (2009, p.55). Similarly, Brocas and Carrillo (2010) claim that "evidence from the brain sciences […] can help uncover the 'true' motivations for the 'wrong' choices and improve the predictive power of the theory"[29]. In this respect, it would be of little import to reiterate (see e.g. Gul and Pesendorfer, 2008) that economists can accommodate a vast array of experimental and empirical results by altering the mathematical formulation of agents' preferences and constraints. For accommodating the available evidence does not amount to accounting for it.

To render this point more vivid, let us compare briefly the Ptolemaic model of planetary motion and Newton's gravitational model. The epicycles postulated to reconcile the

[29] Similar claims were put forward by some behavioural economists. For instance, Rabin holds that "because psychology systematically explores human judgement, behaviour and well-being, it can teach us important facts about how humans differ from the way they are traditionally described by economists" (1998, p.11).

former with celestial observations accommodate the trajectories followed by specific planets, yet do not offer informative insights as to why the planets move as they do. On the contrary, Newton's model provides an explanatorily instructive account of planetary motion in terms of gravitational force. By way of analogy, consider some behavioural economic models such as prospect theory. These models often fit observed choices better than standard EUT. At the same time, they do not provide explanatory insights regarding why the anomalies of EUT emerge in the first place. In this respect, NE promises to enable economists to better account for people's decisions by disclosing the underlying neuro-psychological mechanisms.

To give one example, consider ambiguity aversion (Ellsberg, 1961). Three main explanations of this phenomenon have been proposed in the economic literature. The first suggests that when the odds are unknown agents assume that someone may control the odds to their disadvantage. The second holds that people treat probabilities as if they were outcomes and thus tend to be risk-averse with regard to probabilities as they are concerning outcomes. The third account interprets ambiguity aversion as the result of people's reluctance to bet whenever they think that others possess information they lack. Now, choice behaviour data rarely enable researchers to discriminate between these competing accounts of ambiguity aversion. In a recent study, Hsu et al. (2005) document that several neural areas (e.g. the amygdala and the orbitofrontal cortex) exhibit different activations in conditions of ambiguity as opposed to conditions of risk. This finding does not *per se* discriminate between the first and the third interpretation of ambiguity aversion, but casts doubt on the second explanation (see Keren and Gerritsen, 1999, for some subjective reports in support of those interpretations).

2.E WELFARE ANALYSES

So far, we have examined several respects in which incorporating neural insights can improve *positive* economic theory. NEs, however, advocate the neural enrichment of the economic account of decision making also on *normative* grounds. In particular, several NEs take neuro-psychological findings to challenge traditional economic welfare analyses and policy evaluations. Below I consider the *argument from wants / likes divergences* by means of which some NEs criticize standard welfare analyses. The reasoning, which targets economists' alleged identification between agents' wants and likes, goes as follows:

P.1 Economic welfare analyses focus on people's wants.

P.2 People's well-being depends on their likes.

P.3 People's wants and likes diverge.

C(1) Economic welfare analyses fail to capture agents' well-being.

P.4 NEs' findings accurately measure people's well-being.

P.5 NEs' findings help people choose what promotes their well-being.

C Economic welfare analyses should rely on NEs' findings.

Let us examine the various steps of this argument in turn. As stated by *premise 1*, welfare economists are typically concerned with people's wants. These are usually inferred from observed choices or agents' own reports. According to *premise 2*, however, people's well-being relates not so much to what they struggle to obtain, but to what they actually like. The idea is that satisfying people's preferences does not

necessarily make them better off (Loewenstein et al., 2008, p.666) and that economic welfare analyses should target what individuals like rather than what they happen to desire and choose (Camerer, Loewenstein and Prelec, 2005, p.37). In the words of Gul and Pesendorfer, many NE studies rest on the assumption that "what makes individuals happy […] differs from what they choose" and that economic welfare analyses should be based on what makes individuals happy rather than "the utilities governing choice" (2008, p.3).

Premise 3 states that people's wants and likes diverge. The thought is that people's motivations for action are "not always closely tied to hedonic consequences" (Camerer, Loewenstein and Prelec, 2005, p.37; see also Camerer, 2006). There are various reasons why a mismatch between wants and likes can occur. To give one example, people's preferences are occasionally based on mistaken beliefs and can be prone to manipulation (see e.g. Elster, 1983, and Sen, 1987). Moreover, people's epistemic and computational limitations constrain their ability to judge whether they will like what they want (see e.g. Bernheim and Rangel, 2008, and Loewenstein and Haisley, 2008). Indeed, people often struggle to obtain things they will not like in the long run due to self-control problems (see e.g. Loewenstein, 1996, and Ross et al., 2008, on drug addicts and pathological gamblers).

The purported correspondence between likes and wants was disputed by some cognitive scientists before the advent of NE. For instance, Berridge (1996) challenged the assumption that people struggle to obtain what they will like by pointing out that two separate systems - one responsible for motivation and desire (wanting system), the other

responsible for pleasure and pain (liking system) - underlie people's behaviour. As Berridge's distinction suggests, satisfying people's immediate desires may fall short of maximizing their longer-term hedonic satisfaction. Moreover, the alternative that would maximize an agent's hedonic satisfaction may fail to be among her preferred options.

On the basis of the first three premises, *conclusion 1* doubts that standard economic welfare analyses provide informative insights concerning people's well-being. According to *premises 4* and *5*, identifying the neuro-psychological processes underlying decisions enables NEs to measure well-being and help people choose what promotes their well-being. The thought is that NE findings help us identify both why people often act in ways that do not promote their well-being and what kinds of interventions are likely to correct their actions. In this perspective, NE promises to integrate two major lines of research at the boundary between economics, psychology and neuroscience, namely happiness studies (see e.g. Bruni and Porta, 2005 and 2007) and recent investigations on how to design and alter people's choice architectures. Let me expand on this latter point.

Various ways to correct people's decisions have been proposed in the literature, ranging from restricting agents' choice set to making some options more salient than others. By way of illustration, let us consider what kinds of intervention are respectively advocated by asymmetric paternalists and libertarian paternalists. The proponents of asymmetric paternalism (e.g. Camerer et al., 2003, and Camerer, 2006) argue that it is often possible to benefit people who make suboptimal decisions without imposing substantial costs or restrictions on those who make optimal decisions. For their part, libertarian paternalists

(e.g. Sunstein and Thaler, 2003, and Thaler and Sunstein, 2008) maintain that people can be induced to make better decisions without having their autonomy infringed. In the words of Loewenstein and Haisley, NE research makes it possible to "steer human behavior in more beneficial directions while minimizing coercion" (2008, p.6).

I am not concerned here with settling the merits of these assertions. In section 4.A, I shall put forward some cautionary remarks regarding the relevance of neuro-psychological findings for economists' welfare analyses. In section 7.C, I shall critically assess some NEs' attempts to measure and promote the well-being of economic agents. In doing so, I shall argue that NEs' calls to ground economists' normative analyses on neuro-psychological findings rest on presuppositions which transcend both the scope of traditional decision theory and the evidential reach of NE investigations.

2.F THE META-ARGUMENT IN FAVOUR OF NEUROECONOMICS

As we have seen in the previous sections, economists can improve their models with regard to specific modelling desiderata by incorporating neuro-anatomical and neuro-physiological insights. The arguments we presented above can be combined in a cumulative case in favour of NE in a relatively straightforward way. The idea is to take the conclusion of each of those arguments and include it as premise in a meta-argument whose conclusion is that economists should incorporate NEs' insights into their account of choice behaviour. This meta-argument goes as follows:

P.1 NE insights increase the descriptive accuracy of economic models of choice.

P.2 NE insights overcome the predictive failures of standard economic theory.

P.3 NE insights enable economists to discriminate between competing economic models of choice.

P.4 NE insights help economists to better explain people's decisions.

P.5 NE insights accurately measure and effectively promote people's well-being.

C Economists should incorporate NE insights into their account of choice behaviour.

Prima facie, this meta-argument seems to provide compelling reasons in favour of a neural enrichment of economic theory. As I argue in the following chapters, however, there are cogent reasons to resist this cumulative case in favour of NE. More specifically, I shall articulate several lines of argument which attempt to demonstrate that economists are provisionally justified in retaining a methodologically distinctive

approach to the modelling of decision making. Let me briefly anticipate these arguments in turn.

In *chapter three*, I critically assess the evidential basis on which NEs' findings and inferences rest. In particular, I argue that many NEs' claims can be disputed on purely evidential and epistemological grounds. The reasoning I present in *chapter four* can be summarized as follows. NEs and other economists respectively value different modelling desiderata. These desiderata often pull in opposite directions and make contrasting demands on modellers. Those contrasts, in turn, severely constrain the incorporation of neural insights into economic models of choice.

In *chapter five*, I cast doubt on NEs' attempts to elaborate a single, general theory of choice behaviour on two main grounds. The former concerns the profound dissimilarities (e.g. in terms of employed constructs and pursued explanatory aims) between the economic, psychological and neuroscientific accounts of decision making. The latter relates to some central respects (e.g. how NE is supposed to inform economic theory) in which NEs themselves hold contrasting positions. In *chapter six*, I provide a case study to illustrate how the conceptual differences between NE's parent disciplines constrain the relevance of neuro-psychological findings for the economic theory of choice. In *chapter seven*, I differentiate various senses of the term "revolution" and argue that NEs are unlikely to prompt revolutionary modifications in economic theory in any of these senses.

CHAPTER THREE

ARGUMENT FROM UNCONVINCING EVIDENCE

The recent advances at the interface between economics, psychology and neuroscience have encouraged various NEs to put forward quite ambitious assertions regarding the relevance of NE for its parent disciplines. NEs' claims have prompted a variety of reactions among the practitioners of these disciplines. In particular, some economists welcome the opportunity to enrich specific models of decision making in light of neuro-psychological findings. Still, most remain convinced that NE is *de facto* (e.g. Harrison, 2008a and 2008b, and Rubinstein, 2008) or even *in principle* (e.g. Gul and Pesendorfer, 2008) incapable of triggering revolutionary modifications in the economic theory of choice.

In this chapter, I critically assess the accuracy and the reliability of the evidential basis of brain-imaging and brain-stimulation findings. Moreover, I identify and discuss some epistemological issues which arise concerning the inferences made in those studies. My reasoning can be summarized as follows. NEs criticize standard economic theory in several respects and speak of introducing profound modifications in economic models of choice. For NEs' criticisms and proposals to be effective, their evidential basis must be statistically significant and robust to changes in experimental settings. In many cases, however, NEs fail to show that the evidential basis of their claims is statistically significant and robust to changes in experimental settings. Hence, many NEs' claims can be resisted on purely evidential grounds.

Before proceeding, let me emphasize that my critique is by no means intended to cast doubt on the merits of brain-imaging and brain-stimulation studies indiscriminately. On the contrary, my aim is to identify the main strengths and limitations inherent in those investigations so as to better assess the potential for success in NE research. In sections 3.A and 3.B, I focus on the issues that respectively arise in relation to the *collection* and the *interpretation* of neural data. In section 3.C, I assess the reliability of specific kinds of *inferences* that NEs make in their studies.

3.A COLLECTION OF DATA

The experimental protocols employed in NE studies differ in a number of respects, ranging from the kind of examined organisms to the instruments used to monitor the neural areas of interest. The design of specific experiments includes elements such as: *production procedures*, which prescribe what stimuli are to be presented to the subjects and the temporal distribution of these stimuli; *measurement procedures*, which indicate what variables are to be monitored in the pre-stimulus, inter-stimulus and post-stimulus phase; and *detection procedures*, which specify what value the measured variables must have for the experimenters to legitimately conclude that the phenomenon of interest occurred (Sullivan, 2009, p.514). In this section, I examine some evidential and epistemological issues which arise with regard to the *collection* of raw neural data. More specifically, I discuss in turn the limited availability and representativeness of data, the insufficient spatial and temporal resolution of current NE instruments, and the constrained reliability of the proxies targeted in NE studies.

1) Limited availability and representativeness of data

Availability of raw neural data is often claimed to be a crucial prerequisite for the verifiability and the reproducibility of NE findings. Regrettably, NEs rarely make raw neural data publicly available to other researchers (Harrison, 2008a, and Spiegler, 2008). In this respect, one may well note that other neuroscientists rarely make raw neural data public (Quartz, 2008, p.467). This, however, merely redounds to the disputable character of the current experimental practice, which violates the standards of

most economic practitioners and fails to incentivize the adoption of rigorous procedures for data processing (Ortmann, 2008, p.442). That is to say, precisely because of the difficulties involved in interpreting and replicating NE experimental reports (see below) it is highly advisable that NEs make raw neural data publicly available.

As to the *representativeness* of the published results, some authors (e.g. Glaeser, 2008) worry that the increasing accessibility of brain-imaging and brain-stimulation technology, coupled with the decreasing costs for setting up experimental trials, incentivizes selective presentation of findings. To be sure, inter-laboratory competition discourages strategic misrepresentations of findings and can alleviate the flaws inherent in individual studies (Hubbard, 2003). Yet, the complexity of the experimental designs and the inferential steps involved in NE studies render it more difficult to assess the representativeness of the published results. More generally, the impression remains that the hurry to colonize new areas of investigation led some NEs to overstate the evidence supporting their claims. To render this point more vivid, let us consider the limited *size* of the experimental samples employed in most NE studies.

The number of subjects whose brains are monitored in neuroscientific experiments is usually quite small, i.e. typically less than twenty (Cabeza and Nyberg, 1997 and 2000). The limited size of the experimental samples, in turn, significantly constrains the reliability of the conclusions derived from the collected data. Now, while many practising neuroscientists acknowledge this concern, NEs frequently gloss over it as if it was of negligible importance. For instance, Bhatt and Camerer (2005, p.432) dismiss the complaint that a small number of subjects are typically monitored in fMRI

investigations by claiming that "for most fMRI studies [16 subjects] is usually an adequate sample to establish a result because adding more subjects does not alter the conclusions much". As shown by the history of lesion studies (see e.g. Bechtel, forthcoming), one may occasionally gain informative neuro-physiological insights on the basis of a small experimental sample. However, it is still an open empirical question whether monitoring just a few subjects enables NEs to obtain robust findings in "most fMRI studies" (see Thirion et al., 2007, for discussion). Moreover, the mere fact that the results of some experiments do not considerably vary when the size of the examined sample is increased does not license confidence in the accuracy of those findings. For one might get stable experimental outcomes even in cases where the technology provides rather inaccurate measurements of the investigated phenomena.

2) Insufficient spatial and temporal resolution

The accuracy of brain-imaging and brain-stimulation reports depends on a number of factors. Among these, we find the spatial resolution and the temporal resolution of the employed instruments. *Spatial resolution* refers to a scanner's capacity to detect the elementary units of brain activation (Logothetis, 2008a, p.870). So-called 'voxels' (volume pixels) constitute the smallest box-shaped part of a three-dimensional scan and can be regarded as the basic observational units targeted by neuro-imaging studies. An unfiltered voxel in a typical fMRI study contains up to 6×10^6 neurons, with many of these having to activate to generate a detectable signal (Bechtel and Richardson, 2010, and Logothetis, 2008b). Now, advances in scanner technology will presumably allow for significant improvements in signal specificity and voxel size in the future. Still, the

point remains that the accuracy of current brain-imaging investigations is subject to severe constraints. This, in turn, calls into question NEs' calls to modify economic models of choice in light of the available neuro-physiological findings.

Similar concerns arise regarding the *temporal resolution* of the signals targeted in NE studies. These signals may reflect the integration of up to a few seconds of neuro-physiological activity (Henson, 2005). Given that neural activations may take place within milliseconds, the workings of various neural populations are likely to go unnoticed. This problem has been alleviated thanks to the introduction of scanners with higher temporal resolution, but affects even the so-called event-related fMRI, where several signals can be generated within a second (Rosen et al., 1998). That is to say, even the most advanced scanners yield information that is "orders of magnitude coarser" than neural firings themselves (Roskies, 2008, p.24). These limitations, in turn, call into question the suitability of NEs' current technology for studying the neural substrates of choice behaviour.

3) Unreliable proxies

The signals examined in fMRI and PET studies do not measure neural activity directly. On the contrary, they target physiological factors "that are causally related in a rather complex way to downstream consequences of neural activity" (Roskies, 2008, p.25). In particular, neural activations are usually estimated on the basis of *proxies* whose accuracy and reliability have been questioned on various grounds (e.g. Logothetis et al., 2001, and Logothetis and Pfeuffer, 2004). To render this point more vivid, let us briefly

compare the processes by means of which the signals targeted in PET and fMRI studies are generated.

As shown by the figure below, when a neural area activates, one typically observes a noticeable increase in cerebral blood flow and in glucose utilization (Fox et al., 1988), coupled with a differential increase in oxygen consumption (Fox and Raichle, 1986). PET targets variations in regional cerebral blood flow, while fMRI captures discrepancies between variations in regional cerebral blood flow and changes in oxygen consumption (Kim and Ugurbil, 1997, and Ogawa et al., 1990)[30]. In particular, the BOLD signal targeted by fMRI reflects the mismatch between variations in regional blood flow and changes in the amount of oxygen remaining at the site of brain activation (Raichle, 1998).

Source: Raichle (1998, p.770)

[30] The quantity of blood flowing to a neural area can be estimated thanks to the fact that oxygenated blood is diamagnetic (i.e. it weakly counteracts the applied magnetic field) and that deoxygenated blood is paramagnetic (i.e. it slightly enhances the magnetic field). For further details, see Huettel et al. (2004, ch.1).

Let us focus on BOLD signals. The interpretation of these signals is complicated by the fact that their strength does not always reflect a region's intensity of activation in accurate terms. There are various reasons why this can happen (see e.g. Logothetis and Wandell, 2004). Let us consider three such reasons in turn. Firstly, BOLD signal responses non-linearly vary in strength with the strength of the applied magnetic field and increase only marginally in regions having naturally high blood flow (Bechtel and Richardson, 2010). Secondly, oxygenation variations in large vessels can conceal oxygenation changes in the capillary bed and generate spurious BOLD signals (Klein, 2010a). And thirdly, excitatory and inhibitory interactions between different neural populations (see e.g. Douglas and Martin, 2004) occasionally generate considerable dissociations between actual neural activity and the metabolic variations targeted by fMRI and PET instruments (Buzsaki et al., 2007, and Logothetis, 2008a). Indeed, neuromodulatory interactions can induce larger perturbations in BOLD signals than the sensory inputs themselves, with increases in BOLD signals occurring even without a net excitatory activity in the examined neural populations (Jueptner and Weiller, 1995, and McCormick et al., 2003).

3.B INTERPRETATION OF DATA

In this section, I consider some of the evidential and epistemological issues which arise in relation to the *interpretation* of neuro-physiological data. More specifically, I discuss in turn the alleged arbitrariness, the derived character and the limited generalizability of many brain-imaging and brain-stimulation findings. This list does not exhaust the set of concerns related to the interpretation of NE studies[31]. Yet, it identifies some major respects in which we can evaluate the evidential basis of NEs' investigations.

1) Arbitrary Findings

The definition of the baseline conditions of activation and the identification of activation thresholds constitute two major sources of the purported *arbitrariness* of many NE findings. Let us examine these two issues in turn. In a typical brain-imaging study, the *baseline* conditions "consist of lying quietly but fully awake in [the] scanner with eyes closed or passively viewing a television monitor" (Raichle, 1998, p.768). However, complex physiological and metabolic variations usually take place in the human brain even when one is not engaged in specific cognitive or computational tasks (Gusnard and Raichle, 2001, and Newman et al., 2001). This, in turn, significantly complicates the interpretation of the observed activation patterns (see the *point 3* below).

[31] See e.g. Roskies (2007) and Savoy (2001) on how observed activations can be given different interpretations depending on one's background assumptions regarding brain's structure and functioning.

With regard to *thresholding*, the fact that claims of statistical significance are relative to the choice of a specific significance level has subtle implications for the interpretation of neuroimages. In particular, various authors worry that thresholding at any level of significance tends to generate "artificially sharp barriers between 'active' and 'inactive' regions" (Klein, 2010a, p.270). This concern can be partly mitigated by providing gradations of color representing different magnitudes of the examined test statistics (Hubbard, 2003, p.29). Yet, images reporting dissimilar patterns of activation can be obtained by setting different activation thresholds. Moreover, what activations are detected at particular significance levels often varies depending on the temporal and spatial specificity of the employed scanners (Huettel et al., 2004, and Thirion et al., 2007). In light of these remarks, the precision of the colourful pictures appearing in several NE studies appears to be elusive and to some extent misleading (Hardcastle and Stewart, 2002)[32].

2) Derived Character of Findings

The experimental reports presented in NE articles are typically obtained after a number of *data manipulations* and *statistical adjustments*. These manipulations and adjustments alter raw neural data in ways which often elude the experimenters' control (see e.g. Bullmore et al., 1995, and Uttal, 2001). In what follows, I focus on the BOLD signal targeted in fMRI studies and critically inspect three kinds of intervention required to make raw fMRI data amenable to analysis and intersubject comparisons,

[32] See also Weisberg et al. (2008) on people's tendency to regard specific psychological explanations as more compelling even when irrelevant neuroscientific data are presented in their support.

namely: corrections for the independence of statistical tests, brain averaging and brain normalizations. Let us consider these three interventions in turn.

Neuroimages do not report brain activity directly, but rather indicate regions where the data license some confidence that the examined areas underwent differential activations across experimental conditions (Klein, 2010b, p.187). In a typical brain-imaging study, thousands of *statistical tests for activation* are performed across voxels (Friston et al., 1995a). The significance test for activation consists of two steps. Firstly, one estimates the likelihood that one would observe an area's activation when the experimental tasks do not engage activation in that area (spurious activation). Secondly, one compares the estimated likelihood to a predetermined significance level for each investigated brain region. The results are summarized in statistical parametric maps, which illustrate in what areas the data warrant a confident assertion of differential activity across experimental tasks (Klein, 2010a, p.267-8).

Given that the signals coming from spatially contiguous voxels are often correlated, the statistical tests performed on these regions are not independent and tend to yield false positives (Henson, 2005, p.208). To remedy this problem, researchers correct the obtained statistical estimates in various ways (see e.g. Friston, 2003, and Friston et al., 1995b)[33]. Even so, several authors (e.g. Kiebel et al., 1999) question the validity of those corrections. Indeed, some go as far as to contend that the correlations reported in neuroscience studies are endemically inflated (see e.g. Vul et al., 2009, on social

[33] For instance, one can constrain the search for effects in the main fMRI experiment to anatomically delimited regions by performing a *functional localiser*, i.e. an auxiliary fMRI experiment which aims to isolate a functionally specialised region of interest before the implementation of the main fMRI experiment (Friston and Henson, 2006, and Saxe et al., 2006).

neuroscience investigations)[34]. In such a context, further concerns arise from the fact that most fMRI studies rely on voxel-based analyses of time series data, with estimates from one stage being taken as data in the next stage. This, in turn, calls into question the accuracy of the subsequently reported findings, as standard errors of estimates are likely to propagate at later stages (Harrison, 2008a).

With regard to *brain averaging*, neuroscientists usually average the registered activation signals both over different trials for each subject and across different subjects. In this perspective, brain images are best regarded as generalizations rather than particulars, as there typically is no specific physical brain that those images are intended to represent (Roskies, 2008, p.26). In neuroscientific research, averaging is frequently advocated on the ground that it minimizes the effects of interfering signals on the registered activation patterns (Bechtel and Mundale, 1999). However, due to the interpersonal variability exhibited by several areas' activations, pooling data across subjects usually constrains the signal's spatial precision (Van Orden et al., 2001, p.153). Furthermore, simple averaging can decrease or even suppress signals (e.g. suppose that during the experiment the monitored region activates in some subjects, but deactivates in others). For this reason, neuroscientists usually rely on more complex averaging procedures which take into account the temporal covariance of voxels in the functional regions of interest (Friston et al., 2006, p.1081). These procedures, however, further complicate the interpretation of neural data and constrain the comparability of the results obtained by means of different experimental protocols.

[34] These criticisms have not remained unchallenged. For instance, on the basis of some simulations Lieberman et al. (2009) argue that fMRI analyses are unlikely to report spurious high correlations. Yet, they also concede that "the effect sizes from whole-brain analyses are likely to be inflated" (2009, p.306; see also Aron et al., 2006).

As to *brain normalizations*, the idea is to make the image of distinct brains fit on the image of a standard brain by aligning several intermediate points of each subject's brain with the corresponding points of the standard brain (Aizawa, 2009, p.503). Brain normalizations are typically performed to alleviate the constraints that the anatomical differences between individual brains impose on the interpersonal comparability of brain-imaging findings. Still, the very act of implementing brain normalizations may screen off significant anatomical and functional dissimilarities between the neural architectures of distinct subjects. Moreover, employing different normalization templates can lead one to interpret observed activation patterns in dissimilar ways (see e.g. Ashburner and Friston, 1999, and Gispert et al., 2003).

3) Limited generalizability of findings

A number of concerns arise regarding the *generalizability* of many NE findings. The first worry can be explicated as follows. As we have seen above (point 2), several auxiliary assumptions and experimental manipulations are required to interpret raw neural data. This, in turn, makes it difficult to compare the evidential reports obtained in different experiments. To be sure, not all differences across experimental procedures constrain the comparability and the generalizability of the obtained results (e.g. think of marginal variations in inter-stimuli intervals). Still, the point remains that most NE findings "are highly indexical to the experimental set-up" employed to obtain them (Kuorikoski and Ylikoski, 2010, p.223; see Sullivan, 2009, for similar remarks concerning laboratory experiments in cognitive neuroscience). To render this point

more vivid, below I consider three issues which constrain the generalizability of NEs' findings. These issues respectively concern the functional and anatomical differences between the neural architectures of different species, the interpersonal variability of neural activation patterns, and the experimental setup of NE studies.

NEs investigate the neural architecture of a variety of *species*. The decision to focus on a given species is guided by both ethical and pragmatic considerations. For instance, the ethical concerns related to the non-therapeutic use of invasive techniques limit the applicability of single neuron measurements to humans. To remedy this problem, various authors employ non-human primates as models for human subjects on the alleged ground that "the human brain evolved over time by extending homologous functions, and computations, in predecessor brains" (McCabe, 2008, p.352; see also Gazzaniga et. al., 2002).

Now, cognitive scientists acquired valuable insights concerning the human neural architecture by working on the assumption that the human brain was sculpted by long-lasting evolutionary pressures (e.g. MacLean, 1990, and Tooby and Cosmides, 1994 and 2005). Even so, the neuro-anatomical and neuro-physiological dissimilarities between humans and other species frequently prevent animal studies from serving as an informative basis for investigating higher cognitive functions (Allman et al., 2002). This worry is especially pertinent when it comes to examining the neural substrates of complex decision making tasks that cannot be studied in non-human primates. To be sure, some NEs have recently extended previous findings in non-human primates to

more complex decision problems[35]. Still, even leading NEs occasionally appear to naïvely presuppose that the human brain "is basically the primate brain with extra neocortex", with the primate brain being "a simpler mammalian brain with some neocortex" (Camerer, 2007, C29).

A second constraint on the generalizability of NEs' findings relates to the *interpersonal variability* of neural areas' activation patterns. To explicate this concern, let us consider the plasticity of the human neural architecture. The human brain exhibits various kinds of plasticity. For instance, cortical plasticity obtains when a given psychological function is implemented by anatomically distinct neural areas at different times. Synaptic plasticity, instead, consists in the reinforcement or inhibition of specific synaptic connections between neurons in response to their past interactions (Buonomano and Merzenich, 1998). People's capacity to recover specific computational and cognitive abilities after various traumas prompted some researchers to ascribe high cortical plasticity to the human brain (Liepert et al., 2000, and Richardson, 2009). Yet, cortical plasticity is normally more limited than synaptic plasticity (Gazzaniga, Ivry and Mangun 2002, ch.15)[36]. Now, due to these kinds of brain plasticity most neural areas exhibit some variability in their activation patterns at both the interpersonal and the intertemporal levels (Henson, 2005). This variability, in turn, calls into question some NEs' ambition to formulate wide-scope empirical generalizations regarding the neural substrates of choice behaviour.

[35] See e.g. Dayan and Niv, 2008, and Pessiglione et al., 2006, on the dopaminergic underpinnings of learning in humans; see also Knutson and Peterson, 2005, on how dopamine circuits respond to the receipt of money.

[36] See also Goldberg (2005, ch.14) on how the human neural architecture exhibits a varying degree of distinct kinds of plasticity across developmental stages.

A third constraint on the generalizability of NE findings relates to the *experimental setup* of NEs' studies. The reasoning goes as follows. The majority of NEs have hitherto focused on individual choices and two-person interactions in highly controlled experimental settings, where subjects are "sealed off from advisors, artifacts and other distributed cognitive resources" they have in real life situations (Wilcox, 2008, p.530). This narrow focus, in turn, constrains the generalizability of NEs' findings. For the behaviour of real life economic agents is simultaneously shaped by a number of socio-cultural factors - ranging from agents' axiological commitments to institutional incentives (see e.g. Hutchins, 1995) - that are only partly controlled for in single NE investigations (see also section 4.A).

To be sure, various NEs investigate the neural processes and mechanisms associated with the socio-cultural factors which influence people's behaviour (see e.g. Montague and Lohrenz, 2007, and Spitzer et al., 2007, on the neural substrates of social norms compliance). Yet, while decisions outside the laboratory are influenced by all these factors simultaneously, NEs examine those factors in isolation by screening-off other environmental influences. In this respect, NEs may well object that similar problems affect many studies in experimental economics and that exploring how people's brains interact with their socio-cultural environment can be "valuable" to economists (McCabe, 2008, p.365). Even so, NEs still have to integrate their focus on individual cognition with insights regarding "how individuals work in social groups [and] make use of environmental scaffolding" (Craver and Alexandrova, 2008, p.397)[37].

[37] Various studies in experimental economics (e.g. Smith, 1991) document how specific agents violate rational choice predictions when tested as isolated individuals, yet make consistent decisions in the context of market institutions. As Arrow puts it, "rationality is not a property of

In such a context, the additional worry arises that brain-imaging and brain-stimulation techniques do not enable NEs to monitor subjects in the context of their daily lives (Bechtel, forthcoming). This concern can be explicated as follows. When a subject lies in a scanner, various neural areas (e.g. those involved in spatial awareness and emotional responses) are likely to exhibit activation patterns that differ from the ones they would undergo outside a scanner due to the fact that the subject is aware of being monitored (see e.g. Bechtel and Stufflebeam, 2001). As Ortmann vividly puts it: "Part of the noise may come from subjects engaging in thoughts […] like: Why am I lying here? What's that dizziness? […] Is this harmful to my health? Have they lied to me after all about the health risks? What if I think nasty/naughty thoughts? Can the experimenter decipher them?" (2008, p.437).

The above considerations point to a severe instance of the so-called problem of *observer interference*, according to which the very act of measuring specific phenomena alters them in ways which elude experimental controls (e.g. think of the Heisenberg principle in particle physics). This problem is rarely acknowledged by NEs, but subtly constrains the *external validity* of their studies (on the external validity of experimental results in economics, see Guala, 2003 and 2005, ch.7). To be sure, one might object that the aforementioned biases are unlikely to affect many NE studies. After all - the thought would be - NEs are rarely concerned with the neural substrates of spatial awareness and emotional responses. However, due to the widespread anatomical and functional interconnections between neural populations, variations in

---

the individual alone […] It gathers not only its force but also its very meaning from the social context in which it is embedded" (1987, p.201).

those areas' activations may significantly alter activations in the areas monitored by NEs, thereby confounding the results of their studies. In particular, significant discrepancies are likely to arise both between the activation patterns that some areas respectively display when a subject is and is not monitored (conditions I and II) and when a subject is and is not aware of being monitored (conditions III and IV).

Now, it would be helpful to know which areas are affected by these variations and how exactly these areas' activations change across conditions I and II and conditions III and IV respectively. Regrettably, current brain-imaging and brain-stimulation technology enables us to monitor the neural substrates of choice behaviour only in conditions I and III. Hence, for all we know, most NEs may have recurrently reported highly context-dependent evidence which fails to reliably indicate what activation patterns neural areas undergo in real life situations. In this respect, it would be of little import to conjecture that NEs could supplement their measurements with some account of the neural mechanisms that give rise to the observed activations. For any such account would be presumably based on the context-dependent data provided by the available scanners. In the words of Roskies, the worry arises that the "interpretation of neuroimages takes place within a framework largely established by means of the technique itself" (2008, p.30). Unfortunately, NEs do not seem to take this concern into adequate consideration. Indeed, even leading NEs generally gloss over it when presenting and discussing their findings.

3.C INFERENCES

NEs draw on a rapidly expanding corpus of neuro-anatomical and neuro-physiological findings in developing their models of decision making. Assessing in what respects NE can inform its parent disciplines requires us to accurately characterize the inferential steps on which NEs' experimental reports rests. In this section, I examine the scope and the limitations of the inferences made in many brain-imaging and brain-stimulation studies. More specifically, I focus in turn on functional localizations, the subtraction method, single and double dissociations, forward and reverse inferences, function-to-structure deductions and structure-to-function inductions.

Before considering these inferences in detail, let me put forward some remarks regarding *lesion studies*. Neuroscientists acquired valuable causal and functional insights concerning the human neural architecture by examining subjects with brain lesions. However, the inferences made in lesion studies are frequently characterized by significant limitations. Let us consider three main limitations in turn. To begin with, it is difficult to find a high number of subjects having exactly the same brain lesion in terms of location, gravity, etc. Investigating subjects with lesions in anatomically proximate regions rarely provides informative insights, as minor differences in the location and gravity of lesions can generate dissimilar cognitive and computational impairments. In this respect, some progress has been made thanks to pharmaceuticals which enable researchers to reversibly inactivate spatially delimited neural populations (Hubbard, 2003, p.28). Yet, the interpersonal comparability of the reports obtained in lesion studies is usually quite limited (Bechtel, 2002a, and Mundale, 1998).

Secondly, lesion studies rarely enable one to establish what exactly the damaged regions contribute to normal brain functioning. To appreciate this, suppose that some subjects can properly perform a cognitive task X before suffering a lesion in area α and lose that ability once area α is damaged. It might be tempting to infer that the injured region provides a necessary contribution to the normal execution of task X. Yet, the correlation between the malfunctioning of area α and subjects' failure to perform task X does not license this conclusion. For instance, the malfunctioning of α could result from the impairment of another area β, with the malfunctioning of β precluding the subject from performing task X irrespective of whether α is also damaged (Hubbard, 2003, p.29). Conversely, significant damages to a neural area may not produce any observable cognitive or computational impairment. For example, when some specific region α is damaged, other areas β, γ, etc. may implement the operation initially performed by α and prevent the functional deficit that would otherwise result.

Thirdly, attempts to acquire functional insights by means of lesion studies are complicated by the fact that there are various ways in which lesions impair subjects' ability to perform specific tasks. For instance, a lesion may: damage a neural area whose functioning is necessary to the execution of a task; distort modulatory feedback between distinct neural regions; occlude blood flow in a way that alters the workings of functionally related areas; and so on (see e.g. Bechtel, forthcoming). For this reason, even establishing significant correlations between a lesion and some resulting deficit may provide limited information concerning the functional role played by the damaged area. As I argue below, some of the limitations inherent in lesion studies have been

overcome thanks to brain-imaging and brain-stimulation investigations. Yet, even the inferences drawn in these investigations are affected by several shortcomings.

1) Functional Localizations

The human brain contains a number of regions specialized for processing specific types of signals and contributing to particular cognitive processes. These regions are neither functionally nor anatomically insulated, but carry out their operations in interconnected networks (Hubbard, 2003, and Medler et al., 2005). The aim of functional localizations is not just to identify where specific cognitive operations take place, but also to disclose the principles underlying brain organization (Bechtel and Richardson, 2010). The idea that distinct neural areas play dissimilar functional roles is grounded in the fact that different neural populations have heterogeneous anatomical features (e.g. in terms of cytoarchitecture and patterns of neurotransmission)[38]. Attempts to localize cognitive functions to specific brain locations consist of three stages (see e.g. Bechtel and Richardson, 1993, and Craver, 2007). Firstly, one decomposes the examined experimental task into a set of component cognitive operations that are deemed to be sufficient for its execution. Secondly, one identifies the set of neural areas composing the investigated brain region. And thirdly, one maps each cognitive operation on particular neural areas.

[38] The following asymmetry is worth noting (Aizawa, 2009). On the one hand, developing accurate anatomical maps constitutes a necessary but insufficient condition for elaborating detailed functional categorizations. On the other hand, acquiring a functional understanding of the human neural architecture facilitates the completion of neuro-anatomical maps, but is not necessary for developing them.

Neuroscientists have localized increasingly specific cognitive operations to progressively smaller neural areas (Bechtel, 2002b, and Hubbard, 2003). Even so, severe limitations affect the localizations drawn in current NE studies. Let us consider some of these shortcomings in turn[39]. The first limitation relates to the fact that the human neural architecture exhibits widespread *many-to-many mappings* between anatomical regions and cognitive processes. Two issues are worth mentioning in this respect. On the one hand, even the execution of simple cognitive and computational tasks usually elicits a distributed pattern of activation in several areas (one-to-many mapping). On the other hand, most neural areas respond to diverse stimuli and are engaged by dissimilar experimental tasks (many-to-one mapping). To be sure, the range of functional roles an area can play once its anatomical connectivity has been fixed is quite limited. Yet, a neural population can often perform heterogeneous functions depending on what areas interact with it (Price and Friston, 2005, p.268). In other words, the fact that an area is *involved* in the execution of some tasks does not imply that such an area is *specific* to those tasks[40].

Several studies document the existence of many-to-many interconnections between distinct neural populations. For instance, Cabeza and Nyberg (1997 and 2000) illustrate

[39] A further question arises regarding the alleged *implications* of functional localizations for the thesis that the human neural architecture is modular in character (see e.g. Fodor, 2000, and Mameli, 2001, on different versions of the modularity thesis). In cognitive neuropsychology, intense disputes took place concerning this issue. In particular, various authors argued that identifying the anatomical location *where* mental operations are performed provides limited (e.g. Shallice, 1988) or even no (e.g. Morton, 1984) information for understanding *how* the brain performs those operations. I am not concerned here with assessing these claims. For the purpose of this enquiry, it suffices to note that "the question of what defines a region anatomically is difficult and unresolved" (Henson, 2005, p.197) and that there are no *a priori* reasons to think that anatomically separate regions correspond to functional modules (Harley, 2004).

[40] A related distinction can be made between the concepts of *pluripotentiality* and *redundancy*. Pluripotentiality obtains when an anatomically delimited area can develop so as to perform a range of distinct operations. Redundancy, instead, occurs when two neural populations can perform the same function.

that distributed sets of brain regions are engaged even by simple experimental tasks (see also Anderson, 2006 and 2007). For his part, Gerlach (2007) compares various fMRI studies of visual processing tasks and finds that not a single area consistently activates for a given category across those studies. In such a context, one wonders at what *level of specificity* cognitive operations and anatomically separate areas can be associated. As noted by various authors (e.g. Roskies, 2008, p.27), cognitive processes can be decomposed and associated with specific brain locations only to a limited extent. For instance, researchers are unlikely to localize complex cognitive abilities such as language or attention, as these activate large sets of interconnected areas (see Petersen and Fiez, 1993).

Functional localizations are further complicated by the kind of *connectivity patterns* exhibited by specific neural areas. To give one example, neural circuitries frequently exhibit cascaded (rather than thresholded) and interactive (rather than feedforward) interconnections (Van Orden and Paap, 1997). This, in turn, can render it prohibitively difficult to perform functional localizations. To see this, consider the following passage by Coltheart (2004, p.22):

> "Suppose one believes that some cognitive system includes a sequence of three modules A to B to C, and one wants to localise one of these, say A, by imaging the brains of people as they carry out some task that requires module A. Because of the cascaded nature of the system, modules B and C will be activated if module A is, even if modules B and C are irrelevant as far as the task is concerned. And because of the interactive nature of the system, some of the activity in module A will be due

to feedback from modules B and C. […] How could one ever determine which parts of the brain activity here are specifically associated with module A?".

2) Subtraction Method

Neuroscientists often employ the *subtraction method* in investigating the human neural architecture (see Donders, 1969, and Sternberg, 1969). This method can be characterized as follows. Given an experimental task T_0 whose execution involves several cognitive operations, one attempts to design a control task T_1 requiring the implementation of an additional operation with respect to T_0. The activation patterns of specific neural areas are monitored during the execution of both tasks. The areas exhibiting differential activation, in turn, are deemed to contribute to the execution of the additional operation (Petersen et al., 1989). Various criticisms have been formulated concerning the subtraction method (e.g. Van Orden and Paap, 1997, and Uttal, 2001; see also Friston et al., 1996, and Sartori and Umiltà, 2000, for some attempts to tackle the limitations of subtraction studies). Let us examine two criticisms in turn.

The first criticism targets the *pure insertion* assumption made in most subtraction analyses. According to this assumption, one can precisely partition the processes occurring between stimulus and response into a series of successive operations, so that adding an extra component to a task does not alter the activations that particular areas exhibit during the execution of the other components of the task. However, this assumption is rarely satisfied. Moreover, it is often difficult to ascertain how exactly adding or removing a specific component affects the targeted neuro-cognitive processes across experimental settings (see e.g. Aertsen and Preissl, 1991, and Friston et al.,

1996). This limitation, in turn, constrains the reliability and the robustness of the conclusions drawn in subtraction studies.

Secondly, the activation patterns of the neural areas involved in the execution of a task may change during an experiment, depending on factors such as the extent to which the experimental task is perceived as novel by the subject (Raichle, 1998). In particular, neural adaptation - i.e. the response decrease to a stimulus that some neurons exhibit when the stimulus is repeated (Krekelberg et al., 2006) - can significantly complicate the interpretation of the observed activations. Indeed, standard subtractive designs fail to reveal what variations in the observed neural activations arise from changes in the experimental tasks rather than from adaptive processes in the targeted neural populations.

3) Single and double dissociations

Single dissociations aim to identify what contribution a neural area provides to the execution of a cognitive process by means of the following two-stage procedure. Firstly, one demonstrates that damaging or stimulating a neural area significantly affects the execution of one task, yet not that of another task. Secondly, one infers that the area whose functioning has been altered is involved in the execution of the first - yet not the second - task. There are various reasons to doubt that single dissociations license the conclusion that the examined region contributes to exclusively one of the two tasks. For instance, suppose that two tasks X and Y differ in the demands they respectively make on a neural area α, so that little damage to α results in the impairment of task X alone,

while severe damage to α causes the disruption of both activities. Observing only situations where α suffers minor damages could lead one to erroneously infer that such an area does not contribute to the execution of task Y (see Bechtel, forthcoming, for a similar example).

In order to overcome the inferential limitations associated with single dissociations, neuroscientists often attempt to find *double dissociations*. The idea can be characterized as follows. Consider two neural areas α and β and two experimental tasks X and Y. A double dissociation obtains when (i) altering the workings of neural area α affects the execution of task X (but not Y), and (ii) altering the workings of neural area β affects the execution of task Y (but not X). One can easily identify neural areas exhibiting dissimilar activation patterns across experimental conditions (single dissociations). Double dissociations are more difficult to find, and are often regarded as compelling evidence that different areas contribute to the execution of distinct operations (Robertson et al., 1993). This, however, is not necessarily the case. To see this, suppose that two neural areas α and β are jointly involved in the execution of tasks X and Y. By differently altering the workings of areas α and β, one may respectively affect the execution of tasks X and Y (double dissociation) even if the execution of those tasks involves both neural areas (Hinton and Shallice, 1991).

Most attempts to identify double dissociations rest on two fundamental assumptions, namely (i) that the human neural architecture is composed of functionally dissociable systems, and (ii) that modifying the workings of each of those systems has observable effects on subjects' experimental performance. However, neither of these assumptions is

vindicated by the identification of double dissociations themselves (Van Orden and Paap, 1997). In the words of Van Orden et al. (2001, p.114): "we require a reliable theory of cognitive modules, before the fact, to guarantee that we observe a pure dissociation […] we next require a theory of tasks, to tell us which […] modules are required by which laboratory tasks". In this respect, it is telling how Shallice (1988) dismisses his earlier claim (1979, p.191) that double dissociations demonstrate that distinct experimental tasks make dissimilar demands on different neural populations. As he points out, such reasoning incurs the following fallacy of affirming the consequent: "if modules exist, then […] double dissociations are a relatively reliable way of uncovering them. Double dissociations do exist. Therefore modules exist" (1988, p.248).

4) Forward and reverse inferences

As we have seen in section 2.C, several NEs urge economists to employ neural data for model selection and model evaluation purposes. In the neuroscientific literature, various authors advocate using brain-imaging data to discriminate between competing neuro-psychological theories (Hubbard, 2003). A *forward inference* discriminates between alternative neuro-psychological theories on the basis of the activation patterns exhibited by specific areas across experimental tasks. The idea can be characterized as follows. Given two competing cognitive theories T_0 and T_1, one designs experimental conditions C_1 and C_2 that differ in the engagement of a hypothetical function F only according to T_1. If function F affects the activation patterns of a specific area, then observing

differential activation in that area across conditions C_1 and C_2 favours theory T_1 over theory T_0 (Henson, 2005, p.197).

In making a *reverse inference*, instead, one conjectures that the monitored subject engages in a particular cognitive process on the basis of the activation patterns exhibited by specific brain regions (Henson, 2006, p.64). More specifically, one concludes that activation of area α in a task comparison reveals engagement in cognitive process X after observing that in several studies area α was active whenever the subjects engaged in cognitive process X (Poldrack, 2006, p.59). In other words, one takes the fact that a particular area activates across experimental tasks involving a specific cognitive process to corroborate the hypothesis that the area contributes to the execution of such a process (Poldrack and Wagner, 2004, p.177; see also Bub, 2000, p.475-6).

The strength of a reverse inference crucially depends on the monitored area's *selectivity of activation*, i.e. "the ratio of process-specific activation to the overall likelihood of activation in that area across all tasks" (Poldrack, 2006, p.20). That is to say, if a neural area α is activated in correspondence with a large number of cognitive processes, then its activation provides weak evidence that an individual is engaged in a specific cognitive process X rather than others. Now, for a reverse inference to be deductively valid, the monitored neural area would have to activate if and only if the examined cognitive process is engaged. Regrettably, this is hardly ever the case with the neural circuits related to decision making processes. Indeed, several areas have been shown to activate in relation to a wide range of stimuli (see e.g. Elliott et al., 2000, on the orbitofrontal cortex, and Baxter and Murray, 2002, on the amygdala). As a result,

reverse inferences in NE studies often amount to fairly weak inferences to the best explanation.

Assessing the strength of reverse inferences is complicated by the fact that an area's selectivity of activation is often difficult to estimate precisely. For instance, the size of an area can crucially influence its estimated selectivity of activation, as smaller regions usually activate more selectively than larger ones. Now, one could in principle increase a reverse inference's strength by focusing on anatomically more restricted brain regions. Still, some areas are attributed a greater selectivity of activation for the sole reason that they can be studied at a higher resolution than others. In this respect, even looking at what areas are shown to activate more frequently across studies does not enable one to ascertain their selectivity of activation, as some areas are investigated more often than others just because they can be monitored more easily (see e.g. Poldrack, 2006, on the BrainMap database at http://www.brainmap.org).

5) Function-to-structure deductions and structure-to-function inductions

In making a *function-to-structure deduction*, one infers - upon observing dissimilar activation patterns in two experimental conditions C_1 and C_2 - that performing these tasks involves at least one different function (Henson, 2005, p.197). A *structure-to-function induction*, instead, can be characterized as follows. If the execution of an experimental task C elicits activation in region α relative to some baseline condition, and region α has been independently associated with function F in different studies, then function F is engaged by the execution of C (ibid., p.198).

Structure-to-function inductions rest on stronger assumptions than function-to-structure deductions. These deductions, in fact, just assume that a cognitive process does not originate qualitatively different neural activations across experimental conditions. Structure-to-function inductions, instead, presuppose the existence of one-to-one mappings between functions and structural units across experimental situations. As I argued above (see point 1 in this section), such an assumption is rather demanding and rarely holds with the neural circuits associated with decision making processes. To be sure, whether a one-to-one mapping between functions and structures can be identified partly depends on how exactly the concepts of function and structure are defined. To see this, suppose that no systematic mapping was documented between a specific cognitive function and some anatomically delimited neural area. By redefining the concept of structure in such a way to denote a *network* of interacting areas, one may succeed in redescribing the available evidence as if it supported the existence of systematic function-structure mappings.

One might think that this redefinition of the concept of structure "simply dodges the question of whether there is a one-to-one function–structure mapping" (Henson, 2005, p.217). Yet, *pace* Henson, the possibility of implementing such a redefinition does not render the conjecture of a systematic function–structure mapping "unprovable in a logical sense". For once a precise definition of structure and function is provided, there will typically be some fact of the matter as to whether a systematic mapping exists between those *relata*. In this respect, Henson's claim (2005, p.228) that using imaging data to inform psychological theories requires a one-to-one mapping between

psychological functions and brain structures faces the following dilemma. On the one hand, defining brain structures as anatomically specific locations would render his assertion implausible. On the other hand, taking them to be networks of interacting neural areas would risk making it disappointingly uninformative.

Concluding remarks

To recapitulate, brain-imaging and brain-stimulation studies enable neuroscientists to acquire many valuable insights concerning the anatomical and functional organization of the human neural architecture. At the same time, various evidential and epistemological concerns arise regarding the inferences NEs make in those studies. These concerns do not license an unqualified scepticism concerning brain-imaging and brain-stimulation research. Still, they cast serious doubts on the accuracy and the reliability of the findings that some NEs regard as oracular pronouncements. In the words of Ortmann (2008, p.444): "There are too many open questions that have been pushed aside […] in the last few years people were rushing into NE to stake claims […] This rush is understandable, and individually rational, but came at a cost of questionable practices".

To be sure, showing that *some* NEs' studies lack a statistically significant and robust evidential basis does not license the claim that NE is incapable of informing its parent disciplines. In particular, waving the flag of vulnerability to unnoticed or unknown experimental confounds falls short of calling the validity of NE research into question. Yet, the point remains that many NEs fail to control for "confounds that are known in

the experimental economics literature" (Harrison, 2008a, p.311). Moreover, ambiguities and inaccuracies in task descriptions occasionally undermine the informativeness of the reported findings (Gold and Buckner, 2002). By way of illustration, consider the study of McClure et al. (2004a), where subjects were asked to make choices between immediate and distant rewards. As noted by Ortmann (2008, p.440), the contrast those authors draw between 'immediate' versus 'delayed' rewards is misleading, as the rewards considered in their experiment are all 'delayed' as long as the monitored subjects lie in the scanner.

Faced with these criticisms, a proponent of NE may rebut that progress in scanner technology and experimental design will provide us with more detailed information concerning the neural substrates of decision making. Yet, one should not overemphasize the advantages derivable from those advances. Let me expand on this point. Various limitations in NE studies will be alleviated or overcome thanks to progress in scanner technology and experimental design. For example, NEs can address the concerns related to their reliance on small experimental samples by replicating previous studies with other experimental populations. Similarly, the spatial and temporal resolution of current brain-imaging instruments will soon increase thanks to technological progress. On the contrary, some evidential and epistemological concerns are unlikely to abate with technological advances. By way of illustration, think of the many inferential steps and auxiliary assumptions required to interpret neural data.

One might object that evidence is frequently procured and mediated by means of sophisticated instruments in other scientific disciplines as well (Bechtel and

Stufflebeam, 2001, p.55). Yet, NEs' studies involve a remarkable number of interventions and interpretative steps, which render the subsequent reports very sensitive to the experimental setup. In this respect, the availability of a rapidly improving technology by no means guarantees that NEs will soon provide clearly interpretable insights into the neural substrates of choice behaviour.

In such a context, a further question arises as to whether neuroscience is sufficiently mature to provide NEs with an accurate and reliable basis for informing the economic account of decision making. This concern can be explicated as follows. Experimental findings are getting accumulated, refined and not rarely disconfirmed at a considerably high pace in the neuroscientific literature (see e.g. Vul et al., 2009). These rapid advances do speak in favour of the progressive character of research programs such as neurobiology and computational neuroscience. At the same time, they make it plausible to expect that our currently best neuro-psychological theories will be revised in the near future. In this respect, one wonders why exactly economists should construct their models on the basis of findings and conjectures that are likely to be corrected within *a few* years.

These concerns resemble those raised by various philosophers of science (e.g. Laudan, 1981, and Stanford, 2010), who questioned the plausibility of realistic interpretations of scientific theories on the basis of the profound theory changes occurred in the history of several disciplines. In the words of Laudan (1981, p.47), scientific realism "cannot, even by its own lights, explain the success of those many theories whose central terms have evidently not referred and whose theoretical laws and mechanisms were not

approximately true" (see Poincaré, 1905, van Fraassen, 1980, and Worrall, 1989, for a critical discussion). *Mutatis mutandis*, the worry regarding NE research is that the rapid theory changes occurred in the history of cognitive and computational neuroscience call into question NEs' currently best available neuroscientific theories.

CHAPTER FOUR

ARGUMENT FROM MODELLING TRADE-OFFS

As we have seen in *chapter two*, NEs frequently maintain that economists can improve their models with regard to several desiderata by incorporating neural insights. Their contentions, however, have not been widely accepted by economists, who frequently oppose integrating other disciplines' findings. In particular, most economists remain quite sceptical about NEs' attempts to implement a neural enrichment of the economic theory of choice. As Rubinstein (2008, p.486) provocatively puts it, suppose "we are able to map all brains onto a canonical brain. The functions of the different areas of the brain are crystal clear to us. The machines used in experiments are cheap enough that thousands of subjects can be experimented on. And finally the data is clear and double-checked. The question would still remain: what is the potential role of brain studies in Economics?".

In this context, two issues of great methodological significance deserve to be discussed. Firstly, there is the question why economists are typically unwilling to include neuro-psychological insights into their models. And secondly, one wonders whether economists are justified in adopting such an isolationist attitude. In this chapter, I address these two issues in turn and critically assess NEs' calls to incorporate neural insights into economic models of choice. In particular, I argue that NEs can improve economic models with regard to individual modelling desiderata, yet are unlikely to provide models which supersede the ones employed by economists.

The contents are organized as follows. Section 4.A aims to show that NEs overemphasize the extent to which their contributions improve economic models with regard to specific desiderata. In the other two sections, I focus on the trade-offs between the desiderata that NEs and other economists respectively value, devoting particular attention to the tractability of economic models. In section 4.B, I reconstruct and critique the argument from tractability by means of which economists often resist the incorporation of neural insights. In section 4.C, I articulate and defend a refined argument from tractability which attempts to show that economists are provisionally justified in retaining a methodologically distinctive approach to the modelling of decision making.

Before proceeding, let me provide two preliminary remarks regarding the focus of this chapter. Firstly, I shall consider not just economists' modelling practices, but also the pragmatic and epistemic goals which govern the construction and evaluation of their models. And secondly, I shall predominantly concentrate on neural - rather than psychological - methods and findings. This choice is motivated by the significance that some recent advances at the interface between economics and neuroscience have for the issues I address in this chapter. However, in various places I shall refer to the literature on modelling in other behavioural sciences and highlight some parallel concerns which arise regarding the integration of these disciplines' insights into economic models.

4.A LOCAL AND GLOBAL MODELLING IMPROVEMENTS

The proponents of NE claim that integrating neural insights can improve economic models of choice in a number of respects. In this section, I critically assess NEs' contentions regarding the modelling desiderata we examined in *chapter two*. In doing so, I argue that NEs overestimate the extent to which including neuro-psychological findings improves economists' models. More specifically, I distinguish between *local improvements* - which obtain when a model is improved with regard to some specific desideratum, individually considered - and *global improvements* - which occur when the trade-off between different desiderata is more satisfactorily resolved. I then illustrate that NEs have failed to show that a neural enrichment yields global - as opposed to local - improvements in economic models of decision making.

1) Descriptive accuracy

A model can be regarded as descriptively accurate both if it includes most of the properties or features of its target system and if it provides a detailed characterization of those properties or features (Weisberg, 2007a). According to various NEs, incorporating neuro-psychological insights helps economists to increase the descriptive accuracy of their models in both of these respects (see section 2.A). The idea is that neurally enriched models include a larger subset of the factors influencing people's decisions and posit agents having more plausible cognitive and computational abilities. As I argue below, however, NEs have hitherto failed to substantiate the claim that economists

should include several neural insights into their models. To see this, let us consider the previous two aspects of descriptive accuracy in turn.

To justify their focus on a subset of the variables involved in the investigated choice settings, economists may put forward the following reasoning: "Most economic choices result from the interplay of a high number of causal factors. For this reason, we typically have to include in our models just a small subset of the elements which operate in the examined choice settings. Hence, it is not because of some form of neurophobia that we resist a neural enrichment of our models, but rather because the very act of modelling requires us to elaborate simplified representations of our target systems".

The thought is that a descriptively accurate representation of economists' target systems would usually include so many variables that it would fail to be intelligible or informative. In other words, given that economic phenomena and agents are "too complex to be tractable targets for direct examination", economic models cannot be exact replicas of their target systems, but have to resemble those systems only "in certain respects and to certain degrees" (Mäki, 2005, p.304; see also Hausman, 1992). As Mäki puts it, "on suitable conceptual specifications, a *realist* economist […] is obliged to employ assumptions that are *unrealistic* in many senses of the word" (1992, p.319; see also Mäki, 1994, ch.12).

Analogous considerations apply to the NEs' complaint that economists model agents having implausible cognitive and computational abilities. By way of illustration, consider the claim of Camerer, Loewenstein and Prelec (2005, p.32-3) that even though

economists typically represent agents having stable preferences, people's preferences are subject to widespread intertemporal and across-contexts variations. Economists' reliance on models positing stable preferences is motivated - not so much by their alleged ignorance of preferences' variability, but rather - by considerations of analytical convenience. Moreover, the fact that standard economic theory takes preferences as exogenously fixed does not prevent the construction of models which allow for preference change (see e.g. Dietrich and List, 2011; see also Loomes and Sugden, 1995, on so-called random preference models, where the value of various parameters such as risk aversion is randomly determined at every period).

To give another example, several NEs take people's cognitive and computational limitations to cast doubt on standard economic theory. In particular, some contend that since both intentions and actions are caused by "prior neural events which are inaccessible to consciousness", people often fail to understand the reasons motivating their choices and make decisions which violate the normative axioms of standard economic theory (Camerer, Loewenstein and Prelec, 2005, p.31). Neuro-physiological studies provide informative insights concerning the genesis of conscious experiences and the purported dependence of agents' decisions on neuro-biological determinants (see e.g. Hohwy, 2007, Lloyd, 2002, and Roskies, 2006)[41]. Even so, the mere fact that we often lack awareness as to when or why we make a decision does not *per se* render our choices more prone to violate the axioms of standard economic theory. In particular, one can model an agent's decisions as the solution to a constrained maximization

[41] See also Libet (1983, 1985, and 1996) for some pioneering studies of the neural antecedents of choices. In his view, "even a fully voluntary act is initiated unconsciously […] Cerebral neural activity […] precedes the subject's awareness of his/her intention or wish to act […] Our sense of agency is apparently illusory" (1996, p.95; see e.g. O'Connor, 2009, for a critical evaluation).

problem irrespective of whether the agent possesses the cognitive resources to model her decisions in those terms (see e.g. Bradley, forthcoming, and Ross, 2008b).

In such a context, the more general question arises as to how economists' reliance on descriptively inaccurate assumptions bears on the merits of their models of choice. By way of illustration, consider the idealizing assumption that a perfectly rational agent updates her beliefs in light of novel information instantly and with negligible cognitive costs. This assumption falls short of providing a plausible characterization of human individuals' cognitive performance (Conlisk, 1996). Even so, it offers a sufficiently accurate approximation of people's behaviour in several decision settings (e.g. think of repetitive and simple choices).

Now, it would be excessive to endorse Friedman's (1953) instrumentalist motto that the realism of a model's assumptions is irrelevant to its explanatory and predictive performance (Musgrave, 1981, and Nagel, 1963). Still, the point remains that even economic models resting on descriptively inaccurate presuppositions can be explanatory and predictive (Gruene-Yanoff, 2009). Indeed, economic modellers often deliberately build models which fail to accurately represent the phenomena of interest (see e.g. Mäki, 2002, on models postulating fictitious entities). In the words of Matthewson and Weisberg (2009, p.182), many models aim not so much to resemble any real target system, but rather to "canvass possibility space", as "sometimes exploration of the non-actual helps explain the actual" (see also Wimsatt, 1987, and Odenbaugh, 2005).

To be clear, it is true that - other things being equal - models which provide descriptively accurate characterizations of the investigated target systems are often preferred to models which fail to do so. Still, as I argue in section 4.C, descriptive accuracy often comes at the cost of other modelling desiderata such as tractability. In this respect, many NEs appear to overlook that the possibility of modelling people's decisions without having to incorporate neuro-psychological findings constitutes a strength - not a limit - of standard economic theory. Regrettably, NEs often gloss over this issue and the implications it has for the evaluation of standard economic models of choice.

To render the above remarks more vivid, let us distinguish two ways in which a model can resemble its target system (see Glennan, 2005; see also Matthewson and Calcott, 2011). On the one hand, economic models of choice are meant to be *behaviourally similar* to their target systems in the sense that, given the same inputs (e.g. specific choice problems), they approximate their target systems' outputs (e.g. choices). On the other hand, many NE models aim to be *mechanically similar* to their target systems in the sense of providing descriptively accurate representations of the mechanistic underpinnings of those systems. Now, a model can be behaviourally similar to its target system without being mechanically similar to it. Moreover, increasing the behavioural similarity of a model to its target system does not necessarily require one to increase the model's mechanistic similarity as well. That is to say, economists may succeed in increasing the descriptive accuracy of their models of choice without including any neuro-psychological insight into those models.

2) Predictive power

NEs often claim (e.g. Camerer, 2007, C28) that neuro-psychological findings enable economists to construct more predictive models[42]. *Prima facie*, one may expect NEs' insistence on predictive power to be welcomed by many economists. After all - the thought would be - economists of all stripes ascribe a prominent relevance to predictive considerations. Still, economists *qua* economists are not concerned with every observable implication of their models, but only with those predictions which relate to the phenomena they investigate (see e.g. Friedman, 1953, p.8 and 30; see also Hausman, 2008b). In what follows, I consider two lines of argument which cast doubt on whether constructing more predictive models requires economists to acquire detailed knowledge of the neuro-psychological underpinnings of choice behaviour.

The first reasoning points to the fact that economists can often improve their models' predictions simply by examining the *behavioural correlates* of the neural processes investigated by NEs. The thought is that even when differential activations in specific neural areas correlate with variations in observed decisions, choice behaviour data "screen off the neural details" (Kuorikoski and Ylikoski, 2010, p.223; see also Bernheim, 2009). In this respect, the further concern arises that neural data are often more difficult to obtain precisely in those situations when also the economic data of interest are missing or inaccurate. To put it differently, how likely is it that we can

[42] As specified in section 2.B, I use the expression "predictive power" to refer to both predictive accuracy, which denotes the exactness of a model's observable implications regarding the investigated phenomena, and predictive reliability, which relates to the stability of a model's predictive performance across distinct choice contexts.

observe people's brain activity in detail and yet cannot acquire data about their choices and preferences?

The second reason to doubt that building more predictive models requires economists to import neural data relates to the existence of *multiple levels of description* of human choice behaviour. The reasoning goes as follows. People's decisions can be accounted for in terms of social, psychological, neural etc. mechanisms (e.g. Bechtel, 2008, and Craver, 2007). Now, economists can certainly succeed in constructing more predictive models by incorporating causal and mechanistic insights. Even so, additional argument is needed to substantiate the claim that neural - as opposed to social, psychological, etc. - data should be included into economic models. That is to say, one may concede that economists can formulate more predictive models by adopting a mechanistic approach, and yet resist the claim that the most informative mechanistic insights are acquired at the neural level.

A proponent of NE might advocate the neural enrichment of economic models on the alleged ground that neural insights provide economists with more predictive benefits than other disciplines' findings. Yet, there are at least two reasons to dispute this claim. Firstly, it is doubtful that monitoring people's neural activation patterns yields predictive benefits over sufficiently wide *temporal horizons* to be useful for the economists' purposes. After all, most NEs' predictions apply "to rapid transients in spike rate in the 50-250 millisecond range" and do not extend to the timescales required to inform economists' models and policy analyses (Montague, 2007a, p.223). And secondly, one wonders how *generalizable* the insights offered by specific NE

experiments are. For instance, think of the NE model proposed by Vercoe and Zak (2010) that we examined in section 2.B. As this model suggests, some NE studies provide valuable predictive insights regarding specific economic decisions. Still, it remains hard to see how exactly the results obtained in tightly controlled experimental settings would generalize to the rest of economic theory.

3) Model selection

During a trip to Murano (Italy), you visit a glass craftsman in order to buy a present for your *fiancé*. As it happens, you are holding an expensive colourful vase of glass, and the vase falls out of your hand. There is a wide variety of models by means of which your attempting to catch the vase before it breaks on the ground may be represented. For example, your action may be modelled as if you were trying to minimize purely monetary losses (money-maximizing-agent), the acoustic noise resulting from the vase's destruction (silence-loving-agent), the number of items which will predictably lie on the shop's floor (mereologically-parsimonious-agent), the physical inertia of your muscles (fitness-obsessed-agent), and so on.

As this example suggests, observed choice behaviour can often be modelled in terms of many different constrained utility functions. This, however, does not imply that any such model accounts for observed decisions in plausible terms. On what basis are we to discriminate between competing economic models that are equally compatible with observed choices? Several NEs (e.g. Camerer, 2008a, p.47, and 2008b, p.370) regard neuro-cognitive and neuro-computational plausibility as suitable criteria for model

selection in economics. The thought is to discriminate between distinct economic models of choice in terms of their relative fit with the available neuroscientific evidence concerning decision making processes. Regrettably, neuro-physiological evidence alone rarely enables economists to discriminate between competing models of choice.

To be clear, the mere fact that the available evidence is compatible with dissimilar models does not exclude the existence of convincing reasons to accept one of these models over its competitors (Laudan, 1990, p.270). Indeed, one can find several episodes in the history of science where scientists opted for specific theories even in presence of rival theories that also seemed well supported by the available evidence (see e.g. Okasha, 2000, on the controversy between the Ptolemaic and the Copernican systems, and Zahar, 1973, on the transition between Newtonian and relativistic mechanics). Even so, many NEs appear to overemphasize the extent to which neural evidence discriminates between competing economic models (see Bernheim 2009, sec.1, for a similar remark). To render this point more vivid, let us consider economists' multiple-selves models, which represent people's decisions as the outcome of the interactions of sub-personal entities and processes (see e.g. Ainslie, 1992).

In the economic literature, various kinds of multiple-selves models have been proposed. For example, some of these models represent choice behaviour as the solution of a bargaining game among multiple sub-agents with conflicting objectives (e.g. Schelling, 1978 and 1980). Others, instead, model choices as the sequential equilibrium of a signaling game between multiple selves each controlling the person's behavior in distinct temporal intervals (e.g. Benabou and Tirole, 2003; Prelec and Bodner, 2003).

The proponents of NE take neuro-psychological data to help economists test their multiple-selves models and ground them in neuro-psychological detail. The idea is that brain-imaging data enable economists to associate the sub-personal entities posited by multiple-selves models with specific mental processes or neural areas (Spiegler, 2008, p.520; see also Rustichini, 2009, p.53).

Unfortunately, NEs appear to overlook that standard economic theory does not presuppose that multiple-selves models provide descriptively accurate representations of the neuro-psychological processes underlying choice behaviour (Harrison, 2008a, p.38-9; see also Ross, 2008c). Indeed, it is an open question whether the sub-personal entities postulated by multiple-selves models are more aptly related to psychological - as opposed to neural - entities and processes (Vromen, 2010a, p.27). For instance, Brocas and Carrillo (2008) model intertemporal choices as the result of a conflict between two neural systems that respectively target immediate and delayed rewards. For their part, Frederick, Loewenstein and O'Donoghue argue that modelling behaviour as the interplay of disparate psychological motives enables economists to account for intertemporal choices in terms of more "legitimate" and "stable" constructs (2002, p.393).

Now, one looks with sympathy at the Feyerabendian spirit of those authors who, deeming neuroscience research to be "necessarily speculative" (Camerer 2008b, p.369), advocate testing any sort of neural conjecture. Even so, it remains unclear why exactly economists should discriminate between economic models of choice in terms of their neuro-cognitive and neuro-computational plausibility. After all, those models are not

meant to characterize the activation patterns of specific neural areas or the workings of the human neural architecture. Moreover, the mere fact that economists occasionally face severe problems of evidential underconstraint does not imply that their models should be tested against every kind of evidence. As Mäki puts it, "neurobiological data would be one obvious candidate, *among others*, to be given a role in deciding between observationally equivalent models of choice" (2010, p.115).

At this stage, the proponent of NE may object that economists cannot remain entirely agnostic regarding the neural substrates of choice behaviour on the alleged ground that the available neuro-psychological evidence narrows down the set of economic models compatible with it (Bernheim, 2009, p.17; see also Vromen, 2010b, p.172). This reasoning, however, does not imply that neuro-psychological evidence should be employed to discriminate between competing economic models of choice. To be sure, it is certainly desirable that economists build models whose observable implications are consistent with the findings collected in other behavioural sciences. Yet, this does not *per se* license the demand that economists construct models of choice that are to be "tested simultaneously at the neural, psychological, and economic levels of analysis" (Glimcher, 2010, p.132). For requiring economists *qua* economists to simultaneously account for economic, psychological and neural data places an unreasonably heavy evidential burden on them. In the words of Sunder (2006, p.340), "all sciences must make some assumptions about the phenomena at the level of details they do not wish to delve into […] Insistence that all such assumptions […] be descriptively valid creates a burden that is both unreasonable as well as unproductive" (see also Gul and Pesendorfer, 2008, for similar remarks).

4) Explanatory insightfulness

Various proponents of NE (e.g. McCabe, 2003a and 2003b) argue that neurally enriched models provide explanatorily informative insights regarding choice behaviour. Now, what insights prove to be explanatory for a modeller typically depends on a number of factors, including what constructs she employs and her modelling purposes. To render this point more vivid, let me put forward the following analogy. Suppose that a technologically illiterate philosopher wanted to understand why her brand-new computer crashed on a given occasion. Providing her with a complicated informatic explanation would hardly help her to make sense of her computer's malfunctioning. One might object that such an account does offer an adequate explanation of the computer's breakdown, and that the only reason why the philosopher does not find it informative is because she lacks proper training in informatics. Yet, the point remains that a full-blown informatic explanation is unnecessarily complicated for the philosopher's purposes. As Putnam (1975, p.94) memorably puts it, one does not need a micro-physical explanation to understand why a square peg slightly less than one inch high passes through a one inch high square hole but not through a one inch in diameter round hole.

Regrettably, few NEs seem aware of the variability that criteria of explanatory relevance exhibit across disciplinary boundaries. In particular, several NEs appear to presuppose that understanding how decision making is instantiated at lower levels of description is *ipso facto* explanatorily informative for economists. For instance, after noting that "all economic activity flows through the brain at some point", Camerer

contends that “it is hard to imagine that understanding brain function could not be useful for understanding some aspects of economic choice” (2008a, p.47). This claim does not license the conclusion that economists will find neural insights particularly informative. After all, there is a sense in which all economic activity flows through genes, gluons, etc. at some point. Yet, few NEs would conclude that accounting for choice behaviour requires economists to engage in genetic speculations and hyper-detailed particle physics. Moreover, the point remains that an explanation “is not necessarily improved when the explanans is itself explained” (Kuorikoski and Ylikoski, 2010, p.221). Indeed, a lower-level explanation can even be less explanatory than higher-level ones (see e.g. Weslake, 2010).

To see this, suppose that we were able to provide an accurate description of the neural substrates of agents’ behaviour in a specific choice setting. Such a description, however accurate, may fail to be explanatorily informative for economists. For instance, it may be exceedingly complex, or it may bear no discernible relation to economists’ traditional explanatory goals. In other words, even if NE yields explanatory insights beyond those provided by its parent disciplines, it is an open question whether NE insights will be explanatorily informative for the practitioners of those disciplines. Philosophers of science have provided various remarks along these lines with regard to several disciplines. For instance, Woodward (2003, p.232-3) argues that macroscopic explanations of phenomena can yield information besides that provided by microscopic accounts. Similarly, Weslake (2010, p.290) alleges that explanations can be improved by abstracting away from the causal details of the investigated target systems.

More generally, the point remains that NEs frequently appear to overestimate the explanatory reach of their accounts of choice behaviour. By way of illustration, consider McCabe's assertion (2008, p.348) that NE can help economists to better "understand the disparity of economic growth, and material welfare, both between and within nations". One may tell a story of how neural processes happen to influence specific macroeconomic phenomena by shaping individuals' preferences and actions. Still, when it comes to accounting for macroeconomic growth and inequality, most NE research appears to have marginal relevance. For the correspondences between individual brains' activations and those macroeconomic variables are simultaneously shaped by factors - such as social norms, institutional incentives, and people's axiological commitments - whose influence can be only partly controlled in single NE studies (see section 3.B).

At this stage, a proponent of NE may advocate the neural enrichment of economic theory by appealing to the *causal influence* of neural processes on people's preferences and decisions. In the words of Camerer, Loewenstein and Prelec (2005, p.27), "the traditional economic account of behavior, which assumes that humans act so as to maximally satisfy their preferences, starts in the middle […] of the neuroscience account". Now, it would be implausible to dispute that people's preferences and actions are causally influenced by neural processes and events to which they frequently lack conscious access (Camerer, Loewenstein and Prelec, 2005, p.31; see also Libet, 1965, 1983, 1985, and 1996). Even so, the mere fact that neuro-physiological processes causally influence people's choices does not license the conclusion that providing an adequate explanation of economic phenomena requires one to refer to neural mechanisms or processes.

The proponent of NE might retort that neuro-physiological findings are more informative than biological and psychological ones on the alleged ground that they enable economists to answer more questions regarding hypothetical and counterfactual variations in people's choice behaviour. Regrettably, NEs have hitherto failed to substantiate this assertion. Moreover, some claims to the contrary have been put forward in the literature. For instance, Kuorikoski and Ylikoski (2010, p.223) argue that the range of what if questions that can be answered on the basis of the hitherto identified correspondences between neural areas' activations and observed decisions is "very limited". In their view, modelling agents' choices in psychological terms enables one to answer a "broader range of what if -questions concerning possible alterations in the agent's valuations, knowledge and how the relevant information is presented".

5) Welfare analyses

The *argument from wants / likes divergences* we presented in section 2.E challenges economic welfare analyses by criticizing economists' focus on agents' observed choices. Now, it would be implausible to deny the existence of divergences between people's wants and likes. Even so, it remains doubtful that these divergences challenge the validity of economic welfare analyses. Let me expand on this point. NEs often criticize economists for regarding choice behaviour as a search for pleasure and for assuming that economic agents only strive to obtain what they like (e.g. Camerer, Loewenstein and Prelec, 2005, p.37). These criticisms apply to the conception of behaviour advocated by some early neoclassical economists (e.g. Jevons, 1871). Yet, it

is at least since the Thirties that economists have been reiterating that their theories are "not to be identified with the psychology of the utilitarians, in which pleasure had a dominating position" and that our main motivations for action may not relate to our own or other people's pleasure (Ramsey, 1931, p.173; see also Hicks and Allen, 1934). That is to say, economists realized long ago that economic theory can be articulated in non-hedonistic terms and does not rest on any specific assumption as to why people choose the options they choose (Robbins, 1935, p.93; see also Binmore, 2008)[43].

A proponent of NE may rebut that providing compelling insights concerning people's well-being requires economists to base their analyses on people's likes. Even so, it appears that whether agents obtain what they want is an important constituent of well-being irrespective of whether agents happen to like the things they wanted. For instance, as noted by Sugden (2006, p.217), people often consciously act on preferences that are not stable under experience and reflection, yet attribute a high importance to the opportunity of satisfying those preferences. More generally, the point remains that NEs and other economists respectively endorse fundamentally different approaches to welfare analyses. Let me explicate this contrast. On the one hand, standard economic theory does not rest on specific presuppositions regarding what constitutes agents' objective well-being or what objectives agents should pursue. On the other hand, many NEs relate agents' well-being to their mental states or the activation patterns of particular neural areas. Moreover, they often aim to establish what objectives agents should pursue in specific situations. As I argue in section 7.C, these differences constrain the relevance of NEs' findings for traditional economic welfare analyses.

[43] Indeed, it is since the time of Pareto that economists remark that "it is not an essential characteristic of [economic] theories that a man choosing between two sensations chooses the most agreeable" (1909, ch.3, §11).

To be sure, NEs can provide economists with valuable information regarding the neuro-psychological substrates of choice behaviour, which promises to complement economists' policy evaluations and indexes of well-being. Even so, it remains hard to see how exactly NE descriptive findings are supposed to provide compelling insights regarding what people ought to choose in a given situation. For economists' normative and prescriptive analyses rest - not so much on "empirical hypotheses about how human beings really think and act", but rather - on deductions from a priori assumptions about rational choice (Bruni and Sugden, 2007, p.146-7; see also Knutson and Peterson, 2005, p.305). In the words of Glimcher (2010, p.412), using neural measurements to draw normative conclusions regarding people's well-being seems "unwarranted" not just because of current limitations in our measurements but also because neural data "are explicitly positive in nature"[44].

To recapitulate, the proponents of NE appear to overestimate the extent to which including neural insights helps economists to improve their models of choice. NEs may rebut that advances in scanner technology and experimental practices will enable them to construct models of choice which supersede the ones currently employed by economists. As I argue in the next two sections, however, there are principled reasons to doubt this claim. My reasoning can be summarized as follows. Economists and NEs

[44] A similar point was made by Pareto (1971 [1909], Ch.1, §21-26) in explicating the distinction between pure and applied economics. In his view, applied economics is a "practice" which can be informed fruitfully by the findings of sociology and psychology. Pure economics, instead, is "the science of logical action", and should remain separate from psychology and sociology. See also Hume (2005 [1740], BIII, 1.1) for some famous remarks against deriving prescriptive or normative conclusions from purely descriptive premises.

respectively value dissimilar modelling desiderata, which often make contrasting demands on modellers. The trade-offs between distinct desiderata, in turn, significantly limit the incorporation of neural insights into economic models of choice. To put it differently, NEs persuasively illustrate that including neural insights can foster local improvements in economic models of choice. Yet, they usually gloss over the issue whether a neural enrichment is likely to yield global improvements in economic models. Hence, it remains an open question whether economists will often find it convenient to incorporate neural insights into their models of choice.

To be fair, assessing whether some modelling modification constitutes a global improvement can prove to be quite difficult due to disagreements regarding how the examined desiderata are to be defined and weighed against each other. How one addresses these two issues, in turn, significantly depends on several elements, ranging from particular disciplinary conventions to the aims of modellers. At any rate, showing that including neural insights yields some local improvements for economists is insufficient to substantiate the case for adopting a neurally informed approach to the modelling of decision making. For these local improvements may fail to compensate for the modelling costs involved in a neural enrichment of economic theory. To render this point more vivid, in the next two sections I examine some trade-offs between specific modelling desiderata and illustrate how these trade-offs constrain the incorporation of neural insights into economic models of choice.

4.B THE ECONOMISTS' ARGUMENT FROM TRACTABILITY

In this section, I reconstruct the *argument from tractability* which motivates many economists' reluctance to incorporate neural insights and then argue that two major flaws prevent it from cogently vindicating their position. Before doing so, let me put forward some definitional caveats. Economists frequently cite tractability as a virtue of their models, yet rarely define it in precise terms. The literature on economic modelling nicely illustrates the difficulty inherent in providing an uncontroversial characterization of this desideratum. For instance, some authors (e.g. Kahneman, 2003, p.166) relate tractability to the number of variables which appear in a given model. Others (e.g. Gabaix and Laibson, 2008, p.294, and Hindriks, 2005, p.392, and 2006, p.413) define it in terms of the availability of computable analytical solutions. Still others (e.g. Gibbard and Varian, 1978, p.673) take tractability to be akin to the simplicity of a model's mathematical formalism.

In light of these discrepancies, one may think that tractability is best understood as a cluster concept resembling notions such as parsimony, resolvability and simplicity, which are themselves hard to define unambiguously. This, however, does not prevent us from adopting a sufficiently precise characterization of such a desideratum. In what follows, I regard the number of variables appearing in a model as an approximate indicator of its tractability, as this provides us with a convenient rule of thumb for comparing alternative modelling frameworks[45].

[45] The notion of tractability could be given a more precise definition in specific modelling contexts. I do not engage with this definitional issue here, as my aim is to provide an approximate indicator of tractability which is both sufficiently precise to assess NE models and sufficiently general to capture the heterogeneity of economists' terminological practice.

In everyday discussions, economic practitioners and methodologists often speak of tractability as a fundamental modelling desideratum. Nonetheless, they seem to ascribe different meanings to such an expression. In this respect, it is instructive to distinguish three progressively less stringent senses in which the term "fundamental" may be employed, namely: that economists give tractability lexicographic priority over other modelling virtues (*fundamental-1*); that economists attach a higher weight to tractability than to any other modelling desideratum (*fundamental-2*); and that in spite of their preference for descriptively accurate and predictively powerful models, economists are reluctant to reduce tractability below some minimal threshold (*fundamental-3*). Now, the first two claims are clearly untenable, and fail to fit with a number of episodes in the history of economic theory[46]. The third assertion, instead, accurately reflects the fact that many economists - despite valuing attributes such as descriptive accuracy and predictive power - take these desiderata into account under the overarching constraint that their models remain sufficiently tractable. In what follows, I shall employ the term "fundamental" in the *fundamental-3* sense unless stated otherwise.

Having said that, the economists' argument from tractability can be articulated as follows:

P.1 Economists ascribe fundamental importance to the tractability of their models of choice.

P.2 Building a tractable model of choice prevents one from including a large number of variables in it.

[46] For instance, consider how the predictive gains offered by generalized expected utility models led many economists to alter the more tractable expected utility framework (see e.g. Starmer, 2000, for a detailed review).

C (1) Due to the fundamental importance they ascribe to tractability, economists are prevented from including a large number of variables into their models of choice.

P.3 Fostering a neurally informed revolution in the economic theory of choice requires NEs to include a large number of neural variables into the economists' models.

C NEs will not foster a neurally informed revolution in the economic theory of choice.

Let us examine each step of the argument from tractability in turn. *Premise 1* emphasizes the great importance that economic modellers attribute to tractability. *Premise 2*, in turn, notes the limitations that tractability imposes on incorporating variables into economic models. On the basis of these premises, *conclusion (1)* asserts that the high relevance that economists ascribe to tractability significantly constrains the integration of psychological, neural, biological, etc. constructs into their models. That is to say, importing neural insights may well help economists to improve their models with regard to specific attributes (e.g. think of descriptive accuracy). Yet, there is frequently a trade-off between tractability and other desiderata, and economists' preference for tractability limits the degree to which their models can satisfy them.

The existence and significance of trade-offs between distinct desiderata have been discussed at length in other disciplines besides economics. For example, in the literature on modelling in biology several authors (e.g. Odenbaugh, 2003, and Weisberg, 2004, 2007a and 2007b) argue that different attributes make contrasting demands on

modellers, imposing considerable pragmatic and logical constraints on model construction. To be sure, economists and biologists alike hold diverse positions regarding the definition and the relative importance of particular desiderata[47]. Still, in spite of these discrepancies, both economists and biologists share a relatively precise characterization of various attributes and concur on the existence of specific trade-offs between them. Indeed, some authors even propose quantitative frameworks for weighing distinct desiderata against each other (see e.g. Harless and Camerer, 1994, for a comparison of traditional and generalized expected utility models, and Moscati, 2006, for a similar analysis concerning different versions of the neoclassical theory of demand). Regrettably, not all NEs seem aware of the trade-offs which hold between dissimilar attributes, and various researchers appear to underestimate the constraints that these trade-offs impose on model construction in economics[48].

In advocating the neural enrichment of economic theory, several authors (e.g. Camerer, Loewenstein and Prelec, 2005, p.10, and Rustichini, 2003) speak of introducing revolutionary changes into the economists' account of decision making. In rebuttal to their claims, *premise 3* of the argument from tractability states that if NEs are to foster such a neurally informed revolution, then they have to include several neural variables into the economists' models. As we shall see below, the cogency of this premise rests on how exactly the notion of revolution is interpreted. For now, let me anticipate that there are different ways in which NE research can promote radical modifications in the

[47] See e.g. Mäki, 1988, 1990 and 1992, on different senses in which the terms "realism" and "realisticness" are used in economics; see Weisberg, 2007a, and Matthewson and Weisberg, 2009, on distinct notions of "generality" in biology.

[48] For example, as noted by Glimcher, Dorris and Bayer (2005, p.214), several NEs aim to build neurally enriched models that are highly parsimonious and predictive in a wide range of decision contexts. Parsimony and predictive power so frequently pull in contrasting directions that it is an open question whether NEs will manage to accomplish such an ambitious goal.

economic theory of choice. In this respect, the question arises as to whether revolutionizing such a theory requires NEs to integrate many variables into the economists' models.

Finally, the argument from tractability concludes that the considerable extent to which tractability constrains the import of other disciplines' constructs precludes a neurally informed revolution of the economic account of decision making. The idea is that NEs may provide economists with informative insights, yet will fail to elaborate models which supersede the ones currently employed by economists. In this perspective, what Kahneman (2003, p.166) asserted in relation to behavioural economics seems equally pertinent to NE research. In his view, "the constraint of tractability can be satisfied with somewhat more complex models, but the number of parameters that can be added is small". For this reason, enriched models "cannot stray too far from the original set of assumptions", and "theoretical innovations […] may be destined to be noncumulative"[49].

Now, let us assess the cogency of the argument from tractability. *Prima facie*, such reasoning might seem to provide economists with a compelling rationale for resisting the integration of neural insights into their models. As I argue below, however, two crucial flaws (see points *i* and *ii*) prevent it from constituting a convincing defense of such a conservative position. My first criticism targets the argument's characterization of the relationship between tractability and other modelling desiderata. The second one

[49] Kahneman is far from being a strenuous defender of the traditional economic theory of choice. In fact, he was awarded the 2002 Nobel Prize in Economics "for having integrated insights from psychological research into economic science" (Nobel Press Release, 2002).

casts doubt on the third premise of such an argument. Let us examine each of these criticisms in turn.

i) As stated by *premise 1* of the economists' argument, economic modellers attribute a prominent significance to tractability. This, however, does not *per se* license the claim that they are justified in doing so. To address this concern, economists may attempt to vindicate their reliance on tractable models on pragmatic grounds. In particular, they may point out that the very act of modelling requires one to elaborate simplified representations of the investigated phenomena[50]. The decision concerning which properties or features of the phenomena of interest are to be represented in a model is frequently guided both by specific inclusion rules (see e.g. Weisberg, 2007a) and by the availability of representational techniques such as abstractions, isolations and idealizations. In this latter respect, the issue for economists is not so much whether to use abstractions, isolations and idealizations, but rather which of these it is most convenient to implement in a given modelling context. As Mäki (1992, p.1) aptly puts it: "Faced with the essential complexity of the world, every science is compelled to employ methods of modifying or deforming it so as to make it or the image of it theoretically manageable and comprehensible. Economics is no exception in this regard"[51].

[50] Various authors offer less convincing accounts of economists' predilection for tractable models. To give one example, consider the observation by Gabaix and Laibson (2008, p.295) that economists elevated modelling attributes such as tractability out of people's tendency to "celebrate the things they do best". Invoking this psychological propensity hardly accounts for the great importance economists attach to tractability. In particular, it fails to substantiate the claim that the reason why economists value this desideratum is because of their excellence at building tractable models.

[51] Various kinds of abstractions, isolations and idealizations can be differentiated. For instance, Weisberg (2007a) distinguishes between so-called Galilean and minimalist idealizations as follows. Galilean idealizations consist in simplifying the characterization of a target system to

These considerations persuasively elucidate the pragmatic rationale for why many economists rely on tractable models. In response to those claims, a proponent of NE may concede that it is not feasible to include a large number of neural variables into a single economic model. At the same time, she may wonder in what modelling contexts and to what extent tractability can be legitimately given priority over other desiderata. After all - the reasoning would go - it is true that the trade-offs between distinct attributes often need to be resolved on a case-to-case basis, depending on what choice problem is modelled and the purposes of the modeller. Yet, the argument from tractability does not provide any precise criterion for dealing with those trade-offs. In particular, *conclusion (1)* appears to presuppose - rather than show - that tractability considerations alone outweigh the relevance of the modelling benefits (e.g. increased descriptive accuracy) that economists may derive by importing neural insights.

ii) As anticipated above, one may question the cogency of *premise 3* of the argument from tractability by pointing at distinct ways in which NEs could promote revolutionary changes in the economic account of decision making. To render this point more vivid, let me contrast the following two scenarios. On the one hand, NEs may foster revolutionary modifications in the economic theory of choice by introducing into economic models several variables that were not previously considered by economists[52]. On the other hand, NEs may revolutionize economic theory from the foundations upwards by providing economists with a novel theoretical framework. In this latter

make it tractable and then systematically de-idealizing it so as to obtain increasingly accurate representations. Minimalist idealizations, instead, obtain when one includes in her model only the most important causal factors underlying the investigated phenomena. I gloss over these differences in the remainder of this enquiry.

[52] For example, NEs might replace the economists' intertemporal discount rate with variables representing the activation patterns of neural areas whose operations influence agents' intertemporal preferences.

respect, no momentous achievement has been accomplished by NEs so far. Nonetheless, one can think of a few ways in which NEs might revolutionize economic theory without having to include many neural variables into the economists' models. For instance, NEs might prompt economists to employ altogether different constructs by illustrating that what was regarded as a unitary phenomenon (e.g. hyperbolic discounting) is more plausibly seen as distinct phenomena explained by dissimilar neuro-psychological mechanisms (Craver and Alexandrova, 2008, and Ross et al., 2008). That is to say, implementing revolutionary changes in the economic theory of choice does not necessarily require NEs to integrate many variables into the economists' models. Hence, *premise 3* turns out to be untenable and the argument from tractability fails.

To recapitulate, economists may put forward various considerations to justify their reluctance to include neural constructs into their models. Even so, the argument from tractability proves to be flawed in at least two central respects. In the first place, it remains disappointingly silent regarding in what modelling contexts and to what extent tractability can be given priority over other desiderata. In the second place, it purports - yet fails - to exclude that neurally informed contributions will prompt revolutionary modifications in the economic theory of choice. That is to say, economists often doubt the opportunity to include neural constructs into their models, but do not always vindicate their position by means of a convincing argumentative strategy. Now, these critical remarks do not preclude economists from cogently vindicating their opposition to import other disciplines' insights on the basis of tractability considerations. In the next section, I present one such line of argument which casts doubt on the NEs' case for incorporating neural constructs into economic models of decision making.

4.C A REFINED ARGUMENT FROM TRACTABILITY

In this section, I articulate and defend a *refined argument from tractability* which attempts to demonstrate that economists are provisionally justified in retaining a methodologically distinctive approach to the modelling of choice behaviour. My reasoning can be formalized in different ways depending on which attributes one considers besides tractability. In what follows, I shall prevalently concentrate on tractability and descriptive accuracy, with this latter indicating a model's goodness of fit with the structure or the features of its target system[53]. My focus is motivated both by the contrasting demands that descriptive accuracy and tractability place on economic modellers and by the importance that NEs and other economists attach to these two virtues[54].

Before presenting the refined argument from tractability, let me put forward two preliminary caveats. Firstly, my reasoning by no means denies that importing neural insights may enable economists to build more predictive and descriptively accurate models. On the contrary, such an argument is built on the implicit assumption that NEs can improve the predictive and descriptive performance of economic models of choice. At the same time, it casts doubt on whether these improvements make it convenient for economists to integrate many neural variables into their account of decision making.

[53] As noted in section 2.A, various questions may be raised regarding how descriptive accuracy is most appropriately defined. My approximate characterization is sufficiently precise to enable us to examine the trade-offs I mention in this chapter.

[54] A reasoning along similar lines could be constructed with regard to other modelling trade-offs. For instance, it is an open question whether NE will enable economists to elaborate a framework which is *at once* more tractable and general in scope than rational choice theory. As Vromen (2010b, p.180) puts it, even if "all the neural circuits linking environmental variables to choice behavior can be captured in one general algorithm", such an algorithm would be far more complicated than the utility functions postulated by standard economic models.

Secondly, even though tractability and descriptive accuracy typically pull in opposite directions, economists may succeed in constructing models which provide both tractable and descriptively accurate representations of their target systems. To see this, it is instructive to distinguish between various *levels* - e.g. psychological, neural, biological, and micro-physical - at which the descriptive accuracy of a model can be evaluated. Now, while the neural substrates of decision making are exceedingly complex to be accurately described in tractable terms (neural descriptive accuracy), the psychological underpinnings of choice behaviour can be frequently characterized in a fairly accurate and tractable manner (psychological descriptive accuracy). In this perspective, we can appreciate why economists often exhibit a *selective scepticism* with regard to the incorporation of other disciplines' insights. For one may coherently argue that various psychological findings are "tractable and parsimonious enough that we should begin the process of integrating them into economics" (Rabin, 1998, p.13), and yet resist the inclusion of neural insights into economic models.

Having said that, the refined argument from tractability can be articulated as follows:

P.1 If economists are to construct descriptively accurate and neurally informed economic models of choice, then they have to incorporate several neural variables (tractability cost).

P.2 Economists should incorporate neural variables only insofar as the modelling costs associated with such enrichment are compensated by sufficient modelling benefits.

P.3 NEs have not shown that the modelling benefits derivable from incorporating several neural variables compensate for the modelling costs associated with such enrichment.

C (1) NEs have failed to show that economists should incorporate several neural variables into their models of choice.

P.4 Showing that economists should incorporate several neural variables requires NEs to demonstrate that human choice behaviour is more conveniently modelled at the neural - rather than some other - level.

P.5 NEs have failed to demonstrate that human choice behaviour is more conveniently modelled at the neural - rather than some other - level.

C (2) NEs have failed to show that economists should incorporate several neural variables into their models of choice.

C Economists are provisionally justified in resisting NEs' calls to construct descriptively accurate and neurally informed economic models of choice.

Let us scrutinize the various steps which comprise the refined argument from tractability. *Premise 1* concerns the tractability costs that a neural enrichment of economic models is likely to impose on economists. The idea is that developing descriptively accurate NE models of decision making would require one to build rather intractable representations. To be sure, a proponent of NE may object that some processes and features of the human neural architecture can be modelled in tractable terms (see e.g. Schultz, 1998, and Schultz, Dayan, and Montague, 1997, on the dopaminergic underpinnings of reward evaluation). Yet, accurately representing the neural substrates of choice behaviour would typically impose significant tractability

costs, as numerous cerebral systems are highly interconnected at the anatomical and physiological levels (see e.g. Anderson, 2006 and 2007, and Glimcher, Dorris and Bayer, 2005, p.251).

Two issues are worth distinguishing in this respect. On the one hand, a number of neural areas activate in a wide range of decision contexts (Cabeza and Nyberg, 2000, and Price and Friston, 2005, p.262-5). On the other hand, even the execution of simple experimental tasks typically engages several areas, with additional variability resulting when it comes to solving complex decision problems (Ortmann, 2008, p.442). As Petersen and Fiez (1993, p.513) put it, "a set of distributed functional areas must be orchestrated in the performance of even simple cognitive tasks […] Any task or 'function' utilizes a complex and distributed set of brain areas". To be sure, NEs may occasionally succeed in modelling the neural underpinnings of people's decisions without using many variables. Still, accurately representing the neural substrates of choice behavior usually requires them to employ several variables. In the words of Bernheim (2009, p.7), "precise algorithmic models of decision making of the sort to which many neuroeconomists aspire would presumably map highly detailed descriptions of environmental and neurobiological conditions into choices".

At this point, the proponent of NE may protest that both economists and NEs generally attach a greater importance to predictive power than to descriptive accuracy. Moreover, she may argue that even though tractability considerably constrains the accuracy with which one can characterize the neural substrates of choice behaviour, NEs could build more predictive economic models without incurring substantial tractability costs (see

e.g. Knutson et al., 2007, and Kuhnen and Knutson, 2005, on the possibility of increasing the predictive power of some models of choice by investigating the activation patterns exhibited by a few neural areas). On this basis, the advocate of NE may contend that *premise 1* merely highlights a local conflict between two conveniently selected desiderata which does not hinder the import of neural insights into economic models.

To such a criticism, my rejoinder is three-fold. To begin with, the fact that many economists value predictive power falls short of excluding other desiderata from playing an important role in economic modelling. Secondly, both economists and NEs ascribe a great significance to descriptive accuracy, with some authors going as far as to claim that "the ultimate test of a theory" is "the accuracy with which it identifies the actual causes of behaviour" (Camerer and Loewenstein, 2004, p.4). Finally, it is highly doubtful that invoking predictive considerations *per se* enables NEs to substantiate their case for the neural enrichment of economic theory. For even if NEs may succeed in increasing the predictive power of some economic models with marginal tractability losses, it is an open question whether integrating neural insights will often be so advantageous to economists. Let me expand on this issue.

While the first premise of the refined argument from tractability relates to the tractability costs involved in a neural enrichment of economic theory, *premises 2* and *3* concern the relationship between the modelling costs and the benefits associated with this enrichment. More specifically, *premise 2* asserts that economists should integrate neural variables only insofar as doing so yields modelling benefits which compensate

for the modelling costs imposed by such integration. In this respect, various NEs allege that - by taking neural findings into account - economists could elaborate models of choice which fare better according to their own modelling criteria (Camerer, 2008a, p.59; see Camerer and Loewenstein, 2004, p.3, for an analogous remark concerning psychological findings). Regrettably, as stated by *premise 3*, NEs still have to demonstrate that the benefits derivable by integrating neural variables into economic models typically exceed the associated modelling costs. This being the situation, their pleas in favour of a neural enrichment of the economic theory of choice remain inadequately supported (*conclusion 1*). As Bernheim (2009, p.7) provocatively puts it: "What does a standard economist lose by subsuming all of the idiosyncratic, micro-micro factors that influence decisions, many of which change from moment to moment, within a statistical disturbance term?"[55].

A proponent of NE may rebut that the above considerations place an excessive burden on the NEs' shoulders. In particular, she might contend that substantiating the case for the neural enrichment of economic theory just requires NEs to show that neural data are of some value to economists. However, *premises 4* and *5* of the refined argument from tractability cast serious doubt on such a rebuttal. Let us inspect each of these assertions in turn.

[55] A proponent of NE might point out that there are already some tractable economic models that incorporate neural insights. The existence of these models, however, does not undermine *premise 3*. For such a premise does not deny the existence of neurally enriched and tractable economic models. On the contrary, it is best interpreted as stating a typicality claim which doubts that economists will usually find it convenient to include many neural insights into their models.

According to *premise 4*, it is up to NEs to specify why exactly economists should integrate neural - as opposed to psychological, biological, micro-physical etc. - variables into their models. This claim relates to the fact that human choice behaviour can be modelled at a number of different levels (e.g. psychological, neural, biological, and micro-physical). That is to say, there are various processes and entities in terms of which agents' decisions may be represented, and one can typically employ distinct sets of constructs to account for decision making. For example, while rational choice theorists represent observed choices as the maximization of a utility function under some suitably defined constraints, cognitive psychologists account for those data in terms of specific heuristics and cognitive mechanisms, computational neuroscientists attempt to identify what neural algorithms underlie the observed decisions, and so on. In this perspective, NEs' proposals can be seen as one among several competing candidates for modelling human choice behaviour[56].

One may protest that most NEs advocate the integration of neural variables - not so much as an *alternative*, but rather - in *addition* to other disciplines' insights. In the words of Park and Zak (2007, p.54), "augmented economic models will also likely include results from sociology, anthropology, psychology, and other fields" (see also Camerer 2008b, on the opportunity to triangulate behavioural, psychological and neural data). Yet, the point remains that simultaneously including constructs from various

[56] Other ways to conceptualize model construction in NE have been advocated. In particular, various authors (e.g. Rustichini, 2009) rely on a tripartition originally proposed by Marr (1982) with regard to information-processing systems. The idea is to study human choice behaviour at three different levels. The computational level specifies the problem faced by the examined agents in terms of some input-output mapping. The algorithmic level explicates how the agents represent the input-output mapping and what algorithms transform the inputs into the outputs. The hardware implementation level concerns how the computational and algorithmic processes associated with people's decisions are instantiated at the physical level.

disciplines into a single economic model would often be prohibitively impractical. In other words, given that a limited number of variables can be feasibly imported into one economic model, a criterion is needed for assessing the relative relevance of distinct disciplines' insights for the economic theory of choice. Regrettably, as remarked by *premise 5*, NEs have hitherto failed to convincingly support the claim that choice behaviour is most conveniently modelled at the neural level. Moreover, they rarely indicate what subset of neural data (e.g. neuro-anatomical, neuro-physiological, neuro-biological) economists should incorporate into their models. This being the situation, it is hard to see why economists should integrate neural - rather than higher-level or lower-level - insights into their account of decision making[57].

At this point, one may wonder how the assertion that human choice behaviour is most "conveniently" modelled at the neural level is to be interpreted. In addressing this concern, we should avoid imposing overly rigid restrictions as to how the modelling benefits and the costs associated with a given enrichment of economic theory are to be balanced against one another. For *in primis*, a modeller's views on this matter may significantly vary depending on her modelling goals and on what choice problem she is examining (Landreth and Bickle, 2008, p.420). Secondly, modelling costs and benefits cannot always be compared precisely in quantitative - rather than qualitative - terms. And thirdly, intradisciplinary conventions frequently diverge regarding what modelling attributes are to be valued in a given context, with distinct conventions resolving the trade-offs between different desiderata in dissimilar ways (Mäki, 2010, p.108).

[57] I speak of "higher-level" and "lower-level" insights following an entrenched terminological convention in the philosophy of science literature. My doing so does not commit me to endorse the hierarchical view of the structure of science often held by those who employ these expressions.

To be clear, these concerns do not preclude us from meaningfully comparing the merits of distinct enrichments of economic theory. Still, they persuasively suggest that those evaluations are most sensibly implemented in context-relative terms. To appreciate this, suppose that we wanted to evaluate alternative enrichments of economic theory by means of the following criterion *C*:

> A *target system T* is more conveniently modelled at level *X* rather than *Y* if and only if the difference between the *modelling benefits* (MB) derivable from characterizing *T* in terms of *X* and the associated *modelling costs* (MC) is higher than the analogous difference related to level *Y*, i.e. iff $[MB (T_X) - MC (T_X)] > [MB (T_Y) - MC (T_Y)]$.

Criterion *C* is defined in relation to two distinct levels of description, but can be modified to deal with cases where insights from more levels of description are available. In this way, it can be used to handle situations where distinct modellers hold dissimilar views concerning *which* and *how many* levels of description it is most convenient to integrate into a model. To render this point more vivid, let us assume that insights from three levels of description *X*, *Y* and *Z* were available, with *X*, *Y* and *Z* respectively standing for the behavioural (observed choices), psychological and neural level of description. Ascertaining how many levels of description it is most convenient to include into a model can be seen as the problem of identifying which of the following differences is greatest:

$[MB (T_X) - MC (T_X)]$; $[MB (T_Y) - MC (T_Y)]$; $[MB (T_Z) - MC (T_Z)]$; $[MB (T_{X,Y}) - MC (T_{X,Y})]$;
$[MB (T_{X,Z}) - MC (T_{X,Z})]$; $[MB (T_{Y,Z}) - MC (T_{Y,Z})]$; $[MB (T_{X,Y,Z}) - MC (T_{X,Y,Z})]$.

Now, let us distinguish between those modelling approaches which involve using neural data (namely T_Z, $T_{X,Z}$, $T_{Y,Z}$ and $T_{X,Y,Z}$) and those which do not (namely T_X, T_Y and $T_{X,Y}$). Let us call these two sets *neurally informed* and *neurally free* approaches respectively. Illustrating that adopting some neurally free approach is more convenient than integrating neural insights ($T_{X,Z}$) would fall short of implying that neural insights are of little value to economists. For economists may still find it more convenient to combine neural with behavioural and psychological insights ($T_{X,Y,Z}$) than to rely on any of the neurally free approaches. Conversely, showing that economists should import neural insights would require NEs to demonstrate that employing one of the neurally informed approaches is more convenient than adopting any of the currently available neurally free approaches.

In such a context, it is worth noting the significance that *expected* modelling costs and benefits may be given when assessing distinct enrichments of economic theory. By way of illustration, consider the following comparison between a neural and a psychological enrichment. One may argue that behavioural economists have provided other economists with far greater modelling benefits than NEs, whose success stories are often confined to highly controlled experimental settings. Moreover, she may remark that accurately representing the anatomical or functional interplays between distinct neural areas would impose on economists substantial modelling costs, which exceed those required to construct psychologically realistic models. Now, let us suppose that these observations were correct. One might take them to imply that economists will find it more convenient to implement a psychological - rather than neural - enrichment of

economic theory. However, NEs may object that the modelling benefits promised by their contributions - albeit still hypothetical - are so significant that economists should opt for a neural enrichment of their models. The idea would be to take into account not just current, but also expected modelling benefits and costs, i.e. anticipated benefits and costs discounted for their subjectively estimated probability of actualization.

By doing so, we might be able to better assess what enrichments of economic theory are likely to be more convenient for economists. Yet, we would also introduce some highly conjectural elements into our analysis. To see this, let us briefly consider the explanatory benefits derivable from incorporating neural insights into economic models. On the one hand, NEs already acquired valuable information concerning the neuro-physiological underpinnings of human choice behaviour (see e.g. Barraza and Zak, 2009, Kosfeld et al., 2005, and Zak et al., 2004, for some NE studies of the effects that specific hormones have on agents' decisions). On the other hand, the question remains as to how exactly these neuroscientific advances bear on economic theorising. In particular, it is doubtful that economists will usually find neurally enriched accounts of decision making more explanatorily insightful than psychologically informed ones (see section 4.A).

Regrettably, the impression remains that several NEs exhibit an exceedingly optimistic attitude regarding the achievements that their studies can be plausibly expected to accomplish. To render this point more vivid, let us consider one specific example taken from recent NE research, namely the axiomatic model of learning by Caplin and Dean (2008; see also Caplin et al., 2010). Caplin and Dean provide a set of axioms defining

beliefs and rewards in terms of dopaminergic activity so as to specify "in a simple, parsimonious, and nonparametric way the properties that the dopamine system must have in order to be characterized as encoding a reward prediction error" (2008, p.670). According to Caplin and Dean, their model has a number of "economic applications", which purportedly include "insight into belief formation […] learning theory [and] addiction" (2008, p.665). Now, let us suppose that such a model improves our understanding of the algorithmic underpinnings of learning and enables us to test the reward prediction-error hypothesis of dopamine function (see Caplin et al., 2010). These insights are of great interest to cognitive and computational neuroscientists, yet do not directly bear on economists' traditional concerns. As Caplin et al. (2010, p.953) aptly acknowledge, their model does not "immediately advance our understanding of choice".

To recapitulate, my refined argument from tractability challenges NEs to demonstrate that the modelling benefits offered by a neural enrichment exceed the corresponding modelling costs. Moreover, it provides two main reasons to doubt that choice behaviour is more conveniently modelled at the neural - as opposed to some other - level. The first reason relates to specific anatomical and functional features of the human neural architecture. The idea is that, due to the complexity of the interconnections between distinct neural areas, accurately characterizing the neural substrates of decision making will require NEs to incur inconveniently high modelling costs. The second reason concerns the existence of multiple levels of description of choice behaviour, and calls NEs to specify on what grounds economists should integrate neural - as opposed to other disciplines' - constructs into their models (see also Fumagalli, 2011). Regrettably,

the proponents of NE frequently gloss over these issues and the implications they have for the construction of models spanning different behavioural disciplines

In response to the above remarks, NEs may attempt to substantiate their case for incorporating neural insights into economic models by appealing to *modelling pluralism* considerations. The reasoning can be summarized as follows. Due to the existence of trade-offs, economists are typically unable to construct models that simultaneously possess all the attributes they value. While some of these trade-offs are likely to abate with scientific progress, others will persist in spite of it[58]. Hence, if economists are to achieve their diverse modelling goals, they have to elaborate a variety of models differing in descriptive accuracy, predictive power, included causal factors, etc.

This rationale for building multiple models has been persuasively advocated in the literature on modelling in biology (see e.g. Levins, 1966 and 1968, Roughgarden, 1979, Wimsatt, 1987, and May, 2001). The idea is that modellers can employ a mixed representational strategy, using distinct kinds of models to achieve different pragmatic and epistemic goals (Weisberg, 2007a; see Aydinonat, 2010, p.164, for a similar claim in the economic literature). In this perspective, neurally informed models can be seen as - not so much substitutes for, but rather - complements to traditional economic models of choice.

[58] Among the constraints which will not abate with scientific advances, Matthewson and Weisberg (2009, p.188) include strict trade-offs, increase trade-offs and Levins trade-offs. Two attributes display a strict trade-off "when an increase in the magnitude of one desideratum necessarily results in a decrease in the magnitude of the second, and vice versa". An increase trade-off occurs when the magnitudes of two desiderata cannot be simultaneously increased. Finally, two attributes exhibit a Levins trade-off "when the magnitude of both of these attributes cannot be simultaneously maximized".

Now, by combining neural and other disciplines' insights, economists may succeed in constructing an array of models which - taken collectively - enable them to better attain their predictive and explanatory goals. Moreover, a pluralistic approach appears to be especially worth pursuing when it comes to modelling phenomena - such as human choice behaviour - that are investigated by a range of different disciplines. Nonetheless, the above pluralistic remarks do not license the claim that economists should integrate several neural constructs into their models. For even if such integration enabled them to achieve specific modelling goals in particular contexts, it would remain up to NEs to demonstrate that elaborating a neurally informed account of choice behaviour is, in general, more convenient for economists than relying on traditional modelling frameworks.

To conclude, the emergence of NE raised fundamental methodological issues concerning the relevance of neural data for economic modelling and theorising. At present, most economists remain fairly sceptical about the alleged significance of NE research for the economic theory of choice. As I argued above, including neural insights can improve economic models of choice with regard to several desiderata, individually taken (*local improvements*). Yet, substantiating the case for a neural enrichment of economic models requires NEs to show that such enrichment will foster *global improvements* in the economic account of decision making. Regrettably, the proponents of NE have hitherto failed to do so. Hence, economists are provisionally justified in retaining a methodologically distinctive approach to the modelling of decision making.

CHAPTER FIVE

ARGUMENT FROM DISCIPLINARY HETEROGENEITIES

The proponents of NE often manifest the ambition to combine findings from economics, psychology and neuroscience into a "single, general theory of human behaviour" (Glimcher and Rustichini, 2004, p.447; see also Glimcher, 2010, p.393, Glimcher, Dorris and Bayer, 2005, p.214, and Rustichini, 2005, p.203). Faced with these claims, some economists (e.g. Gul and Pesendorfer, 2008) point to the fact that these disciplines focus on dissimilar evidential bases, employ different constructs and are concerned with distinct explananda. The idea is that the profound differences between NE's parent disciplines in terms of variables of interest, methodological presuppositions and explanatory goals constrain NEs' attempts to foster a genuine unification spanning economics, psychology and neuroscience.

In this chapter, I examine how the aforementioned differences bear on the relevance of NE research for the economic theory of choice. The contents are organized as follows. In section 5.A, I reconstruct and critically assess the *argument from irrelevance* that some economists (e.g. Gul and Pesendorfer, 2008) employ to demonstrate that neuroscientific evidence is evidentially and explanatorily irrelevant to economic theory. In section 5.B, I articulate and defend an *argument from interdisciplinary heterogeneity* which casts doubt on NEs' attempts to develop a unified interdisciplinary framework by emphasizing the differences between the accounts of choice behaviour provided by NE's parent disciplines. In section 5.C, I put forward an *argument from intradisciplinary heterogeneity* which identifies some issues (e.g. how NE is supposed

to inform economic theory) on which NEs themselves hold contrasting positions and shows how these divergences hamper the consolidation of the NE enterprise.

5.A THE ECONOMISTS' ARGUMENT FROM IRRELEVANCE

Faced with NEs' calls to incorporate neuro-physiological insights into economic models of choice, various economists concede that NE evidence can inspire the construction of more predictive and explanatory economic models. At the same time, they deny that neuro-physiological findings have any significant bearing on the economic theory of choice. The idea is that those findings fall outside the domain of economic theory and are mostly orthogonal to the professional interests of economists. As Gul and Pesendorfer put it, "neuroscience evidence cannot refute economic models because the latter make no assumptions and draw no conclusions about the physiology of the brain" (2008, p.4)[59]. The economists' *argument from irrelevance* can be articulated as follows:

P.1 Neuroscientists and other economists employ different theoretical constructs.

P.2 Neuroscientists and other economists target distinct sets of variables.

P.3 Neuroscientists and other economists pursue dissimilar explanatory goals.

C Neuroscientific findings are evidentially and explanatorily irrelevant to economic theory.

Prima facie, this argument seems to provide economists with an appealing defensive strategy, which isolates the economic theory of choice from potentially disconfirming neuro-physiological evidence. As I argue below, however, the above reasoning is vulnerable to two major objections. Firstly, it fails to show that neuroscientists' and

[59] The isolationist attitude of Gul and Pesendorfer is not representative of most economists' position. I focus on it as it nicely exemplifies one extreme position in the literature and provides a useful point of reference for our discussion.

other economists' accounts of choice behaviour are incommensurable, i.e. have non-overlapping evidential bases, lack common concepts and pursue altogether different explanatory aims[60]. And secondly, it seemingly overlooks that even if those accounts were incommensurably dissimilar, it would not follow that NEs are prevented from fostering significant modifications in economic theory. These two objections resemble some of the criticisms that were formulated against the so-called *Non Overlapping Magisteria* model (NOMA) of the relationship between science and religion (see Gould, 1997 and 1999). Let me briefly expand on this parallel before assessing the economists' argument from irrelevance.

According to the NOMA model, science and religion profoundly differ in their methodological presuppositions and investigated objects. These differences, in turn, are said to preclude the possibility of systematic conflict and fruitful interchanges between scientific discoveries and religious beliefs. Now, let us suppose that we could provide sufficiently precise characterizations of science and religion. It would seem implausible to deny that scientists *qua* scientists and religious believers *qua* religious believers respectively address different kinds of questions and rely on dissimilar evidential standards. For instance, scientists *qua* scientists attempt to account for the features and the workings of physical systems by means of highly controlled and replicable experiments. For their part, religious believers *qua* religious believers are concerned with questions of meaning and value that often transcend the empirical realm (see e.g.

[60] In this section, I focus on the evidential bases, the theoretical constructs and the explanatory aims associated with the economic and neuroscientific accounts of choice behaviour, without considering other respects in which different theoretical frameworks can be deemed to be incommensurable (see e.g. Sankey, 1998, on so-called taxonomic incommensurability, which obtains when distinct theories provide inconsistent categorizations of their objects of interest, and Feyerabend, 1975, p.271, and 1981, xi, who regards two theories as incommensurable only if they have inconsistent ontological implications).

Ratzsch, 2009). These differences, however, by no means imply that scientific and religious accounts of reality are incommensurable. In particular, they do not preclude fruitful interchanges of ideas at the frontier between science and religion[61].

Now, let us focus on the economists' argument from irrelevance. It would be hard to deny that neuroscientists and mainstream economists respectively use different vocabularies, target dissimilar variables and pursue distinct explanatory goals. The thought is that economists, even when addressing questions related to those studied in psychology and neuroscience, "have different objectives and target different empirical evidence" (Gul and Pesendorfer, 2008, p.4). In this respect, even leading NEs concede that "what is striking about explanations of choice behavior by economists, psychologists, and neurobiologists is the different levels at which they operate" (Glimcher and Rustichini, 2004, p.448). These remarks, however, do not license the claim that the accounts of choice behaviour provided by economists, psychologists and neuroscientists are *incommensurable*. Let me expand on this point.

Assessing the similarity of different theoretical frameworks is more appropriately seen as a matter of nuanced evaluation than an all-or-nothing judgement (see e.g. Duhem, 1906 [1954]). By way of illustration, let us consider the economic, psychological and neuroscientific accounts of choice behaviour. One may argue that differences in the employed constructs and the pursued explanatory goals are greater between economics and neuroscience than between economics and psychology (e.g. think of the role that psychological constructs such as beliefs and desires play in some economists' accounts

[61] See e.g. Harrison, 2007, on the role that religious beliefs played in fostering the development of several scientific disciplines and the constraints that scientific discoveries impose on specific religious beliefs.

of behaviour). Indeed, some authors (e.g. Kuorikoski and Ylikoski, 2010) go as far as to assert that neuroscientific findings are explanatorily and evidentially irrelevant to economic theory unless they are interpreted in light of psychological background knowledge[62]. Even so, the point remains that Gul and Pesendorfer overstate the heterogeneity of NE's parent disciplines and presuppose - rather than show - that the accounts of behaviour provided by those disciplines are totally disconnected (Vromen, 2010b). In particular, their attempt to block NEs' calls to include neuro-psychological insights into economic models rests on an unsustainably narrow conception of standard economic theory (Moscati, 2008). This, in turn, calls into question the *a priori* opposition of Gul and Pesendorfer to a neuro-psychological enrichment of economic theory. As Mäki (2010, p.115) puts it, "economics has no immutable essence such that […] only data pertaining to observable choice behaviour would be relevant".

Gul and Pesendorfer are not the first authors who overemphasize the conceptual divide between different sciences or theoretical frameworks. By way of illustration, let us consider the debates which took place among philosophers of science regarding how frequently cases of conceptual incommensurability between scientific theories occur. According to Kuhn (1962 and 1982) and Feyerabend (1962), intertheoretic transitions involve radical modifications at both the intensional and the extensional level. The idea is that central theoretical terms are often ascribed altogether different meanings and/or reference by the proponents of rival theories (e.g. compare the concept of mass in Newtonian and relativistic mechanics). Yet, as noted by various authors (e.g. Devitt, 1979, Field, 1973, and Fine, 1967), significant referential continuity can be found across

[62] Similar claims have been put forward with regard to the relationship between folk psychology and neuroscience. For instance, McCauley (1996, p.445) alleges that any influence neuroscience has on folk psychology will be mediated by progress in social and cognitive psychology.

theory change, and scientists can usually compare theories in terms of a shared vocabulary.

Having said that, let us assess the *implications* that the alleged incommensurability of economics and neuroscience can be taken to have for the relevance of NE for economic theory. Suppose - for the sake of argument - that economists and neuroscientists respectively provided incommensurable accounts of choice behaviour, i.e. targeted altogether different variables, employed altogether different theoretical constructs and pursued altogether different explanatory goals. Even this, by itself, would not exclude that neuroscientists may offer informative insights and pose critical challenges to economists. For some NEs may still rely on neuroscientific concepts and findings to foster modifications in the economic account of decision making. In this respect, various authors (e.g. McCabe, 2008, p.350) concede that NEs themselves and other economists are respectively concerned with dissimilar constructs and explananda, and yet insist that NE findings inform fruitfully economic modelling and theorising.

A proponent of the argument from irrelevance might insist that no significant interchange can take place between genuinely incommensurable theoretical frameworks. This rebuttal, however, does not appear to withstand scrutiny. For, as illustrated by several episodes in the history of science (see e.g. Kuhn, 1962, and Feyerabend, 1962), the incommensurability of two theoretical frameworks does not *per se* exclude that significant - or even revolutionary - interplays may occur between them. To put it differently, even if economists and neuroscientists offered incommensurable accounts of decision making, neurally informed contributions could still foster

substantial modifications in economic theory. For instance, NEs may extend the set of questions economists traditionally address or foster an eliminative reduction of the folk psychology constructs underlying some economists' models of choice (I shall expand on these issues in section 7.B).

To recapitulate, the argument from irrelevance emphasizes various respects in which neuroscientists' and economists' accounts of choice behaviour differ. Even so, it fails to show that these accounts are incommensurably dissimilar. Moreover, even if those accounts were incommensurable, it would not follow that NEs cannot foster revolutionary modifications in the economic theory of choice. Having said that, the profound differences in the accounts of choice behaviour that are respectively provided in economics, psychology and neuroscience *do* constrain the relevance of NE findings for these disciplines. In the next section, I articulate and defend an argument which draws on those differences to question the prospects of NEs' attempts to construct a unified interdisciplinary framework for modelling decision making.

5.B AN ARGUMENT FROM INTERDISCIPLINARY HETEROGENEITY

My *argument from interdisciplinary heterogeneity* questions NEs' attempts to develop a unified interdisciplinary framework by emphasizing the differences between the accounts of choice behaviour provided by NE's parent disciplines. Before presenting the argument, let me provide some terminological clarification regarding the notion of unification to which I refer below. The term "unification" can be employed to denote various kinds of accomplishments, ranging from the development of a common vocabulary or taxonomy to the construction of a shared mathematical framework to model the phenomena of interest. In this section, I employ the expression 'unified theoretical framework' to refer to a collection of studies which (i) share a sufficiently precise definition of NE, (ii) are inspired by reasonably similar explanatory aims, and (iii) reflect consistent views concerning the relationship between economics, psychology and neuroscience. My reasoning can be characterized as follows:

P.1 Economists, psychologists and neuroscientists employ different theoretical constructs.

P.2 Economists, psychologists and neuroscientists target distinct sets of variables.

P.3 Economists, psychologists and neuroscientists pursue dissimilar explanatory goals.

C (1) Developing a unified NE account of decision making requires major changes in the models provided by NE's parent disciplines.

P.4 NEs have not shown that developing a unified NE account of decision making yields high modelling benefits to the practitioners of NE's parent disciplines.

C NEs have not shown that the practitioners of NE's parent disciplines should develop a unified NE account of decision making.

Let us examine the various steps of this argument. *Premises 1* and *2* respectively state that economists, psychologists and neuroscientists employ dissimilar constructs and target different variables when building their models. The thought is that economists are concerned with higher-level phenomena than the ones investigated in neuro-psychology (Vromen, 2007, p.162). Moreover, while economists often build their models upon axiomatic foundations and *a priori* assumptions about rational choice behaviour, many psychologists and neuroscientists develop their models on the sole basis of empirical findings (Vercoe and Zak, 2010). To be sure, some NEs (see Caplin and Dean, 2008, and Caplin et al., 2010) have recently developed axiomatic models that target both economic and neural data. Still, NEs and other economists usually make a different use of the axiomatic method. For instance, as noted by Rustichini (2009, p.50), "the functional representation of choice in decision theory is not considered a testable hypothesis, whereas the algorithmic specification [targeted by many NE studies] is. This difference gives a new role to the axiomatic method".

A proponent of NE might deny that the aforementioned differences constrain the relevance of NE research for economic theory on the alleged ground that NEs and other economists frequently rely on the same set of theoretical constructs. After all - the thought would be - both NEs and other economists use concepts such as choices, utility, preferences, and so on. Now, it would be implausible to deny that NEs and other economists employ some common constructs in building their models. Yet, the point

remains that NEs and other economists often use these constructs in different ways and ascribe dissimilar meanings to them. To render this point more vivid, let us consider the concept of rationality. NEs and other economists frequently rely on dissimilar conceptions of what it means for a given decision to be rational. The following contrast appears to be particularly profound. On the one hand, traditional decision theorists relate rationality to the internal consistency of observed choices and remain agnostic regarding what neuro-psychological processes underlie people's decisions (see e.g. Kacelnik, 2006). On the other hand, many NEs attempt to identify the neuro-psychological underpinnings of choices and regard a decision as rational to the extent that it maximizes some specific neuro-psychological measure of well-being.

Equally profound discrepancies can be identified with regard to the concept of utility. More specifically, traditional decision theorists regard utility as a formal representation of preferences to be inferred from observed choices. For their part, many NEs (e.g. Camerer, Loewenstein and Prelec, 2004 and 2005, Glimcher and Rustichini, 2004, and Park and Zak, 2007) relate utility to agents' hedonic experiences or the activation patterns of specific neural areas (see Caplin and Dean, 2008, p.670, for analogous remarks concerning the notion of reward). As I argue in *chapter six*, these conceptual differences constrain the extent to which neuro-psychological findings can inform standard economic theory. Regrettably, NEs often gloss over those differences as if they were of negligible significance for the prospects of NE.

Premise 3 focuses on the explanatory goals that are respectively pursued by economists, psychologists and neuroscientists. This contrast can be explicated as follows. On the one

hand, economists usually target agents' observed choices and aim to identify the solution to specific decision problems without taking a position as to which objectives agents should pursue. On the other hand, NEs investigate the neuro-psychological substrates of observed choices and often evaluate people's decisions according to a normatively oriented perspective. In particular, they aim to ascertain - not just what is the best way to achieve agents' objectives, but also - which objectives agents should pursue in specific situations (Read, 2007, p.58).

I am not concerned here with establishing whether NEs rely on an inappropriate analogy between economists and paternalistic advisors (see e.g. Gul and Pesendorfer, 2008, and Loewenstein and Haisley, 2008, for a debate). For the purpose of this enquiry, it suffices to note that ascertaining what objectives an agent should pursue requires one to address issues which transcend both the scope of traditional decision theory and the evidential reach of neuro-psychological investigations (e.g. who is entitled to define what constitutes agents' well-being? Which welfare criteria should be used to evaluate people's choices? To what extent may paternalistic interventions legitimately interfere with people's decisions?).

On the basis of the first three premises, *conclusion 1* asserts that the NE synthesis spanning economics, psychology and neuroscience is likely to involve major interdisciplinary rearrangements. The thought is that the profound differences between the economic, psychological and neuroscientific accounts of decision making constrain NEs' attempt to integrate evidence, constructs and methods from these disciplines into a unified theoretical framework. Let me expand on this issue.

Over the last few decades, various models combining economic and psychological insights have been developed. By way of illustration, let us consider regret theory (Bell, 1982, Fishburn, 1982, and Loomes and Sugden, 1982 and 1987). In this theory, the utility associated with an option is an increasing function of the payoff yielded by the option and a decreasing function of the payoff given up as a result of one's decision (Bell, 1982). This modelling approach enables one to account for the fact that "the psychological experience of [having an option] can be influenced by comparisons between [that option and the options] one might have had, had one chosen differently" (Sugden, 1991, p.762)[63]. In particular, it offers a psychologically plausible way to account for various violations of standard utility theory (e.g. some instances of cyclical choice). To see this, suppose that an agent exhibits the preference profile $x>y$, $y>z$, $z>x$. If the payoff she derives from choosing option x turns out to be considerably inferior to the payoff she could have obtained by choosing y, the resulting regret may reduce the utility she derives from x so much that she comes to prefer z to x. In this perspective, seemingly intransitive preferences can be reinterpreted as instances of the preference relation $x_1>y$, $y>z$, $z>x_2$.

As this example illustrates, psychological and economic insights can be integrated fruitfully so as to better account for observed choice behaviour. Still, these accomplishments are usually confined to specific models or choice contexts. Moreover, it remains an open question whether this integrative approach can be extended successfully to include neuroscientific - and not just psychological - evidence and

[63] See also the so-called disappointment theory (Bell, 1985, and Loomes and Sugden, 1986), where the utility one derives from an option depends on how the payoff associated with such an option compares with the payoff she previously expected to obtain from it.

constructs. To be sure, various authors have recently tried to reduce the "conceptual gap" between the economic, psychological and neuroscientific accounts of choice behaviour (see e.g. Caplin and Dean, 2008, for an axiomatic model of learning which includes both neural and choice data). However, these integrative efforts have quite a limited scope and do not substantiate NEs' speculations about the development of a single, general theory of choice behaviour spanning economics, psychology and neuroscience. As Glimcher (2010, p.14) aptly notes, "economists, psychologists, and biologists can all offer local explanations [of choice behaviour], but what is striking is the unrelatedness of their explanations".

According to *premise 4*, developing a unified interdisciplinary framework for modelling choice behaviour is unlikely to bring major modelling benefits to the practitioners of NE's parent disciplines. The reasoning can be explicated as follows. Let us suppose - for the sake of argument - that NEs shared a precise view of how insights from economics, psychology and neuroscience should be integrated into a single, general theory of choice behaviour. Even so, it remains an open question whether developing such a theory provides economists, psychologists and neuroscientists with informative insights. In particular, the proponents of NE have hitherto failed to specify why exactly the practitioners of disciplines as diverse as economics, psychology and neuroscience should invest time and resources in this grandiose enterprise. On this basis, the *conclusion* of the argument from interdisciplinary heterogeneity challenges NEs to substantiate their calls to develop a single, general account of decision making spanning different behavioural disciplines.

Below I examine and critique two lines of argument by means of which the proponents of NE have tried to vindicate their attempts to develop an interdisciplinary account of choice behaviour. The first argumentative strategy calls into question the claim that providing a unified account of choice behaviour requires major interdisciplinary rearrangements. The second one attempts to demonstrate that the explanatory benefits offered by such an account make it worth investing in NE research. Before proceeding, let me anticipate that there are various respects in which economists' modelling tools can be of help to NEs and other neuroscientists (e.g. think of the applications that optimization techniques have in computational neuroscience). In this section, however, I focus solely on how NE methods and findings can inform the economic theory of choice.

The first reasoning goes as follows. NEs occasionally speak of promoting a progressive *convergence* between - or even the *unification* of - economics, psychology and neuroscience. For instance, Fehr and Camerer (2007, p.419) allege that NE "hopes to unify mechanistic, mathematical and behavioral (choice-based) measures and constructs". Similary, Camerer, Loewenstein and Prelec (2004, p.573) conjecture that "a biological basis for behavior in neuroscience […] could provide some unification across the social sciences"[64]. Even so, economists, psychologists and neuroscientists could build a unified theory of choice behaviour, and yet retain methodologically distinct approaches to the modelling of decision making. The idea is that economists,

[64] Analogous views have been advocated in relation to other behavioural sciences. For instance, Camerer (1999, p.10575) contends that behavioural economics increases the psychological plausibility of economic models, "promising to reunify psychology and economics". Similarly, Gintis (2004, p.37; see also 2007) argues that behavioural sciences' "incompatible models and disparate research methodologies" will be unified thanks to "theoretical tools […] and data gathering techniques" that transcend disciplinary boundaries.

psychologists and neuroscientists develop a common NE theory of choice behaviour and still continue to investigate their respective objects of interest by means of traditional methods and approaches. In this way, NEs could allegedly improve NE's parent disciplines on their own terms without having to literally unify them (Camerer, 2008a, p.59). As Glimcher and Rustichini (2004, p.452) put it, NE aims to develop "a mechanistic, behavioral, and mathematical explanation of choice that transcends the explanations available to neuroscientists, psychologists, and economists working alone".

To substantiate this line of argument, NEs may point at some successful cases of coevolution occurred at the interface between psychology and neuroscience (see e.g. McCauley, 1996 and 2007). Moreover, they may argue that promising coevolutionary advances are underway between economics and neuroscience (see e.g. Caplin and Dean, 2008). Now, it would be implausible to deny that neuro-psychological and economic accounts of choice behaviour can inform each other in fruitful ways (see e.g. Caplin et al., 2010, on how knowledge of dopaminergic circuits restricts the set of learning processes that can be plausibly associated with people's choices, and Glimcher, 2010, p.234-6, on how observed choices constrain NEs' conjectures about what algorithms underlie people's decisions). Even so, there are some reasons to be cautious concerning the prospects of coevolutionary approaches to the modelling of decision making. To give one example, the mere fact that a mechanistic approach has been adopted fruitfully in some branches of neuroscience does not *per se* imply that economists will find it convenient to build mechanistically informed models. That is to say, even if coevolutionary strategies have fostered significant advances in specific areas of neuro-

psychological research, it remains an open question whether the same will happen with NE.

In such a context, it would be of little import to allege that since behavioural economists managed to include several psychological insights into economic theory, NEs will succeed in their coevolutionary crusades. For in light of the differences between NE and previous research at the interface between economics and psychology (see section 1.B), the results attained by behavioural economists do not *per se* license the claim that NE will foster major advances in economic modelling and theorising. More generally, it is still unclear how informative coevolutionary constraints will be across NE's parent disciplines.

Indeed, there are reasons to be quite prudent in this respect. By way of illustration, consider Marr's (1982) tripartition between the computational, algorithmic and hardware implementation levels. As remarked by Craver and Alexandrova, "many different algorithms can solve the same computational problem, and many different hardwares can implement the same algorithms" (2008, p.393; see also Fernandes and Kording, 2010, p.345). Hence, NE findings concerning the algorithmic underpinnings of people's decisions may fail to be informative to those economists and neuroscientists who aim to provide computational and hardware implementation accounts of choice behaviour.

The second ground on which NEs advocated the construction of a unified interdisciplinary framework for modelling choice behaviour relates to the benefits that

such a framework can be expected to yield to NE's parent disciplines. As we have seen in *chapter two*, several NEs provide alluring characterizations of the predictive and explanatory gains that a neural enrichment of economic theory purportedly offers to economists. Regrettably, as we noted in *chapter four*, the benefits at which many NEs hint are still hypothetical or confined to specific choice settings. To be sure, one welcomes NEs' attempts to integrate the accounts of choice behaviour provided by NE's parent disciplines. Yet, the point remains that economists, psychologists and neuroscientists have *already* achieved considerable successes by relying on highly specialized modelling tools and research methods. This historical record does not *per se* exclude that neuro-psychological findings can inform fruitfully the economic account of decision making. Still, it counsels economists to prudently reflect before embarking on ambitious transdisciplinary Russian campaigns.

In this latter respect, various proponents of NE appear to overestimate the relevance of neural data and findings for the economic account of decision making. For instance, Padoa-Schioppa advocates the adoption of NE models on the alleged ground "that neuroscience can contribute to psychology, and that psychology can contribute to economics" (2008, p.450-1). Regrettably, it remains an open question whether the vague notion of "contribution" to which Padoa-Schioppa refers supports his inference[65]. To give another example, noting that both the equilibrium conditions modelled by economists and the neural computations investigated by NEs are "organized by the general principle of optimization" (McCabe, 2008, p.350) does not license the claim that NE findings are informative to economists. For the practitioners of disciplines

[65] Many notions do not support it. To see this, consider the equivalence relation and the sets (1,2), (2,3) and (3,4). Even though some member of A is equal to some member of B and some member of B is equal to some member of C, no member of A is equal to any member of C.

whose findings are hardly relevant to economic theory employ optimization techniques as a fundamental modelling tool (see e.g. Parker and Smith, 1990, and Sunder, 2006).

More generally, it remains an open question whether the very idea of providing a single, general theory of human behaviour constitutes a viable research project. After all, it is one thing to assert that economists' account of choice behaviour should be consistent with the ones provided by other disciplines. It is quite another thing to claim that economists, psychologists and neuroscientists should employ some common set of constructs to model people's decisions. Now, some NEs (e.g. Caplin and Dean, 2008, and Glimcher, 2010) have recently attempted to complement traditional decision theoretic analyses by identifying and measuring some observable neuro-biological magnitude that is systematically related to decision utilities. Still, the point remains that economics, psychology and neuroscience presently lack the common basic constructs required to provide a unified theory of choice behaviour (I shall expand on this issue in *chapter six*).

To recapitulate, NEs' contributions may well foster the development of interdisciplinary accounts of specific choice patterns or behavioural regularities. Even so, NEs have hitherto failed to provide compelling reasons to think that economists, psychologists and neuroscientists will find it convenient to develop a common framework for modelling decision making. In particular, it remains unclear why the interchanges between NE's parent disciplines should go beyond some limited-scope collaboration. In this respect, what Kahneman (2003, p.165-6) asserted in relation to psychology seems equally pertinent to recent neuroscientific research. As he points out, "the analytical

methodology of economics is stable, and it will inevitably constrain the rapprochement between the disciplines", with "no immediate prospects" of those disciplines "sharing a common theory of human behaviour".

5.C AN ARGUMENT FROM INTRADISCIPLINARY HETEROGENEITY

The recent advances at the interface between economics, psychology and neuroscience have encouraged NEs to raise several criticisms concerning the economic theory of choice. However, the proponents of NE appear to hold contrasting positions with regard to the methodological presuppositions and the explanatory aims of their research. In this section, I assess the scope and the significance of NEs' divergences. In particular, I articulate and defend an *argument from intradisciplinary heterogeneity* which questions the possibility of combining NEs' contributions into a single, general theory of human choice behaviour. Different versions of this argument can be developed depending on which issues one considers. In what follows, I focus on NEs' disagreements concerning (i) the very *definition* of NE, (ii) how NE is expected to *inform* the economic theory of choice, (iii) the *interdisciplinary relationship* that supposedly holds between economics and other behavioural sciences, and (iv) specific *features* of the human neural architecture.

Before proceeding, let me anticipate that the above list does not include all the issues on which NEs hold conflicting positions[66]. Yet, as I argue below, it specifies four respects in which NEs' divergences hamper progress in NE by fragmenting it in a plethora of competing approaches. To be sure, some of these issues are conceptually interconnected. For instance, disagreements about how NE is expected to inform the

[66] To give one example, NEs express heterogeneous views concerning the merits of rational choice theory. For instance, Camerer, Loewenstein and Prelec (2005, p.10) speak of it as a conceptually primitive and empirically inadequate framework for modelling human behaviour. Camerer, instead, alleges that the "rational choice approach has been enormously successful" (1999, p.10575). For their part, Glimcher and Rustichini go as far as to argue that "classical utility theory can be used as a central concept for the study of choice in economics, psychology, and neuroscience" (2004, p.449; see also Glimcher, Dorris and Bayer, 2005, p.253).

economic theory of choice may arise from the fact that different NEs define their discipline in dissimilar terms. Still, they appear to be sufficiently distinct to deserve separate discussion. Having said that, my *argument from intradisciplinary heterogeneity* can be articulated as follows:

P.1 NEs define their own discipline in rather different ways.

P.2 NEs hold dissimilar views as to the extent to which their research is expected to inform the economic theory of choice.

P.3 NEs endorse heterogeneous positions concerning what disciplines will provide the foundations of their framework for modelling human choice behaviour.

P.4 NEs sharply disagree concerning specific features of the human neural architecture.

C Different NEs advocate contrasting modifications of the economic theory of choice.

Let us consider each step of this reasoning in turn. *Premise 1* remarks that NE has been given rather dissimilar characterizations in the literature. As we have seen in section 1.A, some authors (e.g. Glimcher, 2010, and McCabe, 2003a and 2003b) speak of NE as an interdisciplinary enterprise which combines insights from economics, psychology and neuroscience into a single, unified theory of choice behaviour. Other times, NE is presented as a specific application of economic theory to the modelling of the human neural architecture (see e.g. Glimcher, Dorris and Bayer, 2005, and McCabe, 2008). For their part, some NEs (e.g. Camerer, 2003 and 2008a) characterize their discipline as an extension of behavioural and experimental economics. Still differently, other authors

(e.g. Rustichini, 2005, and Zak, 2004) regard NE as an application of neuroscientific techniques and methods to the economic account of decision making.

Premise 2 states that NEs hold heterogeneous views concerning the extent to which NE is supposed to inform the economic theory of choice. To render this point more vivid, let us compare the incremental and the radical approach to NE research we examined in section 1.A. On the one hand, incremental NEs enrich specific economic models in light of neuro-physiological insights, without challenging economists' representation of decision problems as the constrained maximization of some utility function. On the other hand, radical NEs cast doubt on the possibility of modelling decision makers as maximizers of a stable utility function and aim to implement substantial changes in the economic theory of choice.

The incremental/ radical divide is best depicted not so much as an all-or-nothing dichotomy, but as a continuum along which several intermediate positions can be differentiated. In this respect, one may wonder whether the accumulation of incremental modifications amounts to a radical change in economic theory. This question concerns not so much where the boundary between incremental and radical contributions is plausibly set, but the permeability of such a boundary. As we noted in *chapter two*, whether one takes the neural enrichment of specific economic models to constitute a genuine change in economic theory partly depends on how she conceives of the relation between models and theories in economics. Still, the point remains that incremental and radical NE respectively aim to modify the economic theory of choice to a significantly different extent. In this respect, several NE articles appear to face the following

problem. On the one hand, radical contentions are typically too extreme or insufficiently qualified to withstand evidential scrutiny. On the other hand, incremental contributions rarely warrant the propaganda and the excitement that often accompany NE research.

Premise 3 points to NEs' disagreements regarding what disciplines provide the foundation of their hypothesized framework for modelling choice behaviour. Let us consider some examples in support of this assertion. In a recent book, Glimcher argues that it is usually the higher-level abstractions that guide lower-level enquiries and alleges that "insights from economic theory must provide the organizational structure" for NE investigations (2010, p.126; see also Glimcher, 2003). McCabe, instead, contends that elaborating informative NE models requires "both a topdown approach […] from economics and a bottom-up approach […] from cognitive neuroscience" (2008, p.349). Still differently, Camerer argues that "because economics is the science of how resources are allocated by individuals […] the psychology of individual behaviour should underlie and inform economics, much as physics informs chemistry" (1999, p.10575). For their part, Zak and Denzau go as far as to assert that the "methods and findings in the biological sciences need to be incorporated directly into economics if the discipline is to continue to produce relevant insights into human behavior" (2001, p.32).

Now, it is true that NEs may consistently endorse some of these claims. For instance, one could argue - in line with Wilson (1998, p.206) - that "it is in biology and psychology that economists and social scientists will find the premises needed to fashion more predictive models". Still, the previous assertions express quite dissimilar

positions concerning the interdisciplinary relationship that allegedly holds between NE and its parent disciplines (see also Fumagalli, 2010). These divergences, in turn, cast doubt on the possibility of combining NEs' contributions into a cumulative case in favour of NE. In section 7.B, I shall assess the plausibility of NEs' assertions in light of the vast literature on intertheoretic reduction[67]. For now, let me give one reason for being cautious concerning some authors' transdisciplinary fervour. In their articles, the proponents of radical NE fall short of specifying *why* exactly the ongoing cooperation at the interface between economics, psychology and neuroscience would prelude major interdisciplinary rearrangements. In particular, they are disappointingly vague concerning *what constructs* would replace the ones that are currently employed by economists.

As noted by *premise 4*, NEs often disagree also in relation to several features of the human neural architecture. By way of illustration, consider the divergences arisen about the functional localizability of specific cognitive processes. Some NEs (e.g. Camerer, Loewenstein and Prelec, 2004 and 2005) seem to erroneously assume that different neural areas are respectively associated with rational and irrational behaviour. On the contrary, others (e.g. Glimcher, Dorris and Bayer, 2005, and Preuschoff et al., 2006) rightly point out that the neural substrates of rational decision making are unlikely to be topographically localized. In particular, they question the alleged association (*i*) between evolutionarily older neural areas and emotional circuitry, and

[67] Classic works include Nagel, 1961 and 1974, Schaffner, 1967, Fodor, 1974, and Churchland, 1981 and 1985. For some recent publications in the philosophy of neuroscience, see e.g. Bickle, 1998 and 2003, Craver, 2007, Craver and Alexandrova, 2008, and Sullivan, 2009.

(*ii*) between evolutionarily more recent brain regions and higher cognitive functions[68]. As Massey puts it, "the neural anatomy essential for full rationality […] is a very recent evolutionary innovation", and rational abilities "did not replace emotionality as a basis for human interaction", but "were gradually added to pre-existing and simultaneously developing emotional capacities" (2002, p.16; see also Anderson, 2006 and 2007, on how evolutionarily older neural areas are often engaged by many cognitive tasks).

On the basis of the previous premises, the argument from intradisciplinary heterogeneity *concludes* that NEs advocate contrasting modifications of the economic theory of choice. To be sure, not all NEs' contrasts call the advancement of NE research into question. To see this, consider NEs' disagreements over specific features of the human neural architecture. Most of these divergences will be presumably settled thanks to further developments in NE scanner technology and experimental practices. Moreover, the mere fact that some issues in neuroscience are still unresolved does not imply that economists should "wait until neuroscience has grown more mature before we try to accommodate their insights and findings" (Vromen, 2007, p.161). For NEs may succeed in building informative neurally enriched models without taking a definite position on all of those issues.

NEs could put forward additional reasons to resist deriving far-reaching implications from their current disagreements. In particular, they may contend that the existence of

[68] There are various reasons to doubt that two anatomically distinct brain systems respectively underlie rational and irrational decisions. For instance, various areas traditionally associated with emotions contribute to higher cognitive functions such as the encoding of rewards (see e.g. Glimcher, 2009, and McClure et al., 2004b; see also Friston, 2002, on the functional and anatomical interdependences between different areas).

divergences is quite *expectable* given that NE is in its first stages of development. Indeed, they might even argue that the existing contrasts are signs of a lively and promising debate. After all - the reasoning would go - NE is still in its infancy, and it is *desirable* that several approaches compete for defining the canons of its orthodoxy. As it is occasionally claimed in methodological discussions: "Don't bother too much at first about the compatibility of different theories. Just wait, and let inter-theoretic competition decide which candidates will stand the test of time".

Prima facie, the aforementioned recommendation seems to offer sensible advice and nicely fits with the methodological prescriptions provided by some philosophers of science. Consider, for example, the Lakatosian caveat (1970) that research programs often grow in an ocean of anomalies and that adopting an exceedingly severe stance towards novel conjectures might lead one to prematurely abandon promising research avenues[69]. Yet, as Lakatos' critique of degenerating research programs persuasively illustrates, even someone who advocates letting many flowers blossom is still allowed to weed.

A proponent of NE may protest that economics itself, in its early days, was characterized in dissimilar terms by prominent economists, and that nevertheless these discrepancies did not preclude its progress. However, the mere fact that economics progressed in spite of definitional and methodological diversity by no means excludes that NEs' divergences hinder the consolidation of their discipline. To be sure, the existence of profound contrasts between NEs does not *per se* preclude the development

[69] See also Feyerabend (1962, 1970 and 1975) for the claim that developing several competing approaches fosters intradisciplinary progress.

of informative NE models. For even if distinct NEs hold inconsistent positions on several substantial issues, one (or some) of their approaches may still serve as a basis for constructing instructive models. Nonetheless, the divergences we examined above cast serious doubts on NEs' attempts to develop a single, general theory of human choice behaviour. For those divergences concern - not just different terminological options, or secondary aspects of NE research, but - the very definition of NE and how NE is supposed to inform economic theory. To put it differently, it is hard to see how NEs can provide a unified theoretical framework for modelling human choice behaviour, when they agree neither on the explanatory aims of their research nor on what constructs will serve as the foundation of their account of decision making.

An additional reason to think that NEs' divergences hamper progress in their discpline relates to the explanatory goals that some leading NEs claim to pursue. If NEs rested content with proposing a series of unrelated models, each designed to account for a specific phenomenon (see e.g. Kosfeld et al., 2005, and Zak et al., 2005, 2006 and 2007, on how oxytocin may affect agents' trust and generosity in particular choice settings), then reconciling their approaches would not seem to constitute a paramount issue. Yet, when it comes to providing a "single, general theory of human behaviour" (Glimcher and Rustichini, 2004, p.447) and "an entirely new set of constructs" for the analysis of decision making (Camerer, Loewenstein and Prelec, 2005, p.10), reducing the fragmentation which characterizes current NE research becomes a particularly pressing concern. In this respect, various authors aptly warn against the risk of "giving rise to a proliferation of different models that are mutually incompatible not only in terms of the details, but also in terms of the overarching approach" (Caplin, 2008, p.359).

To conclude, one views with favour NEs' short term aspiration to integrate insights and findings from different behavioural sciences. Still, there are reasons to question their long term ambition to provide a unified framework for modelling human choice behaviour. The reasoning I presented above can be summarized as follows. The accounts proposed by different NEs are characterized by profound dissimilarities, which concern the central tenets of NE. By itself, the existence of these contrasts does not prevent NEs from constructing informative neuro-psychological models of choice. At the same time, it casts serious doubts on their attempts to provide a unified theoretical framework for modelling decision making. In particular, the ongoing fragmentation of NE research into heterogeneous approaches makes it doubtful that NEs will foster a grand unification of the behavioural sciences. In this respect, economists should reflect carefully before endorsing NEs' claims regarding the interdisciplinary consilience - not to say unification - of the behavioural sciences (see e.g. Camerer, 1999, p.10575, and Camerer, Loewenstein and Prelec, 2004, p.573). For some of these claims seem inspired more by an unreflective overextension of the methods adopted in particular branches of neuroscientific research than by principled reflections concerning how NE is likely to inform its parent disciplines.

CHAPTER SIX

ON THE FUTILE SEARCH FOR TRUE UTILITY

In this chapter, I provide a case study to illustrate how the conceptual differences between NE's parent disciplines constrain the relevance of neuro-psychological findings for the economic theory of choice. In the literature at the interface between economics, psychology and neuroscience, several authors argue that economists could develop more predictive and explanatory models by incorporating insights concerning agents' hedonic experiences. In particular, some (e.g. Camerer, 2008a, p.45, and Camerer, Loewenstein and Prelec, 2005, p.15) go as far as to contend that agents' utility is literally computed by specific neural areas and urge economists to complement or even substitute their notion of utility with some neuro-psychological constructs.

Economic modellers and methodologists have ascribed a variety of meanings to the concept of utility. Indeed, different authors employed this term in so different senses that one wonders whether we can meaningfully speak of utility as a unified concept[70]. In the recent economic literature, three senses of the word stand out as particularly prominent. Firstly, contemporary decision theorists typically regard utility as a mathematical representation of preferences to be inferred from observed choices[71]. Secondly, other authors employ the term utility to refer to a hedonic magnitude

[70] In this respect, it is telling that several leading economists (see e.g. Pareto, 1909, on "ophelimity", Fisher, 1918, on "wantability", and Pigou, 1920, on "desirability") proposed giving a separate name to utility to differentiate their analyses from the works of psycho-physiologists (see Colander, 2007, p.220, for a similar remark).

[71] As we noted in section 1.B, the idea (see e.g. Savage, 1954) is that if an agent's preferences satisfy specific consistency requirements, then these preferences can be represented by a unique probability function and by a utility function unique up to positive linear transformations such that of any two options the one with higher expected utility will be preferred.

reflecting agents' experiences of pleasure and pain (see e.g. Kahneman, 2000, p.2, and Kahneman, Wakker and Sarin, 1997, p.375). Finally, some speak of utility as a value signal that can be directly measured in the activation patterns of specific neural areas (Camerer, Loewenstein and Prelec, 2004, p.556) and allege that "map-like structures in the brain […] are actually the subject of economic theory" (Glimcher, Dorris and Bayer, 2005, p.238). These three notions are often named *decision utility*, *experienced utility* and *neural utility* respectively. Below I refer to the last two notions as 'true utility', since they are usually employed to indicate some objective quantity rather than a mathematical construct such as decision utility.

The contents of this chapter are organized as follows. In section 6.A, I draw some conceptual distinctions between the three aforementioned notions of utility and outline various methods for measuring experienced and neural utility. In sections 6.B and 6.C, I critically assess the thesis that economists should base decision theoretic analyses on experienced or neural utility. In doing so, I examine some critical issues regarding the definition and measurability of these notions of utility. Moreover, I provide various reasons to doubt that economists should replace decision utility with some notion of true utility as a central concept of decision theory. My critique can be seen as a response to those NEs who advocate the replacement of "the mathematical ideas used in economics" with "more neurally detailed descriptions" (Camerer, 2005) and speak of "substitut[ing] familiar distinctions between categories of economic behaviour" with the ones adopted in other disciplines (Camerer, Loewenstein and Prelec, 2005, p.15).

Before proceeding, let me put forward three preliminary caveats regarding the focus of this chapter. Firstly, I do not explore the historical interrelations between distinct notions of utility, as the cogency of my considerations does not rest on how one reconstructs them. For the purpose of this enquiry, it suffices to note that the possibility of measuring experienced utility was prefigured well before the advent of standard decision theory (see e.g. Edgeworth, 1881 [1967]), whereas the notion of neural utility has appeared in economic discussions only in the last decade. Secondly, the concept of utility is amenable to various interpretations besides the three ones I discuss here (see e.g. Bradley, forthcoming, on subjective and objective interpretations of utility). I do not expand on those interpretations, as they are mostly orthogonal to the focus of my investigation. Finally, decision theory can be given both descriptive and normative interpretations. In this chapter, I focus prevalently on the former. In the next chapter (section 7.C), I shall discuss some normative applications of experienced and neural utility measurements to economic welfare analyses.

6.A FROM DECISION UTILITY TO TRUE UTILITY

Traditional decision theorists treat agents' choices as primitives and regard them as rational to the extent that the underlying preferences satisfy specific consistency requirements (see Bhattacaryya et al., 2011). In such a context, no explicit assumptions are made about the neuro-psychological substrates of agents' preferences. This notion of rationality concerns whether one's preferences are consistent with one another, and places no constraints on "the content of any belief and desire taken in isolation" (Bradley, forthcoming, p.1; see also Sugden, 1991). Yet, as we have seen in section 1.B, even these consistency requirements are occasionally violated by real life decision makers. In the recent NE literature, several authors advocate complementing or replacing the formal notion of decision utility with specific neuro-psychological constructs. The idea is to view agents as "rank[ing] outcomes in terms of some objective measure" (Bruni and Sugden, 2007, p.170) and to ground decision theoretic analyses on such a measure (Glimcher and Rustichini, 2004, p.452, and Kahneman and Sugden, 2005, p.161).

Various candidates for substituting decision utility have been proposed. Below I focus on experienced utility and neural utility in turn, explicating some major conceptual distinctions between them and decision utility. The expression *experienced utility* is usually employed to indicate a hedonic magnitude reflecting agents' experiences of pleasure and pain which is to be measured with psycho-physical methods (Read, 2007, p.58). This usage of the term 'utility' is more restrictive than the one adopted by early utilitarian philosophers (e.g. Bentham, 1789 [1907], propos. I, II and X), who often

speak of utility maximization in both a descriptive and a normative sense. Measures of experienced utility and decision utility diverge in a wide range of circumstances. For instance, for there to be a general correspondence between decision utility and experienced utility, agents' choices must reflect reasonably accurate predictions of the hedonic consequences of their actions (Kahneman and Sugden, 2005, p.167). Still, people frequently fail to anticipate or take into account the effects various actions have on their future preferences (see e.g. Kahneman and Snell, 1990, and Snell et al., 1995).

Three notions of experienced utility can be usefully distinguished, namely: *instant utility*, a moment-based measure which reflects the valence and the intensity of agents' ongoing hedonic experiences (Kahneman, 2000, p.17); *remembered utility*, a memory-based measure "inferred from a subject's retrospective reports of the total pleasure or displeasure associated with past outcomes" (Kahneman, Wakker and Sarin, 1997, p.376); and *anticipated utility*, which reflects "a person's ex ante beliefs about the hedonic quality of future experiences" (Kahneman and Sugden, 2005, p.174). In such a context, measurements of so-called *total utility* can be obtained by taking the temporal integral of instant utility measurements (Kahneman, 2000, p.17, and Kahneman, Wakker and Sarin, 1997, 389). The idea was allegedly anticipated by Edgeworth, who suggested that an imaginary instrument - the hedonimeter - could measure individuals' hedonic experiences[72].

[72] In the words of Edgeworth (1881 [1967], p.101): "imagine an ideally perfect instrument, a psychophysical machine, continually registering the height of pleasure experienced by an individual… The continually indicated height is registered by photographic or other frictionless apparatus upon a uniformly moving vertical plane. Then the quantity of happiness between two epochs is represented by the area contained between the zero-line, perpendiculars thereto at the points corresponding to the epochs, and the curve traced by the index".

As illustrated by several studies, agents' decisions significantly correlate with remembered utility (e.g. Fredrickson and Kahneman, 1993, and Redelmeier et al., 2003), and individuals typically prefer to undergo experiences to which they assign higher remembered utility (Kahneman et al., 1993 and 1997). Two findings regarding the determinants of remembered utility are worth mentioning (see e.g. Kahneman et al., 1993, and Schreiber and Kahneman, 1996). Firstly, the duration of an episode, even when accurately known, has little impact on its remembered utility (*duration neglect*). And secondly, remembered utility highly correlates with the average of the most intense value of instant utility recorded during an episode and the instant utility recorded at the end of such an episode. As a result, the remembered disutility of an aversive episode can be reduced by adding an extra period of discomfort which reduces the peak-end average (*violation of temporal monotonicity*).

The notion of *neural utility* relates to the activation patterns of particular areas in the neural architecture. The idea is that desirability "is realized as a concrete object, a neural signal in the human brain, rather than as a purely theoretical construction" (Glimcher and Rustichini, 2004, p.452). According to the proponents of neural utility, expected utility theory can be employed to represent - not just consistent decisions, but also - specific areas' activation patterns. In their view, "the utility calculations that people were assumed to do really happen in the brain" (Park and Zak, 2007, p.50) and traditional economic models of choice constitute a "limiting case of a neural model with multiple utility types" (Camerer, Loewenstein and Prelec, 2004, p.564). Over the last few years, various authors have suggested that anatomically delimited neural populations compute desirability signals (see e.g. Hare et al., 2008, Kable and Glimcher,

2007, Padoa-Schioppa and Assad, 2006 and 2008). In particular, some document that specific areas' activations track variations in relative expected rewards in various experiments (Platt and Glimcher, 1998 and 1999, and Dorris and Glimcher, 2004; see also Basso and Wurtz, 1997 and 1998).

Two main positions regarding how this evidence should be interpreted can be differentiated. According to some authors, the aforementioned findings demonstrate that the subjects who behave in accordance with the axioms of expected utility theory "do so because they neurally represent something having the properties of utility - a neural activation that encodes the desirability of an outcome in a continuous monotonic fashion" (Glimcher, 2010, p.133-4; see also Preuschoff et al., 2006). Others (e.g. Camerer, 2007 and 2008b, and Quartz, 2008) take the available evidence to provide direct support to standard utility theory. Let us focus on this latter interpretation. As noted by Vromen (2010a, p.33), it is certainly interesting that expected utility theory can be applied to entities and processes that few economists had previously regarded as the target of their predictions. Yet, the mere fact that such a theory can be used to describe the activation patterns of specific neural *areas* falls short of implying that it accurately predicts *agents*' choice behaviour.

Before concluding this section, let me emphasize one important respect in which the notions of experienced utility and neural utility differ. The notion of neural utility refers to a magnitude that is taken to be measurable objectively in specific areas' activation patterns. The proponents of experienced utility, instead, often leave it ambiguous whether agents' experienced utility is determinable solely by means of third-person

measurements or whether it inherently reflects also agents' first-person evaluations (for a recent debate on first-person and third-person perspectives, see e.g. Chalmers, 1996 and 2004, Dennett, 1991, and Papineau, 2002 and 2003). To be sure, some proponents of experienced utility speak of it as a magnitude amenable to objective measurements and allege that "in spite of the immense diversity in the occasions that evoke pleasure […] the hedonic attribute that they share is salient and readily recognized" (Kahneman, Wakker and Sarin, 1997, p.380). Still, these authors rarely specify what exactly the shared hedonic attribute of heterogeneous experiences consists in. As we shall see in the next two sections, the conceptual difference between experienced utility and neural utility has important implications regarding both the measurability and the interpersonal comparability of these notions of utility.

6.B EXPERIENCED UTILITY: CONCERNS

The proponents of experienced and neural utility frequently urge economists to employ neuro-psychological findings and constructs in developing their models of choice. Nonetheless, the notions of decision utility, experienced utility and neural utility differ profoundly, and replacing decision utility with neuro-psychological constructs would involve major changes in traditional decision theory. In this and the next section, I examine several empirical and conceptual concerns arising in relation to experienced utility and neural utility measures respectively. In doing so, I argue that the available neuro-psychological evidence does not substantiate the thesis that economists should substitute decision utility with some notion of true utility. With regard to *experienced utility*, I shall discuss three issues in turn, which respectively concern the measurability, integrability and interpersonal comparability of experienced utility.

i) *Measurability*: the proponents of experienced utility often appear to presuppose that such a magnitude is measurable in reliable and accurate terms. Even so, different authors seem to hold dissimilar positions regarding what methods are to be used to measure it. To see this, consider the characterization of instant utility provided by Kahneman, Wakker and Sarin (1997). In their view, instant utility "can be derived from immediate reports of current subjective experience or from physiological indices" (p.376 and 388). Regrettably, these measures can significantly diverge, and it is unclear which of them should be regarded as more accurate and reliable.

One might rebut that these concerns can be addressed by developing more robust measures of experienced utility. Over the last two decades, several methods for measuring agents' hedonic states have been proposed. Among these, we find the experience sampling method (Brandstatter, 1991, Csikszentmihalyi, 1990, and Stone et al., 1999) and the day reconstruction method (Kahneman et al., 2004). The former consists in asking subjects at random times during the day to answer questions about their current affective state. In the latter, each participant is asked to reconstruct her previous day into short episodes and then rate each episode in terms of various feelings. These methods seem less liable than others to biases and experimental confounds[73]. Even so, subjects' responses can noticeably vary with their circumstances (e.g. the perceived level of experienced utility to which one is accustomed), how the periods composing the examined days are framed and how the provided response scales are interpreted (e.g. distinct persons may employ the response categories differently).

This variability, in turn, constrains the reliability and the interpersonal comparability of experienced utility reports (see also point *iii* below). By way of illustration, consider the following *Garden of Eden* example. Imagine that Adam and Eve have to provide a hedonic report regarding the experience of eating the fruit of the forbidden tree. Suppose that both of them are sincere and have adequate access to their inner hedonic states. Assume further that Adam states that he 'really liked' eating the apple, while Eve just says that she 'liked' eating it. Their reports by no means license the claim that Adam had more hedonic pleasure when eating the apple than Eve did. For Adam and

[73] For example, many happiness reports reflect adaptation to fortunate and unfortunate circumstances (Brickman et al., 1978, Frederick and Loewenstein, 1999, and Ubel et al., 2005) and life satisfaction measures respond to factors such as agents' transient mood and the weather (Kahneman and Krueger, 2006, and Schwarz and Strack, 1991).

Eve may be rating their experiences in terms of dissimilar response scales, i.e. non-equivalent mappings between linguistic expressions of satisfaction and inner hedonic states. Averaging responses over individuals can mitigate the effects of individual variability in the use of response scales. Yet, even though researchers attempt to "anchor response categories to words that have a common and clear meaning across respondents […] there is no guarantee that respondents use the scales comparably" (Kahneman and Krueger, 2006, p.18-9; see also Angner, 2011, for a discussion).

In such a context, a further worry arises regarding the intertemporal variability of experienced utility measures. To explicate this concern, let me provide the following *First Date* example. When remembered shortly thereafter, a first date can give one an unsurpassable pleasure, awakening her noblest and most elevated sentiments. However, the remembered utility associated with such an episode can vary dramatically depending on how the subsequent relationship unfolds. Moreover, one may continue to derive utility (or disutility) from that episode long after it ends by remembering it, by experiencing how her inner life has been affected by it, etc. A proponent of experienced utility might allege that by taking the temporal integral of these remembered utility measures, we could provide a complete measurement of the remembered utility associated with such an episode. Still, it remains unclear how informative it would be to know what the remembered utility associated with an episode at a given moment is. For the value of such a measure can significantly fluctuate depending on factors such as what memories the agent happens to recall, the valence she ascribes to those memories, etc.

ii) *Integrability*: Kahneman (2000, p.6-8) imposes various requirements on how total utility is to be constructed from profiles of instant utility as a measure of what he calls "objective happiness". Among these, we find: *separability*, according to which "the order in which moment-utilities are experienced does not affect total utility"; *time neutrality*, which states that "all moments are weighted alike in total utility"; and so-called *ordinality*, which requires that "any two moments of experience can be compared, to establish which of them carries the higher hedonic value" (Kahneman, Wakker and Sarin, 1997, p.389). As argued by Kahneman (2000), separability and time neutrality are necessary - and, together with inclusiveness (see below) and ordinality, sufficient - for representing utility profiles as a decumulative function showing the amount of time an agent spends at each level of pleasure and pain. Nonetheless, each of these assumptions is vulnerable to severe criticisms. Let us consider those assumptions in turn.

Separability requires that "the contribution of an element to the global utility of the sequence is independent of the elements that preceded and followed it" (Kahneman, 2000, p.7). *Prima facie*, this assumption seems vulnerable to obvious counterexamples. For instance, the hedonic outcomes associated with taking a nap and reading Wittgenstein's *Tractatus* tonight depend on whether you spent the whole day sleeping or compulsively computing Hamiltonians. According to Kahneman, Wakker and Sarin (1997, p.391), separability of instant utilities is justifiable provided that one's measure of instant utility "incorporates all order effects and interactions between outcomes" (*inclusiveness*). Regrettably, the proponents of experienced utility leave it unclear on what grounds we are to ascertain whether inclusiveness is satisfied. Moreover, it seems

doubtful that our measures of instant utility reflect all the affective consequences of previous experiences and the anticipation of future events. To give one example, one's past experiences can influence her current hedonic states in ways which often elude her awareness and existing physiological indicators of instant utility.

Regarding time neutrality, total utility measures differ markedly from decision utility and remembered utility ones, which typically assign more weight to outcomes that respectively occur early and late in a sequence. According to Kahneman, Wakker and Sarin, a time-neutral perspective is "appealing, both as a rule of personal prudence and as a principle of social planning" (1997, p.393). Yet, much controversy surrounds this issue. For instance, one may contend that a decrease in the psychological connectedness between our present self and our future selves gives us a reason to discount the utility of those selves (Parfit, 1982 and 1984). More generally, the inherent temporal situatedness of the position from which decision makers evaluate outcomes makes it dubious that the temporal distance between an outcome and the moment in which such an outcome is evaluated is irrelevant to its evaluation. I am not concerned here with settling the controversy over the rationality of discounting the future. Still, the proponents of total utility cannot gloss over this issue, if they are to provide decision theorists with a convincing case for relying on such a measure.

As to ordinality, Kahneman (2000, p.12) himself acknowledges that whether agents can meaningfully compare distinct hedonic states is ultimately an empirical question. I shall not expand on the vast literature on incomparability cases (see e.g. Chang, 1997 and 2002, and Raz, 1986). For the purpose of this enquiry, it suffices to note that people are

not always able to express definite preferences between heterogeneous experiences (e.g. suppose you had to choose between undergoing some excruciating physical pain and suffering from a profound depression). Indeed, it appears that situations where people are at loss in formulating accurate evaluations of specific "moments of experience" can be multiplied *ad nauseam*.

iii) *Interpersonal comparability*: according to Kahneman, "a distinctive neutral point [neither pleasant nor unpleasant] permits comparisons across situations and persons" (2000, p.7). The idea is that although the stimulus giving rise to a neutral experience varies across contexts, "the neutral experience itself is constant" and can be used to ground meaningful interpersonal comparisons of experienced utility (Kahneman, Wakker and Sarin, 1997, p.380; see also Kahneman and Varey, 1991). However, it is an open question whether the neutral point can serve this purpose. To be sure, one may insist that we often make interpersonal comparisons on the alleged ground that the functions relating hedonic states to physical variables are qualitatively similar across people (Kahneman et al., 1997, p.380). Yet, appealing to this vague notion of qualitative similarity does not license wide-ranging conclusions regarding the interpersonal comparability of experienced utility reports (see e.g. Harsanyi, 1955, Arrow, 1977, and Griffin, 1986, p.75-124)[74].

[74] Indeed, it is unclear whether the mere fact that some experiences are deemed to be neither pleasant nor unpleasant by an agent makes them the subject of an informative categorization. For instance, suppose that you regarded reading *The Karamazov Brothers*, dreaming of walking on Mars and counting blades of grass in your garden as neutral experiences. Is this fact of any particular significance when it comes to assessing the desirability, the choice worthiness, etc. of these experiences?

In such a context, one may wonder whether an agent's experienced utility in a given situation can be measured in purely physical terms. This question relates to an issue that has been discussed extensively by philosophers of mind, namely whether having complete knowledge of an agent's physical states is sufficient for acquiring complete knowledge of her mental states (see e.g. Broad, 1925, Jackson, 1982 and 1986, and Nagel, 1974 and 1986). Now, establishing the objective measurability of experienced utility in purely physical terms would require one to show that, once all the relevant physical data have been collected, there is nothing more to learn regarding what it is like to enjoy a particular level of experienced utility. At present, this constitutes an open question in the philosophy of mind (see e.g. Chalmers, 1996 and 2004, Dennett, 1991, Metzinger, 2000, and Papineau, 2003, for a debate).

6.C NEURAL UTILITY: CONCERNS

Let us focus on *neural utility*. In what follows, I examine in turn some concerns related to the informativeness of neural utility measures, the reducibility of measures of experienced utility to measures of neural utility, and the underdetermination of neuro-psychological theories by the available evidence.

i) My first concern relates to the *informativeness* of neural utility measures. In order to identify which areas generate neural utility signals, one typically has to "rely on correlations with directly interpretable indicators of [well-being]" such as observed choice behaviour or reported expressions of satisfaction (Bernheim, 2009, p.31). In this respect, some authors worry that neuro-physiological measures of utility - being constructed on the basis of observed choices and reported measures of satisfaction - may be incapable of providing insights which correct or disconfirm those data. As Bernheim (2009, p.37) provocatively puts it: "why use brain scans to construct noisy predictions of a subject's answers to questions about happiness and/or satisfaction when we can simply pose those questions directly?".

One pressing challenge for the advocates of neural utility is to offer insights concerning standard economic variables that economists cannot acquire by examining those variables directly (see also Rubinstein, 2008). To address this concern, NEs may attempt to identify the functions linking hedonic states to neural states in some experimental population and then employ these data with other subjects or the same subjects at a later time. Yet, the point remains that identifying which areas generate

desirability signals requires one to use behavioural and psychological data. This, in turn, constrains the extent to which neuro-physiological investigations can inform decision utility and experienced utility measurements.

In recent years, some authors have argued that identifying some neuro-biological magnitude that is systematically related to decision utility provides "fundamental economic insights" (Caplin and Dean, 2008, p.669; see also Glimcher, 2010). Now, it would be implausible to deny that neuro-biological evidence can constrain specific models of choice (see e.g. Caplin et al., 2010, on how knowledge of dopaminergic circuits restricts the set of learning processes that can be plausibly thought to underlie choice behaviour). Yet, it remains unclear how exactly those neuro-biological findings bear on standard decision theory. To see this, suppose - following Glimcher (2010, p.135-7) - that our current instruments enable us to identify anatomically delimited populations of neurons whose mean firing rates are linearly related to decision utility. Even so, profound differences remain between the firing rate of some neurons and an abstract economic construct like decision utility (e.g. only the former is a cardinal object). These differences, in turn, hinder NEs' attempts to bridge the "current conceptual gap" (Caplin and Dean, 2008, p.663) between the economic and the neuroscientific accounts of choice behaviour.

ii) Over the last few years, several insights regarding the neuro-physiological correlates of experienced utility have been provided. Consider, for example, anticipated utility. Some authors (Elliott et al., 2000, and Knutson et al., 2001a) document how particular areas differentially activate during anticipation of specific rewards depending on the

magnitude of the anticipated outcomes. Others (Knutson et al., 2001b) show that various regions exhibit increased activation when subjects anticipate some gain but deactivate by the time subjects receive it. Still others (Knutson et al., 2003) illustrate how the activations of specific neural populations vary depending on whether a subject fails to receive an anticipated reward or expectedly receives no reward.

In commenting over the advancement of neuro-physiological research, Kahneman contends that "the prospects are reasonably good for an index of the valence and intensity of current experience, which will be sensitive to the many kinds of pleasure and anguish in people's lives" (2000, p.13). Now, neuroscientists will certainly provide increasingly accurate characterizations of the neural underpinnings of experienced utility. Still, it is doubtful that measuring the activations of specific neural areas enables us to adequately account for the remarkable diversity of people's hedonic experiences. My point is not just that the neural substrates of some hedonic states are topographically and functionally too complex to be captured by current neuro-physiological investigations. Rather, my main concern is that any claim to have disclosed the neural constituents of people's subjective experiences would be philosophically naïve. After all, one may well reiterate that experienced utility measures supervene on neural utility ones, i.e. that every change in agents' hedonic states reflects some modification in their neural areas' activation patterns. Yet, it is an open question whether experienced utility measures, which relate to people's subjective experiences, can be thoroughly *reduced* to third-person neural utility data. Let me expand on this point.

In the philosophy of mind and of cognitive sciences, several anti-reductionist lines of argument have been developed. For example, some authors (e.g. Fodor, 1974) question the prospects of intertheoretic reductions by pointing to the multiple realizability of mental states at the physical level. Others, instead, deny that materialistic neuroscience can account for the phenomenological features of our mental states (e.g. Nagel, 1974, and Jackson, 1982), the semantic content of our thoughts (e.g. Searle, 1980 and 1990), or the holistic character of the mental (e.g. Davidson, 1980). I am not concerned here with assessing the merits of these arguments. Still, one does not have to be a Cartesian substance dualist to doubt the reducibility of experienced utility measures in terms of neural utility ones. To give one example, the inherent vagueness of many phenomenal accounts of conscious experiences severely constrains their susceptibility to neuro-physiological reduction. In the words of Papineau (2003, p.208-9), scientific research can "narrow down the possible material referents of phenomenal concepts" (e.g. being in pain). Yet, due to the vagueness of our phenomenal concepts, there will typically remain several candidate material referents for each of those concepts.

iii) The practitioners and the philosophers of various behavioural sciences frequently discuss the issue of *underdetermination* of theory by the evidence. Recent neuroscientific research on neural utility faces a particularly severe instance of this problem. The idea is that the available neural evidence typically underdetermines the identification of the processes by means of which distinct areas' activations are combined into desirability signals. That is to say, different models of how neural utility is computed may be elaborated, and the available neural evidence rarely enables us to

discriminate between them[75]. Below I distinguish two issues associated with this problem. The first relates to how many *stages* comprise the generation of neural utility signals. The second concerns which neural *areas* are involved in those computations. Let me consider these two issues in turn.

In exploring the neural underpinnings of reward evaluation, several cognitive and computational neuroscientists advocate the so-called *common currency hypothesis*, according to which the brain ranks outcomes and actions in terms of a unique neural currency (Landreth and Bickle, 2008). The idea can be summarized as follows. Economic agents often have to decide between complex outcomes (e.g. think of various courses of action). For principled choice to be possible, the value of the available options must be represented in terms of some common currency. In the words of Montague and Berns, without an internal currency in the nervous system "a creature would be unable to assess the relative value of different events […] To decide on an appropriate behavior, the nervous system must estimate the value of each of these potential actions, convert it to a common scale, and use this scale to determine a course of action" (2002, p.276)[76].

In such a context, two pressing concerns arise in relation to the task of identifying how many *stages* comprise the computation of neural utility signals. Firstly, there is the

[75] Philosophers speak of underdetermination of theory by the evidence in a variety of senses. For instance, as noted by Okasha (2000), such an expression has been employed to mean that the available data: are logically compatible with more than one theory (Newton-Smith, 1978); can be explained by more than one theory (English, 1973); are entailed by more than one theory (Quine, 1975); and equally support more than one theory (Bergstrom, 1993). In this section, I use the term underdetermination to indicate that the available data can be employed to support more than one theory.

[76] See also Shizgal, 1997, and Shizgal and Conover, 1996, for some evidence in favour of the unique currency hypothesis based on brain-stimulation studies.

problem of ascertaining whether the activation patterns observed at a given moment reflect "aggregation of feelings at different points in time, or immediate feelings driven by anticipated [or remembered] outcomes" (Bernheim, 2009, p.31). And secondly, one faces the challenge of understanding how exactly the observed neural signals are aggregated over time. As noted by Bernheim, both problems appear to be quite challenging. For instance, resolving the former problem is complicated by the fact that the very act of anticipating or reminiscing specific experiences is likely to foster neural activations besides those associated with the hedonic correlates of those experiences. As to the latter challenge, the development of accurate empirical generalizations regarding the intertemporal aggregation of neural signals is constrained by the variability that some areas' activations exhibit across subjects and choice settings.

Additional concerns arise regarding the identification of which neural *areas* are involved in the processing of desirability signals. As we have seen in the first section, researchers (see e.g. Platt and Glimcher, 1998 and 1999) identified various regions whose activations correlate with the relative expected value of some rewards in specific decision contexts[77]. These findings, however, are obtained in highly constrained experimental settings and fall short of implying that the identified areas compute desirability signals across decision contexts. Moreover, various studies suggest that different kinds of reward tend to engage distinct neural circuitries, with dissimilar areas activating across choice settings (see e.g. Elliott et al., 2000, McClure et al., 2004a, and O'Doherty et al., 2002).

[77] Until recently, most neural evidence was obtained in studies of non-human primates engaged in unsophisticated experimental tasks (e.g. repetitive choices between two primary rewards). The limited complexity of the examined tasks, coupled with the anatomical and functional dissimilarities between humans' and other primates' neural architectures (see e.g. Allman et al., 2002), constrain the generalizability of those findings to traditional decision problems.

The identification of an anatomically separate network of areas responsible for computing desirability signals is further complicated by the fact that the areas generating those signals often contribute to other cognitive and computational tasks (see e.g. Anderson, 2006 and 2007, and Cabeza and Nyberg, 1997 and 2000). To be sure, dopaminergic circuits are often taken to play a central role in the computation of reward evaluations (e.g. Schultz et al., 1997, and Schultz, 2000). In particular, several studies suggest that dopaminergic activity encodes deviations between anticipated and obtained rewards (Bayer and Glimcher, 2005, Morris et al., 2006, and Schultz and Dickinson, 2000), with some authors going as far as to regard dopamine as "a key input into the construction of utility" (Caplin and Dean, 2008, p.669). Still, other theories of dopaminergic function have been proposed (e.g. Berridge and Robinson, 1998, and Redgrave and Gurney, 2006). Moreover, natural rewards do not always activate dopaminergic circuits, and other neural populations besides the dopaminergic ones have been shown to contribute to reward evaluations (see e.g. Hare et al., 2008).

To conclude, in the recent literature at the boundary between economics, psychology and neuroscience, various measures of experienced and neural utility have been developed. In light of this diversity, it is highly advisable to distinguish different notions of utility rather than presume that "a single, unifying concept […] motivates all human choices and registers all relevant feelings and experiences" (Kahneman and Krueger, 2006, p.4). As I argued above, there are profound dissimilarities between decision utility and neuro-psychological notions of utility. Moreover, several empirical and conceptual concerns arise in relation to experienced and neural utility measures.

Now, let us suppose - for the sake of argument - that we could measure experienced and neural utility in reliable and accurate terms. One may advocate using these measures to complement traditional decision theory. Even so, additional argument is needed to license the conclusion that decision theoretic analyses should be based on these notions. Indeed, there is a sense in which the search for true utility appears to be - not just incomplete, but also - futile. After all, we can expect to find increasingly accurate neuro-psychological correlates of observed choices across decision settings. Yet, it remains hard to see how those empirical investigations are supposed to foster the replacement of a mathematical construct such as decision utility. Paraphrasing what Davidson (1980, p.231) famously asserted regarding the notion of rationality, the consistency requirements associated with decision utility "have no echo" in neuro-physiological theory.

CHAPTER SEVEN

HOW NEUROSCIENCE *COULD* REVOLUTIONIZE ECONOMICS

As we have seen in the previous chapters, the proponents of NE often manifest the ambition to implement revolutionary modifications in the economic theory of choice (see e.g. Camerer, Loewenstein and Prelec, 2005, p.10, and Rustichini, 2003). In this respect, it would be of little import to object that NEs have failed to accomplish revolutionary achievements. For the issue is whether - and if so, in what respects - they are likely to do so in a reasonably near future. Now, the plausibility of NEs' claims depends on how the notion of revolution is interpreted. In everyday language, the term "revolution" is employed in a number of different senses. What exactly do NEs mean when they speak of revolutionizing the economic account of decision making? What conditions would a modification of economic theory have to satisfy to qualify as revolutionary? In this final chapter, I distinguish four senses of the term "revolution" and argue that NEs are unlikely to foster a revolution in economic theory in any of these senses.

Before proceeding, let me put forward three preliminary remarks regarding the contents of this chapter. Firstly, the list of senses of "revolution" I consider is not meant to be exhaustive. For instance, I shall not discuss here whether NE can prompt significant changes in economists' modelling practices (see e.g. sections 2.B and 2.D on some NE studies which manipulate agents' choice behaviour by means of pharmaceuticals and neurochemicals). Secondly, NEs are not equally unlikely to prompt revolutionary modifications in all the respects I consider below. For example, while NEs may promote

significant increases in the evidential base of economic theory (section 7.A), more severe challenges seem to hinder progress in neuro-psychological measurements of well-being (section 7.C). And thirdly, whether some scientific advance constitutes a genuine scientific revolution is often matter of retrospective judgement (Nickles, 2006). This, however, does not prevent us from providing a principled evaluation of what contributions NEs are likely to offer to NE's parent disciplines.

7.A EVIDENTIAL BASE

A first sense in which the term "revolution" has been used by NEs relates to a significant expansion in the *evidential base* of traditional economic theory. NEs frequently state that a better understanding of the neural substrates of choice behaviour extends the scope of standard economic theory by enabling economists to observe variables that are considered "inherently unobservable" in it (Camerer, 2008a, p.45; see also Bernheim, 2009, p.9 and Caplin and Dean, 2008, p.665). Indeed, some go as far as to characterize traditional economic models of choice as a "limiting case" of NE models containing behavioural, psychological and neural variables (Camerer, Loewenstein and Prelec, 2004, p.564). The idea is that NEs enable other economists to measure not just a greater number of economic variables (*horizontal* expansion in evidential base), but also novel kinds of evidence (*vertical* expansion in evidential base).

Three progressively less restrictive views of the evidential base of economic theory can be contrasted. On a narrow interpretation, only observable choices constitute relevant evidence for the economic account of decision making. This view has been advocated by various leading economists (e.g. Samuelson, 1938 and 1947), but is currently defended only by few authors (e.g. Gul and Pesendorfer, 2008). A more ecumenical perspective takes both observed choices and psychological data (e.g. people's satisfaction reports) to belong to the evidential base of economic theory. This view prompted several behavioural economists (see e.g. Simon, 1955, Kahneman and Tversky, 1979) to combine psychological and behavioural data in constructing their models. According to a third, even more inclusive conception, also neural variables and

findings are to be used in building economic models of choice. In this perspective, NEs' calls to incorporate neuro-physiological data can be seen as the most recent wave of an iterative expansion of the evidential base of economic theory that arguably started with the early works of psycho-physiologists (see e.g. Glimcher, 2010, ch.4). As Camerer puts it, "advances in neuroscience make it possible to measure and causally manipulate many processes and quantities that were not imaginable […] when the foundation of neoclassical economics was being laid […] To ignore these developments entirely is bad scientific economizing" (2008a, p.60-1).

Now, let us assess the cogency of NEs' claims. As we have seen in *chapters two* and *four*, neuro-psychological insights can help economists to predict and explain a wider range of choices (e.g. think of the violations of expected utility theory). In particular, several NEs aim to develop a framework which is more general than standard economic theory in the sense of covering both rational and irrational forms of behaviour (Vromen, 2007, p.159). Yet, the mere fact that NE contributions allow for an increase in the range of phenomena modelled by economists does not imply that they will also foster a major expansion in the evidential base of economic theory. To be sure, it would be implausible to deny that NEs can measure and causally manipulate many variables besides those figuring in standard economic theory. Yet, the reason why economists focus on observed choices is not because they deem neural variables to be inherently unobservable, but rather because the very act of modelling requires them to concentrate on a subset of the features of the examined target systems. In this respect, the question remains as to how far the evidential base of the economic theory of choice should be expanded by economists. Let me explicate this concern.

As we have seen in *chapter four*, NEs have hitherto failed to demonstrate that economists will find it convenient to import many neural insights into their models. In particular, various NEs overstate the extent to which their studies expand the evidential base of economic theory. By way of illustration, consider the claim by Fehr and Camerer that while economists "treat preferences and beliefs as impossible or difficult to observe directly", NE research "rejects the premise of unobservability (2007, p.419). Such an assertion does not appear to withstand scrutiny. For the abstract character of preferences and beliefs makes them an unsuitable target for direct neuro-psychological observation. Analogous remarks apply to those NEs who urge other economists to include neural constructs and findings into their models on the alleged ground that "the study of the brain and nervous system is beginning to allow direct measurement of thoughts and feelings" (Camerer, Loewenstein and Prelec, 2005, p.10).

More generally, the point remains that NEs rarely put forward compelling reasons to integrate other disciplines' insights into economic theory. To see this, let us examine some NEs' contentions concerning the purportedly biological character of economics (e.g. Glimcher, Dorris and Bayer, 2005, p.254) and the alleged need to include biological methods and findings into economic theory (Zak and Denzau, 2001, p.32). *Prima facie*, it might seem that economics, defined as the study of relationships between ends and scarce resources having alternative uses (Robbins, 1935, p.15), has significant conceptual affinities with biology (see e.g. Marshall, 1890 [1961], p.772). Still, it is one thing to maintain that biological insights could help economists to improve their models of choice. It is quite another thing to assert that "ultimately, economics is a biological

science" (Glimcher, Dorris and Bayer, 2005, p.254). This latter assertion appears to face the following dilemma. On the one hand, interpreting the term "biological" broadly - so as to imply that economic models represent the behaviour of living organisms of some sort - would render the claim trivial. On the other hand, opting for a narrower interpretation of such a term makes it unclear why exactly economics would constitute a biological - as opposed to psychological, neuroscientific, etc. - discipline[78].

Zak and Denzau's (2001, p.32) assertion that the "methods and findings in the biological sciences need to be incorporated directly into economics" is even more disputable. To be sure, biological insights may well inspire and inform economists' models of decision making (see e.g. Fumagalli, 2011, on some parallel debates about modelling trade-offs in biology and economics). Yet, this falls short of implying that the advancement of economics is *conditional* upon the *direct* integration of biological methods and findings. Furthermore, one wonders what sort of biological "methods" and "findings" economists would have to include into their models. After all, no sensible economist would find it necessary to incorporate hyper-complicate sets of genetic and phenotypic traits' frequencies into her models of choice. Furthermore, it is hard to think of biological methods and findings whose incorporation is necessary for the advancement of economic theory. That is to say, economists are advised to resist the call of these biologically informed sirens, if they are to make the most of their modelling odysseys.

[78] It is interesting to note that various authors (e.g. Ghiselin, 1978, p.233) characterize biology as an economic discipline. Indeed, some go as far as to present their biological investigations as economic studies concerning "unusual sets of entities maximizing a rather unusual utility function" (Tullock, 1979, p.2).

At this point, one might wonder on what basis the proponents of NE put forward so disputable assertions. At first, it might be tempting to ascribe exaggerations and mistakes to the rhetoric of the discipline and to some authors' desire for visibility and funding. Upon inspection, however, NEs' overstatements can also be seen as symptoms of a deeper methodological *malaise*, which can be given a more philosophically informative diagnosis. As I argued elsewhere (Fumagalli, 2010), several NEs appear to presuppose that neuro-anatomical and neuro-physiological data constitute more *explanatorily basic* evidence concerning choice behaviour than the variables considered by economists. In the words of Camerer, Loewenstein and Prelec (2005, p.27), "the traditional economic account of behavior, which assumes that humans act so as to maximally satisfy their preferences, starts in the middle […] of the neuroscience account".

As we noted in section 4.A, this assertion is vulnerable to several objections. In particular, there are reasons to doubt many NEs' presupposition that insights from lower-level disciplines such as neurobiology and cognitive neuroscience are *ipso facto* more explanatorily informative for economists than observed choices. Moreover, insights from disciplines other than neurobiology and cognitive neuroscience (e.g. think of genetics) may well be more explanatorily informative than NE findings under the criteria of explanatory relevance that are implicitly presupposed by some leading NEs. For, as Camerer, Loewenstein and Prelec (2005, p.27) concede, "even the neuroscience account begins […] in the middle" of other disciplines' investigations.

7.B INTERTHEORETIC REDUCTION

A second respect in which NE has been said to revolutionize the economic theory of choice relates to *intertheoretic reduction*. In the NE literature, several authors advocate the adoption of a reductionist approach to the analysis of decision making and urge economists, psychologists and neuroscientists to engage in a common reductive unification of the decision sciences (Glimcher and Rustichini, 2004, p.448; see also Glimcher, 2010, xv). The NEs' case in favour of a reductive approach can be seen as the combination of two argumentative steps. The *pars destruens* goes as follows. The economic account of decision making heavily relies on folk psychology constructs such as beliefs and desires. Folk psychology, however, constitutes a conceptually primitive and predictively flawed theoretical framework that will be eventually reduced or eliminated by neuroscience (Churchland, 1981, 1986 and 1988; see also Feyerabend, 1963, and Rorty, 1965)[79]. Hence, to the extent that economists build their models on folk psychology foundations, their explanatory and predictive efforts are doomed to fail. As Rosenberg puts it, "beliefs and desires […] do not describe natural kinds. They do not divide nature at the joints. They do not label types of discrete states that share the same manageably small set of causes and effects and so cannot be brought together in causal generalizations" (1992, p.235).

The *pars construens* of the NEs' reductive case aims to show that NE research provides a neural microfoundation of socio-economic behaviour (Fehr and Camerer, 2007,

[79] According to eliminative materialists, "our commonsense conception of psychological phenomena [is] so fundamentally defective that both the principles and the ontology of that theory will eventually be displaced, rather than smoothly reduced, by completed neuroscience" (Churchland, 1981, p.67; see McCauley, 1996, p.437-440, on how the Churchlands came to qualify their eliminativist position).

p.419), thereby fostering the replacement of the theoretical constructs employed by economists. The proponents of NE often manifest the ambition to "replace the mathematical ideas used in economics with more neurally detailed descriptions" (Camerer, 2005) and "substitute familiar distinctions between categories of economic behaviour" with neural ones (Camerer, Loewenstein and Prelec, 2005, p.15). The idea is that although "conventional economic language can indeed approximate a lot of neural phenomena […] at some point it is more efficient to simply adopt constructs as they are defined and understood in other fields" (Camerer, 2008a, p.45). Now, what sort of conceptual modifications may be promoted by NEs? As we noted in section 4.B, one possibility is that NEs prompt economists to employ altogether different constructs by showing that what was thought to be a unitary phenomenon (e.g. hyperbolic discounting) is more plausibly regarded as several distinct phenomena brought about by dissimilar mechanisms (Craver and Alexandrova, 2008, p.402).

As noted by various authors (e.g. Bickle, 1996, p.64, and Churchland and Churchland, 1996, p.424), historical cases of reduction line up on a spectrum, ranging from retentive cases where the vocabulary and the ontology of the reduced theory are at least partially preserved (e.g. think of the reduction of Kepler's laws of planetary motion to Newton's laws of motion) to eliminative cases where the reduced theory is displaced (e.g. think of the phlogiston theory of combustion). Now, how profoundly NE will reshape the economic theory of choice is an open empirical question to be settled by actual research rather than armchair speculations. This, however, does not preclude us from providing methodologically informed reasons to doubt that the import of neural insights should be massive (see Fumagalli, 2011). To be sure, one might think that NE accounts of

decision making will reduce or even supplant higher-level ones, leading economists to relinquish their beloved utility functions in favour of hyper-complicated vectors of neural areas' activation patterns. Still, there are various reasons to resist NEs' reductive or eliminativist calls. Before examining these reasons, let us consider some models of intertheoretic reduction proposed by philosophers of science.

According to the so-called 'unity view' of science, scientific disciplines are structured as a layered edifice of levels connected via intertheoretic reductive relations (McCauley, 1996, p.432-3; see also Wimsatt, 1976). In *The Structure of Science*, Nagel characterizes reduction as "the explanation of a theory or a set of experimental laws established in one area of inquiry, by a theory usually […] formulated for some other domain" (1961, p.338; see also Oppenheim and Putnam, 1958). Nagel distinguishes two types of reductions, namely homogenous and heterogeneous reduction. Homogenous reductions occur when all the descriptive terms of the reduced theory are contained in the reducing theory and have approximately the same meanings in the two theories. Heterogeneous reductions, instead, take place when the reduced theory contains terms that do not appear in the reducing theory or when the meaning of central terms varies across the two theories[80].

Nagel's view has been the subject of several criticisms. For instance, some authors (e.g. Sklar, 1967) doubted that genuine homogeneous reductions occured in the history of science on the ground that what can be derived from the reducing theory are just approximations to the laws of the reduced theory. Others (e.g. Feyerabend, 1962 and

[80] Implementing heterogeneous reductions requires one to employ so-called bridge laws which connect the terms of the higher-level theory to those of the lower-level one. For a debate over the status of bridge laws, see e.g. Sklar, 1967, Fodor, 1974, and Schaffner, 1976.

1987, and Kuhn, 1962) challenged the purported meaning equivalence of terms in the reducing and the reduced theory by arguing that the meaning of a term crucially depends on the role it plays in a theory (see Papineau, 1996, for a critical discussion). In response to criticisms, some authors amended the Nagelian model by relaxing the requirement of exact nomological derivability between the reduced and the reducing theory (e.g. Schaffner, 1976 and 1977). Others developed alternative models of intertheoretic reduction which more radically depart from the Nagelian model. For instance, on the so-called 'new wave' reductionist account the reductive derivation does not target the to-be-reduced theory directly, but concerns an analog structure framed in the vocabulary of the reducing theory (Hooker, 1981, and Bickle, 1996 and 1998; see Endicott, 1998 and 2001, for some criticisms).

At present, no consensus has been reached among NEs as to *what* conception of intertheoretic relations most accurately reflects scientific practice and theorising across distinct behavioural sciences. Indeed, one wonders whether *any single* picture faithfully accounts for the complexity of those interrelations. In this respect, various authors maintain that the very idea of reducing the behavioural sciences to a single theoretical framework imposes an exceedingly restrictive constrain on interdisciplinary progress (Sunder, 2006, p.323)[81]. As we noted in section 6.C, several anti-reductionist lines of argument have been developed in the philosophy of cognitive sciences. In what follows, I provide three additional reasons to question NEs' reductionist claims.

[81] Whether scientific theories are reducible to one fundamental theoretical framework is a conceptually distinct issue than the one concerning what kind of regulative role, if any, reductionism plays in scientific theorising. According to some authors (e.g. Fodor, 1974, p.128-9), reducibility to physics is often taken to constrain the acceptability of theories in the special sciences (see also Horgan, 1993). Others, instead, deny that reductionism imposes restrictions on theory choice in the special sciences (e.g. Bickle, 1996, Churchland, 1986, ch.9, and Hooker, 1981). I do not expand on this debate for the purpose of my enquiry.

A first line of argument points to *NEs' divergences* concerning which disciplines are supposed to provide the fundamental constructs for their still-to-come reductive framework. My reasoning goes as follows. Many NEs take their findings to yield a neural microfoundation to the economic theory of choice. Yet, as we have seen in section 5.C, it remains an open question "where the appropriate microfoundations have been reached" (Mäki, 2010, p.109). Moreover, many NEs' reductionist claims rest on highly speculative presuppositions concerning the relationship between economics and other behavioural sciences. By way of illustration, consider Camerer's assertion (1999, p.10575) that "because economics is the science of how resources are allocated by individuals […] the psychology of individual behaviour should underlie and inform economics, much as physics informs chemistry". In his paper, Camerer does not cogently substantiate this parallel. Moreover, his simplistic comparison would hardly impress anyone having some knowledge of the complexity of intertheoretic - not to say interdisciplinary - reductive relations.

My second argumentative strategy relates to the existence of *multiple levels of description* of human choice behaviour. As we have seen in sections 2.C and 4.A, several NEs seem to think that accounting for choice behaviour in terms of lower-level mechanisms and processes constitutes an obvious explanatory advancement. Yet, as noted by Kuorikoski and Ylikoski, higher-level explanations "do not [always] inherit their explanatory qualities from lower-level descriptions" and "an explanation is not necessarily improved when the explanans is itself explained" (2010, p.220-2; see also Ylikoski and Kuorikoski, 2010). Moreover, in the absence of principled reasons for

thinking that choice behaviour is most conveniently characterized at the neural - rather than some lower - level of description, one could employ reductionist considerations to advocate the inclusion of - not so much neural, but rather - genetic, micro-physical, etc. insights into economic models. In this respect, advocating the adoption of a reductionist approach may even backfire against the proponents of NE. For many NE experiments are grounded in folk psychology's vocabulary and constructs (Quartz, 2008, p.468).

A third anti-reductionist line of argument relates to the *depth* of NEs' intended reduction. My reasoning can be explicated as follows. As illustrated by the history of science, fruitful coevolutionary interactions have taken place between distinct behavioural disciplines (see e.g. Craver, 2007, Ch.7, and Richardson, 2009, on psychology and neuroscience; see also Darden and Maull, 1977, on interfield relations in the biological sciences)[82]. However, some NEs advocate - not just some coevolutionary interaction between NE's parent disciplines, but also - the elimination of various economic constructs in favour of neuro-psychological ones (see e.g. Camerer, 2008a, p.45, and Camerer, Loewenstein and Prelec, 2005, p.15).

Now, intertheoretic transitions occasionally involve the elimination of particular theoretical entities (e.g. think of the crystal spheres of Ptolemaic astronomy and the luminous ether of pre-Einsteinian mechanics). In light of this historical record, some authors assert that our introspective certainty concerning the existence of beliefs and desires will turn out to be "as badly misplaced as was the classical man's visual certainty that they star-flecked sphere of the heavens turns daily" (Churchland, 1981,

[82] See also Kincaid (1997, p.6) for a case in favour of non-reductive unifications according to which "scientific unity comes from integrating the special sciences with their lower-level counterparts […] using one to develop explanatory constraints for the other".

p.70). Still, it is doubtful that progress in psychology and neuroscience is more plausibly reconstructed in eliminativist rather than coevolutionary terms (see Bickle, 2003, Churchland and Churchland, 1996 and 1998, McCauley, 1986 and 1996, and Mundale and Bechtel, 1996, for a debate). More generally, one may concede that biological, psychological and economic entities are composed of lower-level micro-physical entities, and yet deny that "biology, psychology, and economics are […] reducible as explanatory theories to their lower-level counterparts" (Kincaid, 1997, p.6)[83].

Now, the proponents of NE rarely put forward convincing reasons for substituting economists' constructs. By way of illustration, consider the remarks that some NEs formulated regarding the concepts of risk preferences, time preferences and social preferences. Most economists treat these as distinct types of preference. According to Fehr and Camerer, however, it is "important" for economists to ascertain whether these preferences have the same neural substrates (2007, p.426). Now, let us suppose that significant overlap was found between the neural substrates of risk preferences, time preferences and social preferences. Even this would fall short of implying that economists should stop regarding these types of preferences as distinct. For the reasons why economists differentiate between those preferences relate - not so much to the alleged dissimilarity of their neural underpinnings, but rather - to the fact that they concern distinct economic phenomena.

[83] See also Ross (2010, p.661) for a critique of the claim that molar scale phenomena (e.g. people's observed choices) are in principle fully explicable by reference to molecular phenomena (e.g. neural activation patterns).

To recapitulate, several NEs argue that the NE framework for modelling human choice behaviour will be forged by an iterative reduction of economics to psychology and then to neuroscience (Glimcher, 2010, xv). Indeed, some appear to presuppose that constructing neuro-psychological models of choice would *per se* constitute "a reductive unification of the decision sciences" (Glimcher and Rustichini, 2004, p.448). Even so, there are various reasons to doubt that NEs will foster a reduction of the economic account of decision making in neuro-psychological terms. In particular, NEs have hitherto failed to specify how exactly neuro-psychological evidence and insights are supposed to prompt the reduction - not to say the elimination - of higher-level economic constructs.

7.C WELFARE ANALYSES

A third respect in which NE has been said to trigger a revolution relates to economic *welfare analyses* and *policy evaluations*. As we have seen in the previous chapter, a number of concerns arise regarding some NEs' proposal to replace decision utility with neuro-psychological constructs such as experienced and neural utility. Now, let us suppose - for the sake of argument - that we could measure experienced utility and neural utility in reliable and accurate terms. Even so, the question would remain as to whether - and if so, why - decision theoretic analyses should be based on these notions. Various proponents of NE advocate the adoption of experienced utility and neural utility on the alleged ground that these constitute a better approximation to agents' *well-being* than decision utility. The idea is that maximizing experienced utility or neural utility is objectively better for agents than satisfying their actual or informed preferences. In the words of Loewenstein and Haisley, recent NE research contributes to "the measurement of welfare and the design of economic and social systems that maximize welfare" (2008, p.44). Let us assess the cogency of this line of argument.

Suppose we are facing a situation where there is some fact of the matter as to what an agent's well-being is. One might allege that well-being is somehow *reducible* to some measure of experienced utility or neural utility, i.e. that well-being ultimately consists of specific hedonic experiences or some set of neural activation patterns. This rebuttal, however, does not appear to withstand scrutiny. After all, well-being depends not just on hedonic experiences or their neuro-physiological correlates, but also on "other aspects of life, such as autonomy, freedom, achievement" (Kahneman and Sugden,

2005, p.176; see also Nozick, 1974, p.42-45). In this respect, one may insist that people's hedonic experiences and their neuro-physiological counterparts are important constituents of well-being (Kahneman, 2000, p.19). Even so, we do not have to embrace a radical form of eudaimonism to deny that experiences are valuable solely in virtue of their liability to generate pleasurable hedonic states or particular neural activations. As Bernheim vividly puts it, "we often consider ourselves better off when we have actual autonomy, liberty, and a firm grasp on reality even if, as a consequence, we must relinquish appealing illusions and experience less pleasurable neurobiological sensations" (2009, p.30).

The proponents of true utility may protest that one's intuitions concerning this issue presumably depend on what conception of well-being she endorses. In particular, they might contend that some measure of experienced utility or neural utility *approximates* the well-being of the agent more accurately than decision utility. After all - the thought would be - there are various reasons why an agent may fail to choose options which promote her objective well-being (see e.g. Elster, 1983, and Sen, 1987, on preferences based on mistaken beliefs)[84]. Nonetheless, also actions yielding high experienced utility or neural utility may fall short of maximizing well-being. To give one example, think of drug addicts and of pathological gamblers. A person can experience very pleasant sensations if her dopamine neurons suppress competing serotonergic circuits. Regrettably, this process often results in addictive forms of behaviour which have a disastrous impact on people's well-being (Ross, 2008, p.132). That is to say, on many

[84] Similar remarks apply to the proposal to regard only choices based on informed preferences as indicators of well-being, for many suboptimal decisions result from - not so much inaccurate information, but rather - people's computational limitations and self-control problems (see e.g. Bernheim and Rangel, 2008).

conceptions of what it is good for someone to do, the best actions are not the ones that maximize experienced utility or neural utility. As Read puts it, “a separate value judgment is necessary before Benthamite utility can be identified with the good” (2007, p.58).

In such a context, the further question arises as to why economists should commit themselves to some particular conception of well-being. After all, traditional decision theory remains agnostic as to whether agents’ objective well-being consists in experiencing pleasure and avoiding pain, satisfying their actual preferences, etc. Now, let us assume - for the sake of argument - that decision theory did rest on the assumption that well-being consists in maximizing the value of some specific measure of utility. One may wonder which notion of utility should be regarded by economists as normatively fundamental, i.e. why agents should choose actions that maximize the value of that - rather than some other - measure. By way of illustration, suppose that agents’ objective well-being consisted in maximizing the value of some measure of experienced utility. Even so, several questions would remain such as: *qua* rational agent, should you maximize the value of instant, anticipated or remembered utility? Over what time intervals? Or maybe should you compute a weighted average of distinct kinds of experienced utility, each being measured over various time spans? If so, how should the various elements of such a function be weighed? Providing principled grounds to specify the maximandum of this optimization exercise appears to be disconcertingly difficult.

At this point, the proponents of experienced utility may allege that their analyses apply to "situations where a separate value judgment designates experienced utility as a relevant criterion for evaluating outcomes" (Kahneman, Wakker and Sarin, 1997, p.377). Yet, the issue at stake is precisely when this happens to be the case, and by means of what criterion we are supposed to identify these situations. To be sure, one might speculate that well-informed agents would often deem maximizing experienced or neural utility to be in their own best interest. Still, this conjecture leaves us in the dark in cases where different people have contrasting intuitions regarding what constitutes or promotes their well-being. In this respect, dissimilar views have been advocated as to whether experienced and neural utility provide a more suitable basis for policy evaluations than decision utility. For instance, Kahneman, Wakker and Sarin (1997, p.389) allege that consumer sovereignty is called into question when observed decisions fall short of maximizing experienced utility. For his part, Sugden (2004 and 2008) questions the normative appeal of experienced utility as a criterion for policy evaluation (see also Sugden, 2006, p.217, for the claim that an agent may attribute a high importance to the opportunity of satisfying her own preferences irrespective of whether these preferences are stable under experience and reflection).

I am not concerned here with evaluating these assertions. For the purpose of this enquiry, it suffices to note that the previous calls to ground economists' analyses on experienced utility or neural utility measures involve presuppositions which transcend the scope of traditional decision theory and the evidential reach of neuro-psychological investigations (e.g. who is entitled to establish what constitutes agents' well-being? By means of what criteria we are to adjudicate disagreements regarding this issue?). That is

to say, advances in neuro-psychological research may promote the development of indicators of well-being which complement traditional economic welfare analyses. At the same time, it remains an open question whether our growing ability to measure neural activation patterns will enable us to "objectively compare mental state between individuals" and implement "direct inter-individual comparisons of welfare" (Glimcher, 2010, p.425).

7.D INTERPRETATION OF MODELS

Another respect in which NE is claimed to prompt revolutionary modifications relates to the *interpretation* of economic *models* of choice. As we have seen in section 2.A, the proponents of NE often criticize standard economic theory for failing to provide a descriptively accurate representation of the neural substrates of choice behaviour and for positing agents having implausible cognitive and computational abilities. Moreover, they advocate the construction of mechanistic models of choice and urge economists to employ neuro-psychological findings in building their models. The reasoning of NEs can be explicated as follows.

Standard economic theory treats the human brain as a black box and does not make specific assumptions regarding the neuro-psychological substrates of agents' decisions. If economic theory provided accurate predictions across choice settings, then identifying the neuro-psychological underpinnings of decisions could arguably be unnecessary for the economists' purposes. Yet, economic theory faces widespread predictive failures. Moreover, identifying what neuro-psychological processes underlie decisions enables economists to improve the explanatory and predictive performance of their models. Hence, economists should make use of neuro-psychological findings in constructing their models of choice.

As we have seen in *chapter four*, it is doubtful that the trade-offs between the modelling benefits and the modelling costs associated with a neural enrichment of economic theory make it convenient for economists to include many neural insights into their

models. For the purpose of this section, let us contrast the interpretations that NEs and other economists respectively give to their models of choice. On the one hand, economists usually remain agnostic regarding the actual number and the features of the neuro-psychological processes underlying people's decisions. In particular, they do not take their models to provide a descriptively accurate representation of these processes. On the other hand, many NEs give to their models a realistic interpretation, according to which there *really is* such and such neuro-psychological process at work when the subjects make a particular choice (Glimcher, 2010, p.126 and 133). The idea is that: (i) the neuro-psychological processes postulated by NE models exist; (ii) they possess the features that NE models ascribe to them; and (iii) these features are characterized by NE models in fairly accurate terms[85].

In what follows, I shall use expressions such as "realistic interpretation of models" and "realistic representation" in the sense defined by these three conditions unless specified otherwise. Those conditions relate to three different respects in which the realisticness of scientific representations can be evaluated. A similar categorization was proposed by Mäki (1992, p.329) with regard to the realisticness of idealizing assumptions in economic theory. According to Mäki's categorization, an assumption is: referentially realistic when it can be taken to refer to a non-fictitious target system; representationally realistic when it represents a feature that is actually possessed by the modelled target system; and veristically realistic when it characterizes those features in accurate terms. Adopting Mäki's terminology, many NE modellers appear to regard their models as

[85] The calls in favour of a realistic interpretation of neuro-psychological models are often framed in a mechanistic vocabulary. However, this interpretation may be advocated on independent, non-mechanistic grounds. Moreover, one may adopt a mechanistic approach without thereby being committed to give a realistic interpretation to her models (see e.g. Matthewson and Calcott, 2011, p.737; see also Craver and Alexandrova, 2008).

referentially, representationally and veristically realistic representations of the neuro-psychological substrates of choice behaviour.

Indeed, some proponents of NE do not rest content with advocating such a realistic interpretation of their own models, but also urge economists to relinquish their as if representations of choice behaviour in favour of neurally informed models. For example, Camerer alleges that NE "replaces the […] fiction of a utility-maximising individual which has a single goal, with a more detailed account of how components of the [human neural architecture] interact and communicate to determine individual behaviour" (2007, C28; see also Camerer, 1998, p.177). Similarly, Glimcher (2010, p.142) invites economists to "take [their] powerful mathematical models at face value and begin to ask whether the hidden elements that they propose actually exist", and Camerer, Loewenstein and Prelec contend that neuroscientific evidence "suggests specific functional forms to replace 'as if' assumptions that have never been well supported empirically" (2005, p.10).

In recent years, a growing body of empirical findings have been taken to show that human choice behaviour results from the interplay of multiple neuro-psychological processes (see e.g. Loewenstein et al., 2008, p.647, and McCabe, 2008, p.355). Regrettably, NEs seem to hold heterogeneous positions regarding what neuro-psychological processes underlie people's decisions. Below I identify some of these divergences and assess what implications they can be taken to have for the interpretation of NE models of choice. In particular, I argue that the availability of multiple NE models positing dissimilar neuro-psychological processes does not

undermine the merits of these models, yet calls into question the realistic interpretation many NEs give to those models.

NEs postulate a number of distinct neuro-psychological processes in their models of decision making. For instance, McCabe (2008, p.355) argues that many decisions result from the interplay of two neural circuits, namely a stimulus-response system which encodes correlations between actions and rewards, and a goal directed system which enables the agent to evaluate options in terms of anticipated motivational states. For their part, Bernheim and Rangel (2004) model decisions as the outcome of the interactions of brain processes operating in either 'cold' or 'hot' mode. Still differently, Fudenberg and Levine (2006) represent agents' choices as the result of the interplays of a long-run and a short-run self, and Loewenstein and O'Donoghue (2004) propose a model where deliberative and affective systems jointly underlie choice behaviour.

The proponents of NE hold heterogeneous positions regarding not just *what* neuro-psychological processes underlie human choice behaviour, but also the *role* allegedly played by specific processes. To give one example, consider the claim by Camerer, Loewenstein and Prelec (2005, p.11) that human decisions result from the interactions occurring both between controlled and automatic processes and between cognitive and affective systems[86]. The affective system has been respectively claimed to help (e.g.

[86] *Cognitive* and *affective* processes have been differentiated both in psychology (e.g. Zajonc, 1980 and 1984, and Zajonc and McIntosh, 1992) and neuroscience (e.g. LeDoux, 1996, and Panksepp, 1998). As to *controlled* and *automatic* processes, the main distinction can be explicated as follows (Schneider and Shiffrin, 1977, and Shiffrin and Schneider, 1977). Controlled processes typically involve serial computations and activate when a person encounters novel decision problems or unexpected events (Hastie, 1984, and Libet, 1985). Automatic processes, instead, frequently operate in parallel, are rarely accessible to agents'

Damasio, 1994), constrain (e.g. Loewenstein, 1996) and prevent (e.g. Baumeister, 2003) the cognitive system from making optimal choices (see also Benhabib and Bisin, 2005, and Brocas and Carrillo, 2008, for different positions regarding how often controlled processes interfere with automatic ones).

Indeed, NEs do not even concur on the issue of how many stages allegedly comprise human decision making. According to some (e.g. Glimcher, Dorris and Bayer, 2005, p.246; see also Glimcher, 2009), decision making is most appropriately regarded as a two-stage process whereby expected utilities are computed and then compared for all the available actions. Others, instead, argue that utility maximization can be plausibly characterized as a one-stage process (e.g. Vromen, 2010a, p.24).

In such a context, the question arises as to what *implications* the aforementioned divergences have for the interpretation of NE models. The existence of multiple models representing a given target system does not *per se* cast doubt on the merits of those models. After all, distinct modellers often study specific decision problems at different levels of detail and for dissimilar purposes. Moreover, scientists frequently employ multiple models which ascribe different properties and features to the investigated target systems (see e.g. Morrison, 2000, and Weisberg, 2007a). Indeed, modellers occasionally acquire valuable predictive and explanatory insights by combining models which make inconsistent assumptions about the phenomena of interest (see e.g. Woodward, 2006)[87]. Even so, the availability of multiple NE models of choice which posit dissimilar neuro-

consciousness and prompt most of our daily, repetitive behaviour (Bargh and Chartrand, 1999, and Baumeister and Sommer, 1997).

[87] See also Friedman (1953) and Nagel (1963) on how descriptively inaccurate models can be predictive and explanatory (see Wimsatt, 1987, and Odenbaugh, 2005, for analogous remarks in the literature on modelling in biology).

psychological processes does not fit well with the realistic interpretation many NEs give to those models. In particular, the diversity of the neuro-psychological processes postulated in NE models - coupled with the evidential concerns arising in relation to NEs' observational tools (see *chapter three*) - makes it doubtful that these models provide descriptively accurate representations of the neural substrates of choice behaviour.

A proponent of NE may rebut that the aforementioned distinction exaggerates the differences between the interpretations that NEs and other economists respectively give to their models. In particular, she might allege that both NEs and other economists are concerned with the algorithmic level of description of choice behaviour (see Marr, 1982). Now, it is true that at an abstract level many economic and NE models alike represent choice processes as an algorithmic procedure linking environmental variables and decisions (Rustichini, 2009, p.48). Yet, even at this abstract level NEs and other economists respectively give to their models rather different interpretations. For instance, as noted by Vromen (2010b, p.174), NEs often take their models to "get the decision-making process right" at the algorithmic level of analysis. By contrast, standard economic theory does not rest on the presupposition that the functional forms posited by economic modellers provide a descriptively accurate representation of the algorithmic underpinnings of people's decisions[88].

[88] In the words of Rustichini (2009, p.49), the utility functions posited by standard decision theory are "a purely conceptual device, and testing whether the decision maker really selects his choice by maximizing [a specific functional form] is a misunderstanding of the method of economic theory".

In this respect, it is telling that various leading NEs emphasize the contrast between the interpretations that NEs and other economists respectively give to their models of choice. To see this, consider the distinction between a "soft theory" and a "hard theory" of choice behaviour proposed by Glimcher (2010). On the one hand, economists' "soft theory" abstracts from the mechanistic underpinnings of observed choices and does not take a position as to what neuro-psychological processes underlie people's decisions. On the other hand, NEs' "hard theory" posits that choosers who behave in accordance with the axioms of expected utility theory "do so because a group of neuronal firing rates in a valuation circuit encodes the cardinal subjective values (and/or expected subjective values)" of each available option (Glimcher, 2010, p.129 and 138).

The proponents of NE may object that other economists as well routinely model specific decision problems in dissimilar ways. To give one example, economists postulated a variety of sub-personal entities to account for observed choices (see e.g. Fudenberg and Levine, 2006, on far-sighted and short-sighted selves, and Benhabib and Bisin, 2005, on controlled and automatic processes). Even so, the availability of multiple models positing dissimilar entities and processes does not have equally problematic implications for NEs and other economists. For while NEs take their models to provide descriptively accurate characterizations of the neuro-psychological substrates of choice behaviour, economists remain agnostic as to whether the processes postulated by their models have precisely identifiable neuro-psychological counterparts and have the features hypothesized by those models. Hence, the fact that distinct economic models posit dissimilar entities and processes to account for observed decisions does not *per se* cast doubt on the merits of those models.

Regrettably, NEs frequently regard economic models as if they were intended to provide descriptively accurate characterizations of the neuro-psychological processes underlying people's decisions. This, in turn, leads them to formulate misplaced criticisms of standard economic models. By way of illustration, consider the contention of Loewenstein et al. (2008, p.647) that NE challenges "the standard economic assumption that decision making is a unitary process" and suggests instead that choice behaviour "is driven by the interaction between automatic and controlled processes". Such a claim apparently overlooks that economists, in assuming that a rational agent behaves as if she maximizes her expected utility, do not take a position as to the number of neuro-cognitive processes underlying her choices. Furthermore, the alleged fact that automatic and controlled processes interactively underlie decision making does not directly bear against the economists' conjecture that agents behave as if they were maximizing expected utility. For one can behave consistently with such an assumption even if decision making is not a "unitary process". Hence, pointing out that choice behaviour results from the interaction of heterogeneous neuro-psychological processes does not *per se* undermine standard economic models.

In such a context, a further worry arises regarding the evidential basis on which additional neuro-psychological processes are postulated in the literature. This concern can be explicated as follows. Neuro-imaging studies are often based on task comparisons that attempt to associate individual tasks with particular cognitive processes (Poldrack and Wagner, 2004). Regrettably, these studies often face severe problems of evidential underconstraint (see e.g. Van Orden et al., 2001). That is to say,

one can frequently account for the collected neural data by refining the characterization of formerly posited processes and by postulating additional processes. The problem is that almost any profile of neural activation patterns can be accommodated by means of these procedures. Hence, it is dubious that the models that are thereby constructed can be plausibly taken to provide a descriptively accurate representation of the neuro-psychological substrates of choice behaviour (see Fox et al., 1998 and 2005, and Henson, 2006, for similar remarks).

To be sure, convergent evidence from psychology and neuroscience can usefully constrain NEs' conjectures regarding the neuro-psychological substrates of choice behaviour. Still, the point remains that neuroscientific findings have been hitherto employed prevalently for confirmatory purposes. As Kuorikoski and Ylikoski (2010, p.226) put it, many authors regard finding differential activations across tasks as supporting the existence of distinct cognitive processes, yet "the failure to find such contrasts is not regarded as disconfirming such hypotheses". In this respect, one may well allege that "ontology is rarely if ever handed to us on a silver platter in any science" (Roskies, 2008, p.28) and that NEs will develop increasingly accurate characterizations of neuro-psychological processes. Even so, the existence of multiple NE models of choice which postulate dissimilar neuro-psychological processes casts doubt on the realistic interpretation that many NEs give to those models.

CONCLUSIONS

The recent advances at the interface between economics, psychology and neuroscience have encouraged various NEs to advocate substantial modifications in the economic theory of choice. As we have seen in the previous chapters, there are several reasons to doubt that NE research is going to foster revolutionary modifications in economic theory. Moreover, the proponents of NE have hitherto failed to provide compelling evidence or reasons in support of such an ambitious thesis. This, however, does not exclude that NE can promote significant progress in economic theory. In what follows, I evaluate the prospects of NE in light of some criteria of scientific progress, devoting particular attention to Lakatos' distinction between progressive and degenerating research programs. I then conclude by summarizing the main objections we raised against NEs' attempts to substitute economists' constructs and develop a unified interdisciplinary framework for modelling decision making.

As we noted in the *Introduction*, Kuhn (1962) provides an innovative and controversial account of scientific progress. In his view, psychological and sociological factors exert a pervasive influence on theory choice in science (1970b, p.6). To be sure, Kuhn does not exclude that scientists may formulate principled comparisons of the predictive and explanatory virtues of competing paradigms. Still, he takes rational discussion to play only a limited role in determining which paradigms emerge and denies the existence of shared superparadigmatic standards for assessing whether specific intertheoretic transitions constitute objective scientific progress.

Various authors follow Kuhn in doubting the existence of objective criteria for theory choice in science (see e.g. Barnes, 1974, and Bloor, 1984, on the so-called strong programme in the sociology of knowledge). Others, instead, criticize him for severely underestimating the extent to which rational debate can promote scientific progress (see e.g. Lakatos, 1971, p.104-5; see also Laudan, 1984, and McMullin, 1993). In the philosophy of science, various criteria for assessing scientific progress have been advocated. Let us consider the ones proposed by Popper and Lakatos in turn.

According to Popper, a novel theory constitutes progress in science when it has "new and testable consequences" and successfully predicts phenomena that have not been previously observed (1963, p.241-3; see also Popper, 1959). Similarly, Lakatos (1970) defines a series of theories as theoretically progressive if each successive theory has some excess empirical content over its predecessors, i.e. it predicts some previously unexpected fact. A theoretically progressive series of theories is also empirically progressive if some of its excess empirical content is corroborated. Finally, a series of theories is progressive if it is "consistently theoretically progressive" - that is each successive theory predicts some new fact - and at least "intermittently empirically progressive" - that is, "every now and then the increase in content should be […] retrospectively corroborated" (1970, p.134)[89].

[89] There are at least two respects in which Lakatos' criteria for assessing scientific progress are less demanding than the ones proposed by Popper. Firstly, Lakatos requires intermittent - rather than continuous - empirical success. And secondly, while Popper holds that only previously unknown facts count when it comes to assessing theories' progressiveness, Lakatos allows that a theory can be supported by previously known facts, provided that those facts were not employed in constructing the theory (Lakatos and Zahar, 1975). As Worrall puts it, "one can't use the same fact twice; once in the construction of a theory and then again in its support" (1978, p.48).

Now, let us focus on Lakatos' criteria for evaluating scientific progress. Over the last few decades, intense debates have taken place as to whether the development of economic theory can be plausibly reconstructed as the steady growth of a progressive research program (see e.g. Hutchison, 1978, and Hands, 1985a and 1985b). In particular, various authors have assessed specific episodes in the history of economic theory in light of Lakatos' criteria of scientific progress (see e.g. Blaug, 1975, on Walrasian and Keynesian economics, Latsis, 1976, on the theory of the firm, and Weintraub, 1979, on general equilibrium theory). A proponent of NE may draw on Lakatos' distinction between progressive and degenerating research programs and put forward the following challenge to other economists: "standard economic theory has obtained remarkable predictive and explanatory successes. Still, it also faces widespread explanatory shortcomings, and its predictions fail to cohere with the findings that have been accumulated in other behavioural sciences. Fortunately, NE has already achieved significant accomplishments and promises to foster significant progress in economic theory. Hence, economists should invest efforts and resources in NE research".

Prima facie, this reasoning provides NEs with a plausible basis to advocate the neural enrichment of economic theory. As I argue below, however, such reasoning is in need of two major qualifications. To begin with, it remains an open question whether NE research is as progressive as many NEs appear to think. And secondly, even if NE was shown to be progressive, this would not *per se* license the conclusion that NEs will implement revolutionary modifications in the economic account of decision making. Let me expand on these two issues in turn.

As we have seen in *chapter two*, NEs can improve economic models with regard to several desiderata, ranging from descriptive accuracy to predictive power and explanatory insightfulness. In recent years, some authors have begun to reduce the "conceptual gap" (Caplin and Dean, 2008, p.663; see also Caplin, 2008) between economics, psychology and neuroscience by integrating findings and modelling tools across these disciplines. In light of these contributions, it would be ungenerous to claim that most NEs "are in the dark" about how their research "will reshape economics" (Rubinstein, 2008, p.486-7). Still, as we noted in *chapters four* and *five*, NEs' integrative efforts have a limited scope and do not substantiate their speculations about the completion of a single, general theory of choice behaviour.

Our previous observations concerning Lakatos' criteria of scientific progress suggest an additional reason for questioning NEs' calls to implement a neuro-psychological enrichment of the economic theory of choice. This reason relates to the fact that economists rely on additional criteria of scientific progress besides those endorsed by NEs. By way of illustration, consider the transition from cardinal to ordinal utility theory. The latter has no additional empirical content with respect to the former. Still, many economists deem ordinal utility theory to constitute a major improvement in economic theorising, as it enables one to represent consistent choice behaviour without making any assumption regarding agents' psychological states.

What about the link between the alleged progressiveness of NE research and its revolutionary potential? Let us suppose - for the sake of argument - that NE constitutes a progressive research program in Lakatos' sense. Even this, by itself, falls short of

implying that NE is likely to foster a revolution in the economic theory of choice. Let me provide two remarks in support of this claim. My first remark points to the fact that many NEs' contributions, while being of great interest to neuroscientists, have a limited or indirect bearing on standard economic theory (see e.g. Rubinstein, 2008). In particular, few NEs have satisfactorily addressed the challenge that Bernheim (2009, p.27) formulates in relation to NE research: "Provide an example of a novel economic model derived originally from neuroeconomic research that improves our measurement of the causal relationship between a standard exogenous environmental condition […] and a standard economic choice". My second remark is that many advances in NE research build not so much on NEs' original contributions, but rather on previous results in behavioural economics. In the words of Craver and Alexandrova, much of the recent NE literature "focuses on behavioural, not neural, economics", with the neural component being frequently "limited to scanner evidence showing what areas of the brain light up when one performs some behavioural-economic task" (2008, p.383).

Regrettably, many NEs appear to overstate their own past and potential achievements. To render this point more vivid, let us compare the ambitious claims initially made by some leading NEs and what they assert in light of their attained results. Consider, for example, Camerer, Loewenstein and Prelec's manifesto *How Neuroscience can Inform Economics*. On the one hand, the authors prophesize that a "radical departure from current theory will become necessary, in the sense that the basic building blocks will not just consist of preferences, constrained optimization and […] equilibrium" (2005, p.54). On the other hand, when it comes to assessing how NEs have *already* informed economic theory, they put forward rather elusive claims such as: "Perhaps knowing

more about basic neural mechanisms […] can help explain these puzzles"; "there is no reason other models starting from a very different basis could not be constructed"; "it is hard to believe that some neuroscientific regularities will not help explain some extant anomalies" (2005, p.53-55).

As I noted elsewhere (Fumagalli, 2010), it is striking how *much more* moderate NEs have become just *a few* years after their initial announcements. As Camerer (2008a, p.44) has recently conceded, "these early neuroeconomics papers should be read as if they are speculative grant proposals which conjecture what might be learned from studies which take advantage of technological advances"[90]. Now, I am aware that one should not derive momentous implications from literally interpreting isolated statements, and that some exaggerations may be explained in light of the need to obtain public attention and funding. Even so, one expects NEs to advance much more measured claims in the future. For some authors' propensity to overstate their own achievements has generated a lot of unnecessary confusion in the literature, making many economists needlessly sceptical about the prospects of NE research. As Glimcher aptly acknowledges, NE "has rocketed into the public awareness at a rate completely out of proportion to its accomplishments" (2010, xii)[91].

[90] See also Spiegler (2008, p.520) and Jamison (2008, p.407), who allege that it would be a remarkable accomplishment if NEs succeed in inspiring new economic models. Irrespective of its plausibility, such a claim constitutes a significant downplay with respect to the promises initially put forward by many NEs.

[91] Analogous remarks apply to other NEs' claims. To give one example, consider the assertion by Vercoe and Zak (2010) that NE promotes a methodological reversal from deductive to inductive economics which prompts economists to base their models on empirical findings rather than axiomatic speculations. Such a claim grossly underestimates the relevance of many inductive models that had been proposed by behavioural economists since the Seventies (see e.g. Kahneman and Tversky, 1979, and Tversky and Kahneman, 1992, on various versions of prospect theory).

Having said that, let us briefly recapitulate the main objections we put forward in the preceding chapters against NEs' revolutionary ambitions. As I argued in *chapter three*, several concerns arise regarding the accuracy, reliability and robustness of many NE findings. Moreover, the inferences NEs make in their investigations do not always warrant confidence in the subsequently reported results. Some of these evidential and epistemological concerns will be resolved thanks to advances in NEs' scanner technology and experimental practices. Others, instead, are likely to persist in spite of these progresses. Moreover, as illustrated in *chapter four*, NEs overestimate the extent to which neurally informed contributions enable economists to satisfy their modelling desiderata. In particular, the trade-offs between the desiderata that NEs and other economists respectively value severely constrain the incorporation of neural insights into economic models.

These remarks appear to be all the more significant in light of the profound differences in the evidential bases, the theoretical constructs and the explanatory aims associated with NE's parent disciplines. As we have seen in *chapter five*, these differences cast doubt on the relevance of many NE studies for the economic theory of choice (see also *chapter six* for a case study on the concept of utility). NEs' attempts to develop a unified interdisciplinary account of decision making are further constrained by the divergences occurred among NEs themselves regarding how NE is to be conceptualized, how it should be expected to inform economics, and what disciplines will provide the fundamental constructs for the NE theoretical framework. As to NEs' revolutionary claims, their cogency rests on whether NE research brings an innovative perspective on questions that have been intractable for, or beyond the reach of, other economists

(Smith, 2007, p.313; see also Camerer, Loewenstein and Prelec, 2005, sec.4, and Gul and Pesendorfer, 2008). As I argued in *chapter seven*, NEs have failed to show that their studies are likely to prompt revolutionary changes in the economic theory of choice.

Over the last few years, NEs' attempts to combine insights from economics, psychology and neuroscience have prompted intense debates among the practitioners and the philosophers of these sciences concerning issues such as the relevance of different disciplines' findings for economic theory and how to solve the trade-offs between specific modelling desiderata (see e.g. *Economics and Philosophy*, 2008, Vol.24, no.3; the *Journal of Economic Methodology*, 2010, Vol.17, no.2; and *Biology and Philosophy*, 2011, Vol.26, no.5). In such a context, the further question arises as to how NE should organize itself to maximize its potential for success. According to some authors (Craver and Alexandrova, 2008, p.384), NEs will find it more fruitful to pursue integrative - rather than revolutionary - research projects involving NE's parent disciplines. The idea is that combining findings and modelling tools from these disciplines enables economists to build more predictive and explanatory models. As I argued in this enquiry, the greatest promise for the advancement of NE lies in adopting this pluralistic approach to the modelling of decision making, with economists, psychologists and neuroscientists pursuing integrated - yet not unified - modelling approaches.

In commenting over the potential for success in NE research, several authors maintain that it is premature to judge NEs' achievements (Quartz, 2008, p.466, and Smith, 2007, p.313), that "the ultimate proof is in the pudding" (Bernheim, 2009, p.38; see also

Schotter, 2008, p.77), and the like. For their part, Craver and Alexandrova (2009, p.382, italics mine) contend that any bets on NE's long-term prospects are "not so much *premature* as *ill defined*". In their view, "the field is too young for definition let alone wagering", as we currently lack answers "to even the most fundamental questions for defining neuroeconomics". As we have seen in the previous chapters, various conceptual and empirical issues wait to be carefully sorted out and clarified in the NE literature. These concerns, however, do not license wholesale methodological anarchy. On the contrary, they make it especially pressing for the practitioners and the philosophers of NE's parent disciplines to assess the potential for success in NE research.

To conclude, it is true that NE is still in its infancy and that considerable achievements may await its pioneers in the years to come. Yet, the time is ripe for beginning to distinguish between alluring marketing hype and well-founded hopes. As I argued in this enquiry, NEs can provide valuable incremental contributions by enriching specific economic models of choice with neuro-psychological insights. Still, they have hitherto failed to demonstrate that economists will usually find it convenient to include several neural insights into their models or substitute economic constructs with neuro-psychological ones. Whatever success NEs will have in informing and constraining the accounts of choice behaviour developed by economists, psychologists and neuroscientists, NE is unlikely to provide a grand revolutionary synthesis spanning its own parent disciplines.

REFERENCES

Achinstein, P. 1968. *Concepts of Science. A Philosophical Analysis.* Baltimore: Johns Hopkins Press.

Aertsen, A. and Preissl, H. 1991. Dynamics of Activity and Connectivity in Physiological Neuronal Networks, *Nonlinear Dynamics and Neuronal Networks*. Schuster, H. Ed. Verlag (Weinheim), p.281-301.

Ainslie, G. 1992. *Picoeconomics: The Strategic Interaction of Successive Motivational States Within the Person*. Cambridge University Press.

Aizawa, K. 2009. Neuroscience and multiple realization: a reply to Bechtel and Mundale. *Synthese*, Vol.167, no.3, p.493-510.

Allais, M. 1953. Le Comportement de l'Homme Rationnel devant le Risque: Critique des Postulats et Axiomes de l'Ecole Americaine. *Econometrica*, 21, p.503-46.

Allais, M. and O. Hagen, eds. 1979. *Expected Utility Hypotheses and the Allais Paradox*. Dordrecht: Reidel.

Allman, J., Hakeem, A. and Watson, K. 2002. Two phylogenetic specializations in the human brain. *The Neuroscientist*, Vol.8, no.4, p.335-346.

Anderson, M.L. 2006. Evidence for massive redeployment of brain areas in cognitive function. *Proceedings of the Cognitive Science Society, 28*, p.24–29.

Anderson, M.L. 2007. The massive redeployment hypothesis and the functional topography of the brain. *Philosophical Psychology, 21*(2), p.143–174.

Angner, E. 2011. Are Subjective Measures of Well-Being 'Direct'? *Australasian Journal of Philosophy*, 89 (1), p.115-130.

Aron, A.R., Gluck, M.A. and Poldrack, R.A. 2006. Long-term testretest reliability of functional MRI in a classification learning task. *NeuroImage*, 29, p.1000-1006.

Arrow, K.J. 1977. Extended Sympathy and the Possibility of Social Choice. *American Economic Review*, 67(1), p.219-225.

Arrow, K.J. 1987. Rationality of Self and Others in an Economic System. In Hogarth, R.M. and Reder, M.W. (eds.). *Rational Choice.* Chicago: The University of Chicago Press.

Ashburner, J. and Friston, K.J. 1999. Nonlinear spatial normalization using basis functions. *Hum. Brain Mapping*, 7, p.254–266.

Aumann, R., 1962. Utility theory without the completeness axiom. *Econometrica*, 30, p.445-462.

Aydinonat, N.E. 2010. Neuroeconomics: more than inspiration, less than revolution. *Journal of Economic Methodology*, 17, No. 2, p.159-169.

Bargh, J. A. and T. L. Chartrand. 1999. The unbearable automaticity of being. *American Psychologist* 54: 462-79.

Barnes, B. 1974. *Scientific Knowledge and Sociological Theory*. London, Routledge.

Baron-Cohen, S. 1995. *Mindblindness: an essay on autism and theory of mind*. MIT Press/Bradford Books.

Barraza, J.A., and Zak, P.J. 2009. Empathy Toward Strangers Triggers Oxytocin Release and Subsequent Generosity. *Annals of the New York Academy of Sciences*, 1167, p.182-189.

Basso, M.A. and Wurtz, R.H. 1997. Target uncertainty modulates neuronal activity. *Nature*, 389, p.66-69.

Basso, M.A. and Wurtz, R.H. 1998. Modulation of neuronal activity in superior colliculus by changes in target probability. *J. Neurosci*, 18 (18), p.7519-7534.

Baumberger, J. 1977. No Kuhnian Revolutions in Economics. *Journal of Economic Issues* 11: 1-20.

Baumeister, R.F. 2003. The Psychology of Irrationality: Why People Make Foolish, Self-Defeating Choices. In *The Psychology of Economic Decisions. Vol. 1: Rationality and Well-Being*, Brocas I. and Carrillo, J.D. (ed), p.3-16. Oxford: Oxford University Press.

Baumeister, R.F. and Sommer, K.L. 1997. Consciousness, free choice, and automaticity. In *Advances in social cognition*, Vol. X, ed. R. S. Wyer Jr., 75-81. Mahwah, NJ: Lawrence Erlbaum.

Baumgartner, T., Heinrichs, M., Vonlanthen, A., Fischbacher, U. and Fehr, E. 2008. Oxytocin shapes the neural circuitry of trust and trust adaptation in humans. *Neuron*, 58(4): 639-650.

Baxter, M.G. and Murray, E.A. 2002. The amygdala and reward. *Nat Rev Neurosci* 3:563–73.

Bayer, H.M. and Glimcher, P.W. 2005. Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron*, 47: 129-141.

Bechtel, W. 2002a. Aligning multiple research techniques in cognitive neuroscience: Why is it important? *Philosophy of Science*, 69, S48-S58.

Bechtel, W. 2002b. Decomposing the Mind-Brain: A Long-Term Pursuit. *Brain and Mind, 3*, p.229-42.

Bechtel, W. 2008. *Mental mechanisms: Philosophical perspectives on cognitive neuroscience*. London: Routledge.

Bechtel, W. Forthcoming. The epistemology of evidence in cognitive neuroscience. In R. Skipper Jr., C. Allen, R. A. Ankeny, C. F. Craver, L. Darden, G. Mikkelson, and R. Richardson (eds.), *Philosophy and the Life Sciences: a Reader*. Cambridge, MA: MIT Press.

Bechtel, W. and Mundale, J. 1999. Revisiting multiple realization. *Philosophy of Science, 66*, 175-205.

Bechtel, W. and Richardson, R.C. 1993. *Discovering complexity: Decomposition and localization as scientific research strategies*. Princeton, NJ: Princeton University Press.

Bechtel, W. and Richardson, R.C. 2010. Neuroimaging as a tool for functionally decomposing cognitive processes. In S. J. Hanson and M. Bunzl, *Foundational issues in human brain mapping*, p.241-262. Cambridge, MA: MIT Press.

Bechtel, W. and Stufflebeam, R.S. 2001. Epistemic Issues in Procuring Evidence about the Brain: the Importance of Research Instruments and Techniques. In W.P. Bechtel, P. Mandik, J. Mundale & R.S. Stufflebeam (eds.), *Philosophy and the Neurosciences: A Reader*, p.55-81. Blackwell.

Bell, D. 1982. Regret in Decision Making under Uncertainty. *Operations Res*. 20, p.961- 81.

Bell, D. 1985. Disappointment in Decision Making under Uncertainty. *Operations Res*. 33, p.1-27.

Belliveau, J.W., Kennedy, D.N., McKinstry, R.C., Buchbinder, B.R., Weisskoff, R.M., Cohen, M.S., Vevea, J.M., Brady, T.J. and Rosen, B.R. 1991. Functional mapping of the human visual cortex by magnetic resonance imaging, *Science*, 254, p.716-719.

Benabou, R. and Tirole, J. 2003. Willpower and personal rules. *Journal of Political Economy*, 112, p.848-886.

Benhabib, J. and Bisin, A. 2005. Modeling internal commitment mechanisms and self-control: a neuroeconomics approach to consumption-saving decisions. *Games Econ. Behav.*, 52 (2), p.460–92.

Bentham, J. 1789 [1907]. *An introduction to the principles of morals and legislation*. Oxford: Clarendon Press.

Bergstrom, L. 1993. Quine, underdetermination and scepticism. *Journal of Philosophy*, 90, p.331-358.

Berker, S. 2009. The moral insignificance of neuroscience. *Philosophy and Public Affairs*, 37 (4), p.293-329.

Bernheim, B.D. 2009. On the Potential of Neuroeconomics: A Critical (but Hopeful) Appraisal. *American Economic Journal: Microeconomics*, 1(2), p.1-41.

Bernheim, B.D. Rangel, A. 2004. Addiction and cue-triggered decision processes. *American Economic Review*, 94, p.1558–90.

Bernheim, B.D. and Rangel, A. 2008. Choice-Theoretic Foundations for Behavioral Welfare Economics. In *The Foundations of Positive and Normative Economics: A Handbook*, ed. Andrew Caplin and Andrew Schotter, p.155–92. Oxford: Oxford University Press.

Berns, G.S., Chappelow, J., Zink, C.F., Pagnoni, G., Martin-Skurski, M.E. and Richards, J. 2005. Neurobiological correlates of social conformity and independence during mental rotation. *Biology Psychiatry*, 58: 245-53.

Berridge, K.C. 1996. Food Reward: Brain Substrates of Wanting and Liking. *Neuroscience and Biobehavioral Reviews*, 20(1): 1–25.

Berridge, K.C. and Robinson, T.E. 1998. What is the role of dopamine in reward: Hedonic impact, reward learning, or incentive salience? *Brain Research Reviews*, 28: 309–369.

Bhatt, M. and Camerer, C.F. 2005. Self-Referential Thinking and Equilibrium as States of Mind in Games: fMRI Evidence. *Games and Economic Behavior*, 52: 424-459.

Bhattacharyya, A., Pattanaik, P.K., and Xu, Y. 2011. Choice, Internal Consistency and Rationality. *Economics and Philosophy*, 27, p.123-149.

Bickle, J. 1996. New Wave Psychophysical Reductionism and the Methodological Caveats. *Philosophy and Phenomenological Research*, vol.56, no.1, p.57-78.

Bickle, J. 1998. *Psychoneural Reduction: The New Wave*. Cambridge, MA: MIT Press.

Bickle, J. 2003. *Philosophy and Neuroscience: A Ruthlessly Reductive Account*. Dordrecht: Kluwer Academic Publishers.

Bielsky, I., Hu, S., Ren, X., Terwilliger, E. and Young, L. 2005. The V1a vasopressin receptor is necessary and sufficient for normal social recognition: A gene replacement study. *Neuron*, 47, p.503–513.

Binmore, K. 1999. Why experiment in economics? *Economic Journal*, 109(453): F16–F24.

Binmore, K. 2008. *Rational Decisions. Gorman Lectures in Economics.* Princeton University Press.

Black, M. 1962. *Models and Metaphors. Studies in Language and Philosophy*. Ithaca, New York: Cornell University Press.

Blaug, M. 1975. Kuhn versus Lakatos, or paradigms versus research programmes in the history of economics. *History of Political Economy*, vol.7, 41, p.399-433.

Blaug, M. 1992. *The methodology of economics, or, How economists explain*, 2nd ed. Cambridge University Press: Cambridge and New York, NY.

Bloor, D. 1984. The Sociology of Reasons, or why 'Epistemic Factors' are really 'Social Factors'. In J.R. Bloor, D. 1991. *Knowledge and Social Imagery*, 2nd ed. Chicago: University of Chicago Press.

Bone, J., Hey, J. and Suckling, J. 1999. Are Groups More (or Less) Consistent than Individuals? *J. Risk Uncertainty*, 18, p.63- 81.

Boyd, R. 1984. The Current Status of Scientific Realism. In J. Leplin, ed. *Scientific Realism*. Berkeley: University of California Press, p.41-82.

Bradley, R. Forthcoming. *Decision Theory with a Human Face*. CUP.

Braitenberg, V. and Schuez, A. 1998. *Cortex: Statistics and Geometry of Neuronal Connectivity*, 2nd ed. Springer, Berlin.

Braithwaite, R. 1953. *Scientific Explanation*. Cambridge: Cambridge University Press.

Brandstatter, H. 1991. Emotions in Everyday Life Situations: Time Sampling of Subjective Experience. In F. Strack, M. Argyle, and N. Schwarz, eds., *Subjective Well-Being: An Interdisciplinary Perspective* (Oxford: Pergamon,), p.173-92.

Brickman, P., Coates, D. and Janoff-Bulman, R. 1978. Lottery Winners and Accident Victims: Is Happiness Relative? *Journal of Personality and Social Psychology*, 36, 917-927.

Broad, C.D. 1925. *The Mind and its Place in Nature*. New York: The Humanities Press Inc, London: Routledge & Kegan Paul LTD.

Brocas, I. and Carrillo, J. D. 2008. The Brain as a Hierarchical Organization. *American Economic Review*, 98, p.1312-1346.

Brocas,I. and Carrillo, J.D. 2010. Neuroeconomic theory: Using neuroscience to understand the bounds of rationality. Available at: http://www.voxeu.org/index.php?q=node/4758, 18 March 2010.

Bruni, L. and Porta, P.L. 2005. *Economics and Happiness: Framings of Analysis*. Oxford: Oxford University Press.

Bruni, L. and Porta, P.L. 2007. *Handbook on the Economics of Happiness.* Cheltenham: Edward Elgar.

Bruni, L. and Sugden, R. 2007. The road not taken: how psychology was removed from economics and how it might be brought back. *The Economic Journal, 117*, 146-73.

Bub, D.N. 2000, Methodological issues confronting PET and fMRI studies of cognitive function. *Cognitive Neuropsychology*, 17 (5), p.467–484.

Bullmore, E., Brammer, M.. Williams, S., Rabe-Hesketh, S.; Janot, N., David, A., Mellers, J., Howard, R. and Sham, P. 1995. Statistical Methods of Estimation and Inference for Functional MR Image Analysis. *Magnetic Resonance in Medicine*, 35, 261-77.

Buonomano, D. and Merzenich, M. 1998. Cortical plasticity: From synapses to maps. *Annual Review of Neuroscience*, 21, p.149-186.

Buxton, R.B. 2002. *Introduction to functional magnetic resonance imaging: principles and techniques* Cambridge University Press (UK).

Buzsaki, G., Kaila, K. and Raichle, M. 2007. Inhibition and brain work. *Neuron*, 56, p.771-783.

Cabeza, R. and Nyberg, L. 1997. Imaging cognition: An empirical review of PET studies with normal subjects. *Journal of Cognitive Neuroscience*, 9, p.1-26.

Cabeza, R. and Nyberg, L. 2000. Imaging cognition II: An empirical review of 275 PET and fMRI studies. *Journal of Cognitive Neuroscience, 12*, p.1-47.

Camerer, C.F. 1998. Bounded rationality in individual decision making. *Experimental Economics*, Vol.1, No.2, p.163-183.

Camerer, C.F. 1999. Behavioral Economics: Reunifying Psychology and Economics. *Proceedings of the National Academy of Sciences of the United States of America*, 96(19): 10575-7.

Camerer, C.F. 2003. What is Neuroeconomics? Available online at: http://www.neuro-economics.org/

Camerer, C.F. 2005. What is Neuroeconomics? Available online at: http://www.hss.caltech.edu/~camerer/web_material/n.html

Camerer, C.F. 2006. Wanting, Liking, and Learning: Speculations on Neuroscience and Paternalism. *The University of Chicago Law Review*, 73(1): 87-110.

Camerer, C.F. 2007. Neuroeconomics: Using Neuroscience to Make Economic Predictions. *The Economic Journal*, 117(519), C26–C42.

Camerer, C.F. 2008a. The Case for Mindful Economics. In *The Foundations of Positive and Normative Economics. A Handbook*. Caplin, A. and Schotter, A. Eds. p.43-69.

Camerer, C.F. 2008b. The Potential of Neuroeconomics. *Economics and Philosophy*, 24, 369-379.

Camerer, C., Issacharoff, S., Loewenstein, G., O'Donoghue, T. and Rabin, M. 2003. Regulation for conservatives: Behavioral economics and the case for "asymmetric paternalism." *University of Pennsylvania Law Review*, 151, p.1211-1254.

Camerer, C.F. and Loewenstein, G. 2004. Behavioral Economics: Past, Present, Future. In *Advances in Behavioral Economics*. New York: Princeton University Press, p.3-51.

Camerer, C.F., Loewenstein, G., Prelec, D. 2004. Neuroeconomics: Why Economics Needs Brains. *Scandinavian Journal of Economics*, 106(3): 555-79.

Camerer, C.F., Loewenstein, G., Prelec, D. 2005. Neuroeconomics: how neuroscience can inform Economics. *Journal of Economic Literature*, 43(1), p.9-64.

Caplin, A. 2008. Economic Theory and Psychological Data: Bridging the Divide', in *The Foundations of Positive and Normative Economics: A Handbook*, eds. A. Caplin and A. Schotter, New York: Oxford University Press, p.336–371.

Caplin, A. and Dean, M. 2008. Dopamine, reward prediction error, and economics. *The Quarterly Journal of Economics*, 123(2), p.663–701.

Caplin, A., Dean, M., Glimcher, P.W. and Rutledge, R.B. 2010. Measuring beliefs and rewards: A neuroeconomic approach. *The Quarterly Journal of Economics*, 125(3), p.923-960.

Carlson, N.A. 1992. *Foundations of Physiological Psychology*. Needham Heights, Massachusetts: Simon & Schuster.

Carnap, R. 1938. Foundations of Logic and Mathematics. In Otto Neurath, Charles Morris and Rudolf Carnap (eds.), *International Encyclopaedia of Unified Science. Vol. 1*. Chicago: University of Chicago Press, p.139-213.

Cartwright, N. 1983. *How the Laws of Physics Lie*. Oxford: Oxford University Press.

Cartwright, N. 1994. *Nature's Capacities and Their Measurements*. Oxford: Oxford University Press.

Cartwright, N. 1998. Capacities. In Davis, D. et al. (eds.). *The Handbook of Economic Methodology*. Edward Elgar: Cheltenham.

Cartwright, N. 1999. The Vanity of Rigour in Economics. CPNSS Discussion Paper.

Cartwright, N. 2007. *Hunting Causes and Using Them: Approaches in Philosophy and Economics*. Cambridge: Cambridge University Press.

Chalmers, D. J. 1996. *The Conscious Mind: Toward a Fundamental Theory*, Oxford: Oxford University Press.

Chalmers, D. J. 2004. How Can We Construct a Science of Consciousness? In M. Gazzaniga (ed.), *The Cognitive Neurosciences III*, MIT Press.

Chang, R. 1997. *Incommensurability, Incomparability, and Practical Reason*, Cambridge: Harvard University Press.

Chang, R. 2002. *Making Comparisons Count*. New York: Routledge.

Chew, S.H. 1983. A Generalization of the Quasilinear Mean with Applications to the Measurement of Income Inequality and Decision Theory Resolving the Allais Paradox. *Econometrica*, 51, p.1065- 92.

Chu, Y. and Chu, R. 1990. The Subsidence of Preference Reversals in Simplified and Marketlike Experimental Settings: A Note. *A.E.R.* 80: 902-11.

Churchland, P.M. 1981. Eliminative Materialism and the Propositional Attitudes. *The Journal of Philosophy*, Vol. 78, No.2, p.67-90.

Churchland, P.M. 1985. Reduction, Qualia, and the Direct Introspection of Brain States. *The Journal of Philosophy*, Vol.82, No.1, p.8-28.

Churchland, P.M. and Churchland, P.S. 1996. Intertheoretic Reduction: a Neuroscientist's field guide. In *Philosophy and the Neurosciences*. 2001. W. Bechtel, P. Mandik, J. Mundale, and R. Stufflebeam (eds.), Oxford: Blackwell Publishers.

Churchland, P.M. and Churchland, P.S. 1998. Intertheoretic Reduction: A Neuroscientist's Field Guide. In *On the Contrary*. Cambridge, MA: MIT Press, p.65-79.

Churchland, P. S. 1986. *Neurophilosophy: Toward a unified science of the mind/.brain*. Cambridge, MA: MIT Press.

Churchland, P. S. 1988. Replies to Corballis and Bishop. *Biology and Philosophy*, 3, p.393-402.

Cohen, D. and Halgren, E. 2004. Magnetoencephalography. In *Encyclopedia of Neuroscience*, Adelman G., Smith B., Ed. Elsevier.

Colander, D. 2007. Edgeworth's Hedonimeter and the Quest to Measure Utility. *Journal of Economic Perspectives*, Vol.21, No.2, p.215-225.

Coltheart, M. 2004. Brain Imaging, Connectionism, and Cognitive Neuropsychology. *Cognitive Neuropsychology*, 21(1): 21-25.

Coltheart, M. 2005. What has functional neuroimaging told us about the mind (so far)? *Position paper presented to the European Cognitive Neuropsychology Workshop, Bressanone.*

Conlisk, J. 1996. Why Bounded Rationality? *Journal of Economic Literature*, 34, p.669-700.

Cox, J.C. and Grether, D.M. 1996. The Preference Reversal Phenomenon: Response Mode, Markets and Incentives. *Econ. Theory*, 7, pp.381-405.

Craver, C.F. 2006. When mechanistic models explain. *Synthese*, 153, p.355-376.

Craver, C.F. 2007. *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. Clarendon: Oxford.

Craver, C.F. and Alexandrova, A. 2008. No Revolution Necessary: Neural Mechanisms for Economics. *Economics and Philosophy*, 24, p.381-406.

Csikszentmihalyi, M. 1990. *Flow: The psychology of optimal experience.* New York: Harper and Row.

Da Costa, N.C. and French, A. 2003. *Science and Partial Truth*. New York: Oxford University Press.

Damasio, A.R. 1994. *Descartes' Error: Emotion Reason, and the Human Brain*. NY: G. P. Putnam.

Darden, L. and Maull. N. 1977. Interfield Theories. *Philosophy of Science*, Vol.44, No.1, p.43-64.

Davidson, D. 1963. Actions, Reasons and Causes. *Journal of Philosophy*, 60, p.685-700.

Davidson, D. 1980 [1974]. Psychology as Philosophy. In *Essays on Actions and Events*, p.229-238. New York: Clarendon Press. Originally Published in: S.C. Brown (Ed.) 1974. *Philosophy of Psychology*. London: MacMillan.

Dayan, P. and Niv, Y. 2008. Reinforcement Learning: The Good, The Bad and the Ugly. *Current Opinion in Neurobiology*, 18(2): p.185–96.

Delgado, M.R., Frank, R.H., Phelps, E.A. 2005. Perceptions of moral character modulate the neural systems of reward during the trust game. *Nature Neuroscience*, 8: 1611-18.

Dennett, D.C. 1991. *Consciousness Explained*. Boston: Little, Brown & Co.

Detre, J.A., Leigh, J.S., Williams, D.S., and Koretsky, A.P. 1992. Perfusion imaging. *Magnetic Resonance in Medicine*, 23(1), p.37-45.

Devitt, M. 1979. Against Incommensurability. *Australasian Journal of Philosophy*, 57, p.29-49.

De Vroey, M. 1975. The transition from classical to neoclassical economics: a scientific revolution. *Journal of Economic Issues*, 9, p.415-440.

Diamond, P.A. and Hausman, J.A. 1994. Contingent Valuation: is some number better than no number? *Journal of Economic Perspectives*. 8(4): 45-64.

Dietrich, F. and List, C. 2011. A model of non-informational preference change. *Journal of theoretical politics*, 23 (2), p.145-164.

Dixit, A.K. 1990. *Optimization in Economic Theory*. Oxford University Press.

Dobbs, D. 2005. Hard Science or "Technicolor Phrenology"? The Controversy over fMRI. *Scientific American Mind*, 16: 24–31.

Donders, F.C. 1969. On the speed of mental processes, *Acta Psychologica*, 30, 412-431.

Dorris, M.C. and Glimcher, P.W. 2004. Activity in Posterior Parietal Cortex Is Correlated With the Relative Subjective Desirability of Action. *Neuron*, 44, 365–378.

Douglas, R.J. and Martin, K.A. 2004. Neuronal circuits of the neocortex. *Annual Review of Neuroscience*, 27, p.419-451.

Duhem, P. 1906. *The Aim and Structure of Scientific Theories*. Transl. by P. Wiener. Princeton: Princeton University Press, 1954.

Dupré, J. 1983. The disunity of science. *Mind*, 92, p.321-346.

Dupré, J. 1993. *The Disorder of Things. Metaphysical Foundations of the Disunity of Science*. Cambridge (MA): Harvard University Press.

Edgeworth, F.Y. 1881 [1967]. *Mathematical Psychics: an essay on the application of mathematics to the moral sciences*. New York: Kelley.

Eichner, A. 1983. Why Economics is not yet a science. In A. Eichner, ed. *Why Economics Is not yet a Science*. Armonk, New York: M.E. Sharpe, p.205-241.

Elliott, R., Friston, K.J. and Dolan, R.J. 2000. Dissociable neural responses in human reward systems. *Journal of Neuroscience*, 20: p.6159–65.

Ellsberg, D. 1961. Risk, ambiguity and the Savage axioms. *Quarterly Journal of Economics*, 75: 643-69.

Elster, J. 1983. *Sour Grapes: Studies in the Subversion of Rationality*. Cambridge: Cambridge University Press.

Endicott, R. 1998. Collapse of the new wave. *Journal of Philosophy*, 95, p.53–72.

Endicott, R. 2001. Post-structuralist Angst-critical notice: John Bickle, psychoneural reduction: The new wave. *Philosophy of Science*, 68, p.377–393.

English, J. 1973. Underdetermination: Craig and Ramsey. *Journal of Philosophy*, 70, p.453-462.

Fehr, E. and Camerer, C.F. 2007. Social neuroeconomics: the neural circuitry of social preferences. *Trends in Cognitive Sciences*, Vol.11, No.10, p.419-426.

Feigl, H. 1970. The 'Orthodox' View of Theories: Remarks in Defense as well as Critique. In M. Radner and S. Winokur (eds.), *Minnesota Studies in the Philosophy of Science*, p.3-16. University of Minnesota Press.

Fernandes, H.L. and Kording, K.P. 2010. In Praise of 'False' Models and Rich Data. *Journal of Motor Behavior*, 42 (6), p.343-349.

Feyerabend, P.K. 1962. Explanation, Reduction, and Empiricism. In *Scientific Explanation, Space, and Time*, H. Feigl and G. Maxwell (eds.), Minneapolis: University of Minnesota Press, p.28-97.

Feyerabend, P.K. 1963. Materialism and the Mind-Body Problem. *Review of Metaphysics*, 18 (1), 65, p.49-66.

Feyerabend, P.K. 1970. Against Method: Outline of an Anarchistic Theory of Knowledge. In M. Radner and S. Winokur (ed.), *Analysis of Theories and Methods of Physics and Psychology*. Minneapolis: University of Minneapolis Press, p.17-130.

Feyerabend, P.K. 1975. *Against Method. Outline of an Anarchistic Theory of Knowledge*. London: New Left Books.

Feyerabend, P.K. 1981. *Realism, Rationalism and Scientific Method. Philosophical papers*, Cambridge: Cambridge University Press.

Feyerabend, P.K. 1987. Putnam on Incommensurability. *The British Journal for the Philosophy of Science*, 38, p.75-81.

Field, H. 1973. Theory of Change and Indeterminacy of Reference. *Journal of Philosophy*, 70, p.462-81.

Fine, A. 1967. Consistency, Derivability, and Scientific Change. *Journal of Philosophy*, 64, p.231-240.

Fishburn, P.C. 1982. Nontransitive Measurable Utility, *J. Math. Psych*. 26, p.31-67.

Fisher, I. 1918. Is 'Utility' the Most Suitable Term for the Concept It is Used to Denote? *American Economic Review*, 8(2): 335–37.

Fodor, J. 1974. Special Sciences (or: The Disunity of Science as a Working Hypothesis). *Synthese*, 28, p.97-115.

Fodor, J. 2000. *The Mind Doesn't Work That Way: The Scope and Limits of Computational Psychology*, MIT Press, Cambridge, MA.

Fox, P.T., Laird, A.R., Fox, S.P., Fox, M., Uecker, A.M., Crank, M., Koenig, S.F., Lancaster, J.L. 2005. BrainMap taxonomy of experimental design: Description and evaluation. *Hum Brain Mapp*, 25, p.185-98.

Fox, P.T., Parsons, L.M., Lancaster, J.L. 1998. Beyond the single study: Function-location meta-analysis in cognitive neuroimaging. *Curr Opin Neurobiology*, 8, p.178-87.

Fox, P.T. and Raichle, M.E. 1986. Focal physiological uncoupling of cerebral blood flow and oxidative metabolism during somatosensory stimulation in human subjects, *Proceedings of the National Academy of Science Usa*, 83(4), p.1140-1144.

Fox, P.T., Raichle, M.E., Mintun, M.A. and Dence, C. 1988. Nonoxidative glucose consumption during focal physiologic neural activity, *Science*, 241(4864), p.462-464.

Frederick, S. and Loewenstein, G. 1999. Hedonic adaptation. In *Well-Being: The Foundations of Hedonic Psychology*, eds. Kahneman, Daniel, Diener, Ed and Schwarz, Norbert, p.302-329. New York: Russell Sage Foundation.

Frederick, S., Loewenstein, G. and O'Donoghue, T. 2002. Time discounting and time preference: a critical review. *Journal of Economic Literature*, 40, p.351-401.

Fredrickson, B.L. and D. Kahneman. 1993. Duration Neglect in Retrospective Evaluations of Affective Episodes. *Journal of Personality and Social Psychology*, LXV, 45–55.

Freeman, W.J. and Watts, J.W. 1942. *Psychosurgery in the Treatment of Mental Disorders and Intractable Pain*. Springfield: Thomas.

Friedman, M. 1953. The Methodology of Positive Economics. In *Essays in Positive Economics*. Chicago: Chicago University Press.

Friedman, M. 1974. Explanation and Scientific Understanding. *Journal of Philosophy*, 71, p.5-19.

Friston, K.J. 2002. Beyond phrenology: What can neuroimaging tell us about distributed circuitry? *Annual Review of Neuroscience, 25*, p.221–250.

Friston, K.J. 2003. Statistical parametric mapping. In Kotter, R. (ed). *Neuroscience Databases: A Practical Guide*, ch.16.

Friston, K.J. and Henson, R.N. 2006. Commentary on: Divide and conquer; a defence of functional localisers. *NeuroImage*, 30, p.1097-1099.

Friston, K.J., Holmes, A.P., Poline, J-B., Grasby, P.J., Williams, S.C., Frackowiak, R.S. and Turner, R. 1995a. Analysis of fMRI time-series revisited. *Neuroimage*, 2, p.45-53.

Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.B., Frith, C.D. and Frackowiak, R.S.J. 1995b. Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, *2*, p.189-210.

Friston, K.J., Price, C.J., Fletcher, P., Moore, C., Frackowiak, R.S. and Dolan, R.J. 1996. The trouble with cognitive subtraction, *Neuroimage*, 4(2), 97-104.

Friston, K.J., Rotshtein, P., Geng, J.J., Sterzer, P. and Henson, R.N. 2006. A critique of functional localisers. *NeuroImage*, 30, p.1077-1087.

Fudenberg, D. and Levine, D.K. 2006. A dual-self model of impulse control. *American Economic Review*, 96, p.1449–1476.

Fumagalli, R. 2010. The Disunity of Neuroeconomics: a Methodological Appraisal. *Journal of Economic Methodology*, vol.17, no.2, p.119-131.

Fumagalli, R. 2011. On the Neural Enrichment of Economic Models: Tractability, Trade-offs and Multiple Levels of Description. *Biology and Philosophy*, 26, p.617–635.

Gabaix, X. and Laibson, D. 2008. The Seven Properties of Good Models. In *The Foundations of Positive and Normative Economics*, Caplin, A. and Schotter, A. (ed). Oxford University Press, p.292-299.

Gazzaniga, M.S., Ivry, R. and Mangun, G. 2002. *Cognitive Neuroscience*, Second Edition, W.W. Norton & Company, New York, NY.

Gazzaniga, M.S. and LeDoux, J.E. 1978. *The Integrated Mind*. New York: Plenum.

Gerlach, C. 2007. A review of functional imaging studies on category specificity. *Journal of Cognitive Neuroscience* 19, p.296–314.

Ghiselin, M.T. 1978. The economy of the body. *American Economic Review*, 68: 233–7.

Gibbard, A and Varian, H.R. 1978. Economic Models. *The Journal of Philosophy*, 75 (11), p.664-677.

Giere, R. 1988. *Explaining Science: A Cognitive Approach*. Chicago: University of Chicago Press.

Gigerenzer G., Todd P.M. and the ABC Research Group. 1999. *Simple heuristics that make us smart*. Oxford University Press, p.119-140.

Gintis, H. 2004. Towards the unity of the human behavioral sciences. *Politics, Philosophy and Economics*, 3: 37–57.

Gintis, H. 2007. A framework for the integration of the behavioral sciences. *Behavioral & Brain Sciences*, 30, 1-61.

Gispert, J.D., Pascau, J., Reig, S., Martinez-Lazaro, R., Molina, V., Garcia-Barreno, P., and Desco, M. 2003. Influence of the normalization template on the outcome of statistical parametric mapping of PET scans. *NeuroImage*, 19(3), p.601-612.

Glaeser, E. 2008. Researcher Incentives and Empirical Methods. In *The Foundations of Positive and Normative Economics: A Handbook*, ed. Andrew Caplin and Andrew Schotter. Oxford University Press, Ch.13.

Glennan, S.S. 2005. Modeling mechanisms. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 36(2), p.443–464.

Glimcher, P.W. 2003. *Decisions, Uncertainty, and the Brain: The Science of Neuroeconomics*. Bradford Books.

Glimcher, P.W. 2009. Choice: Towards a standard back-pocket model. In P.W. Glimcher, C. Camerer, E. Fehr and R. Poldrack, eds., *Neuroeconomics: Decision Making and the Brain*, p.503-521. London: Elsevier.

Glimcher, P.W. 2010. *Foundations of Neuroeconomic Analysis*. Oxford University Press.

Glimcher, P.W., Dorris, M.C. and Bayer, H.M. 2005. Physiological utility theory and the neuroeconomics of choice. *Games and Economic Behavior*, 52: 213–256.

Glimcher, P.W. and Rustichini, A. 2004. Neuroeconomics: The Consilience of Brain and Decision. *Science*, 306(5695): 447–52.

Glymour, C. 1994. Methods of cognitive neuropsychology. *British Journal for the Philosophy of Science*, *45*, 815-835.

Gold, B.T., and Buckner, R.L. 2002. Common prefrontal regions coactivate with dissociable posterior regions during controlled semantic and phonological tasks. *Neuron*, 35, p.803-12.

Goldberg, E. 2005. *The wisdom paradox, how your mind can grow stronger as your brain grows older*. New York: Gotham Books.

Goldstein, D.G. and Gigerenzer, G. 2002. Models of ecological rationality: The recognition heuristic. *Psychological Review*, 109, p.75-90.

Gould, S.J. 1997. Nonoverlapping magisteria. *Natural History*, 106, p.16-22.

Gould, S.J. 1999. *Rocks of Ages: Science and Religion in the Fullness of Life*. New York: Ballantine Publications.

Greene, J.D. 2007. The secret joke of Kant's soul. In W. Sinnott-Armstrong (Ed.), *Moral Psychology, Vol.3: The Neuroscience of Morality: Emotion, Disease, and Development*, p.35-79. MIT Press: Cambridge, MA.

Greene, J.D., Sommerville, R.B., Nystrom, L.E., Darley, J.M., and Cohen, J.D. 2001. An fMRI investigation of emotional engagement in moral judgment. *Science*, Vol. 293, p.2105-2108.

Griffin, D. and Tversky, A. 1992. The Weighing of Evidence and the Determinants of Confidence. *Cognitive Psychology*, 24(3), p.411-35.

Griffin, J. 1986. *Well-Being: its Meaning, Measurement, and Moral Importance*. Clarendon Press.

Gruene-Yanoff, T. 2009. Learning from minimal economic models. *Erkenntnis*, 70, p.81-99.

Guala, F. 1998. Experiments as mediators in the non-laboratory sciences, *Philosophica*, 62, p.901-918.

Guala, F. 2002. Models, Simulations, and Experiments. In L. Magnani and N. J. Nersessian (eds.) *Model-Based Reasoning: Science, Technology, Values*. New York: Kluwer, p.59-74.

Guala, F. 2003. Experimental localism and external validity. *Philosophy of Science*, 70, p.1195-1205.

Guala, F. 2005. *The Methodology of Experimental Economics*. Cambridge: Cambridge University Press.

Gul, F. 1991. A Theory of Disappointment in Decision Making under Uncertainty. *Econometrica*, 59, pp. 667-86.

Gul, F. and Pesendorfer, W. 2008. The case for mindless economics. In Caplin, A. and Schotter, A. eds. *The foundations of positive and normative economics*, p.1-40. New York: Oxford University Press.

Gusnard, D.A. and Raichle, M.E. 2001. Searching for a baseline: functional imaging and the resting human brain. *Nat Rev Neurosci*., 2(10), p.685-94.

Guth, W., Schmittberger, R. and Schwarze, B. 1982. An Experimental Analysis of Ultimatum Bargaining. *Journal of Economic Behaviour and Organization*, 3(4): 367-88.

Hacking, I. 1983. *Representing and intervening*. Cambridge, UK: Cambridge University Press.

Halonen, I. and J. Hintikka. 1999. Unification –it's magnificent but is it explanation? *Synthese*, 120, p.27-47.

Hands, D.W. 1985a. Karl Popper and Economic Methodology. *Economics and Philosophy*, 1, p.83-100.

Hands, D. W. 1985b. Second Thoughts on Lakatos. *History of Political Economy*, 17, p.1-16.

Hanks, T.D., Ditterich, J. and Shadlen, M.N. 2006. Microstimulation of macaque area LIP affects decision-making in a motion discrimination task. *Nature Neuroscience*, 9, p.682-689.

Hansen, P.C., Kringelbach, M.L. and Salmelin, R. 2010. *MEG: An Introduction to Methods*. New York: Oxford University Press Inc.

Harada, Y. and Takahashi, T. 1983. The calcium component of the action potential in spinal motoneurones of the rat, *The Journal of Physiology*, 335, p.89-100.

Hardcastle, V.G. and Stewart, C.M. 2002. What Do Brain Data Really Show? *Philosophy of Science*, 69, p. S72–82.

Hare, T.A., O'Doherty, J., Camerer, C.F., Schultz, W. and Rangel, A. 2008. Dissociating the Role of the Orbitofrontal Cortex and the Striatum in the Computation of Goal Values and Prediction Errors. *Journal of Neuroscience*, 28(22): 5623–30.

Harless, D.W. and Camerer, C.F. 1994. The Predictive Utility of Generalized Expected Utility Theories. *Econometrica*, 62, p.1251-89.

Harley, T. 2004. Does Cognitive Neuropsychology Have a Future? *Cognitive Neuropsychology*, 21(1), p.3-16.

Harrison, G. 2008a. Neuroeconomics: A Critical Reconsideration. *Economics & Philosophy*, 24, 303-344.

Harrison, G. 2008b. Neuroeconomics: Rejoinder. *Economics & Philosophy*, 24, 533-544.

Harrison, G. and Ross, D. 2010. The methodologies of neuroeconomics. *Journal of Economic Methodology*, Vol.17, no.2, p.185-196.

Harrison, P. 2007. *The Fall of Man and the Foundations of Science*. Cambridge: Cambridge University Press.

Harsanyi, J.C. 1955. Cardinal Welfare, Individualistic Ethics and Interpersonal Comparisons of Utility. *Journal of Political Economy*, 63 (4), p.309-321.

Hastie, R. 1984. Causes and Effects of Causal Attribution. *Journal of Personality and Social Psychology*, 46(1): 44–56.

Hausman, D.M. 1992. *The Inexact and Separate Science of Economics*, Cambridge: Cambridge University Press.

Hausman, D.M. 2008a. Mindless or Mindful Economics: a Methodological Evaluation. In *The Foundations of Positive and Normative Economics: A Handbook*, ed. Andrew Caplin and Andrew Schotter. Oxford University Press, Ch.6, p.125-152.

Hausman, D.M. 2008b. Why Look Under the Hood. In Hausman, D., ed. 2008. *The Philosophy of Economics: An Anthology*. 3rd. ed. Cambridge: Cambridge University Press, p. 217-221.

Hempel, C.G. 1962. Deductive-Nonlogical vs. Statistical Explanation. In H. Feigl and G. Maxwell (eds.) *Minnesota Studies in the Philosophy of Science*, Vol.111. Minneapolis: University of Minnesota Press.

Hempel, C.G. 1965. *Aspects of Scientific Explanation and other Essays in the Philosophy of Science.* New York: The Free Press.

Henson, R. 2005. What can functional neuroimaging tell the experimental psychologist? *Quarterly Journal of Experimental Psychology*. A 58, 193-234.

Henson, R. 2006. Forward inference using functional neuroimaging: dissociations versus associations. *Trends in Cognitive Science*, vol. 10, no.2, p.64-9.

Hicks, J. and R. Allen. 1934. A Reconsideration of the Theory of Value. *Economica*, 1, p.52-76 and 196-219.

Hindriks, F.A. 2005. Unobservability, tractability and the battle of assumptions. *Journal of Economic Methodology*, 12 (3), p.383-406.

Hindriks, F.A. 2006. Tractability assumptions and the Musgrave-Mäki typology. *Journal of Economic Methodology*, 13 (4), p.401-423.

Hinton, G. E. and Shallice, T. 1991. Lesioning a connectionist network: Investigations of acquired dyslexia. *Psychological Review*, *98*, p.74-95.

Hirshleifer, D.A. and Shumway, T. 2003. Good Day Sunshine: Stock Returns and the Weather. *Journal of Finance*, 58, p.1009-32.

Hodgkin, A.L. and Huxley, A.F. 1952. A quantitative description of membrane current and its application to conduction and excitation in nerve. *Journal of Physiology*, 117, p.500–544.

Hohwy, J. 2007. The Search for Neural Correlates of Consciousness. *Philosophy Compass*, 2/3, p.461-474.

Hooker, C. 1981. Towards a general theory of reduction. Part I: Historical and Scientific Setting. Part II: Identity in Reduction. Part III: Cross-Categorical Reduction. *Dialogue*, 20, p.38–60, 201–236, and 496–529.

Horgan, T. 1993. Nonreductive Materialism and the Explanatory Autonomy of Psychology. In S. Wagner and R. Warner (eds.). *Naturalism: A Critical Appraisal*. Notre Dame, Indiana: University of Notre Dame Press.

Houser, D., Schunk, D. and Xiao, E. 2007. Combining Brain and Behavioral Data to Improve Econometric Policy Analysis. *Analyse & Kritik*, 29: 86-96.

Hsu, M., Bhatt, M., Adolphs, R., Tranel, D. and Camerer, C.F. 2005. Neural Systems Responding to Degrees of Uncertainty in Human Decision-Making. *Science*, 310: 1680-3.

Hubbard, E.M. 2003. A discussion and review of Uttal (2001) *The New Phrenology*, Cognitive Science Online, Vol.1, p.22-33.

Huettel, S.A., Song, A.W. and McCarthy, G. 2004. *Functional Magnetic Resonance Imaging*. Sinauer Associates.

Hume, D. 1740. *A Treatise of Human Nature*. Oxford: Oxford University Press, 2005.

Hutchins, E. 1995. *Cognition in the Wild*. Cambridge, MA: MIT Press.

Hutchison, T.W. 1978. *On Revolutions and Progress in Economic Knowledge*. Cambridge: Cambridge University Press.

Jackson, F. 1982. Epiphenomenal Qualia. *Philosophical Quarterly* 32: 127-136.

Jackson, F. 1986. What Mary Didn't Know. *Journal of Philosophy* 83: 291-295.

Jamison, J.C. 2008. Well-being and Neuroeconomics. *Economics and Philosophy*, 24, p.407-418.

Jevons, W. 1871. *The Theory of Political Economy*. First Edition. London and New York, MacMillan and Co

Jezzard, P., Matthews, P.M. and Smith, S.M. 2002. *Functional magnetic resonance imaging: an introduction to methods*. Oxford University Press, Oxford New York.

Jones, R.M. 2005. Idealization and Abstraction: A Framework. In M.R. Jones and N. Cartwright (eds.), *Idealization XII: Correcting The Model. Idealization and Abstraction in the Sciences* (Amsterdam: Rodopi).

Jueptner, M. and Weiller, C. 1995. Review: does measurement of regional cerebral blood flow reflect synaptic activity? Implications for PET and fMRI, *Neuroimage*, 2, p.148-156.

Kable, J.W. and Glimcher, P.W. 2007. The Neural Correlates of Subjective Value During Intertemporal Choice. *Nature Neuroscience*, 10(11), p.1625-33.

Kacelnik, A. 2006. Meanings of rationality. In S. Hurley and M. Nudds (ed.) *Rational Animals?* Oxford University Press, Oxford, p.87–106.

Kahneman, D. 2000. Experienced Utility and Objective Happiness: A Moment-based Approach. In A. Tversky and D. Kahneman, eds., *Choices, Values, and Frames*, ch. 37. Cambridge: Cambridge University Press and the Russell Sage Foundation.

Kahneman, D. 2003. A Psychological Perspective on Economics. *The American Economic Review*, 93(2): 162-8.

Kahneman, D., Fredrickson, B.L., Schreiber, C.A. and Redelmeier, D.A. 1993. When More Pain Is Preferred to Less: Adding a Better End. *Psychological Science*, IV, p.401-405.

Kahneman, D. and Krueger, A. 2006. Developments in the Measurement of Subjective Well-Being. *Journal of Economic Perspective*, 20(1): 3–24.

Kahneman, D., Kreuger, A., Schkade, D., Schwarz, N. and Stone, A. 2004. A Survey Method for Characterizing Daily Life Experience: The Day Reconstruction Method. *Science*, 306, p.1776–1780.

Kahneman, D., Slovic, P. and Tversky, A. 1982. *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press.

Kahneman, D. and Snell, J. 1990. Predicting Utility. In R. Hogarth, eds., *Insights in Decision Making*, Chicago: University of Chicago Press.

Kahneman, D. and Sugden, R. 2005. Experienced utility as a standard of policy evaluation. *Environmental and Resource Economics*, 32, p.161-181.

Kahneman, D. and Tversky, A. 1979. Prospect Theory. An analysis of decision under risk. *Econometrica*, 47(2): 263-291.

Kahneman, D. and Varey, C. 1991. Notes on the Psychology of Utility. In J. Elster. J. Roemer, eds., *Interpersonal Comparisons of Well-Being*, p.127–163. Cambridge: Cambridge University Press.

Kahneman, D., Wakker, P. and Sarin, R. 1997. Back to Bentham? Explorations of experienced utility. *The Quarterly Journal of Economics*, 112: 375-406.

Keren, G. and Gerritsen, L.E. 1999. On the robustness and possible accounts of ambiguity aversion. *Acta Psychologica*, Vol.103, no.1, p.149-172.

Kiebel, S.J., Poline, J.B., Friston, K.J., Holmes, A.P. and Worsley, K.J. 1999. Robust Smoothness Estimation in Statistical Parametric Maps Using Standardized Residuals from the General Linear Model. *NeuroImage*, 10, p.756-766.

Kim, S.G. and Ugurbil, K. 1997. Comparison of blood oxygenation and cerebral blood flow effects in fMRI: estimation of relative oxygen consumption change, *Magnetic Resonance in Medicine*, 38(1), p.59-65.

Kincaid, H. 1997. *Individualism and the Unity of Science. Essays on Reduction, Explanation and the Special Sciences.* Lanham: Rowman and Littlefield.

King-Casas, B., Tomlin, D., Anen, C., Camerer, C.F., Quartz, S.R. and Montague, P.R. 2005. Getting to know you: reputation and trust in a two-person economic exchange. *Science*, 308: 78-83.

Kitcher, P. 1981. Explanatory Unification. *Philosophy of Science*, Vol.48, No.4, p.507-531.

Kitcher, P. 1989. Explanatory Unification and the Causal Structure of the World. In *Scientific Explanation*, P. Kitcher and W. Salmon, p.410-505. Minneapolis: University of Minnesota Press.

Klein, C. 2010a. Images Are Not the Evidence in Neuroimaging. *British Journal for the Philosophy of Science*, 61, p.265–278.

Klein, C. 2010b. Philosophical Issues in Neuroimaging. *Philosophy Compass 5/2*: 186-198.

Knight, F. 1935. Economics and Human Action. From *The Ethics of Competition and Other Essays*. Knight, F. New York and London: Harper and Brothers.

Knutson, B., Adams, C.M., Fong, G.W. and Hommer, D. 2001a. Anticipation of increasing monetary reward selectively recruits nucleus accumbens. *Journal of Neuroscience*, 21: RC159.

Knutson, B., Fong, G.W., Adams, C.M., Varner, J.L., Hommer, D. 2001b. Dissociation of reward anticipation and outcome with event-related FMRI. *NeuroReport*, 12: 3683–87.

Knutson, B., Fong, G.W., Bennett, S.M., Adams, C.M. and Hommer, D. 2003. A region of the mesial prefrontal cortex tracks monetarily rewarding outcomes: characterization with rapid event-related fMRI. *Neuroimage*, 18: 263-72.

Knutson, B. and Peterson, R. 2005. Neurally reconstructing expected utility. *Games and Economic Behavior*, 52, p.305–315.

Knutson, B., Rick, G.S., Wimmer, E., Prelec, D. and Loewenstein, G. 2007. Neural Predictors of Purchases. *Neuron*, 53: 147-56.

Kosfeld, M., Heinrichs, M., Zak, P.J., Fischbacher, U. and Fehr, E. 2005. Oxytocin Increases Trust in Humans. *Nature*, 435: 673-6.

Krekelberg, B., Boynton, G.M. and van Wezel, R.J. 2006. Adaptation: from single cells to BOLD signals. *Trends in Neuroscience*, 29, p.250-256.

Kuhn, T.S. 1962. The Structure of Scientific Revolutions. Chicago : University of Chicago Press.

Kuhn, T. 1970a. *The Structure of Scientific Revolutions*. 2nd Edition. Chicago: University of Chicago Press.

Kuhn, T. 1970b. Logic of Discovery or Psychology of Research. In I. Lakatos and A. Musgrave (eds.), *Criticism and the Growth of Knowledge*. Cambridge University Press.

Kuhn, T. 1974. Second Thoughts on Paradigms. In F. Suppe, ed., *The Structure of Scientific Theories*, p.459-482, Urbana: University of Illinois Press.

Kuhn, T. 1981. What are Scientific Revolutions?. In L. Krüger, L. Daston and M. Heidelberger, eds., *The Probabilistic Revolution: Ideas in History*, p.7-22, Cambridge Mass.: M.I.T. Press.

Kuhn, T.S. 1982. Commensurability, Comparability, Communicability. In *PSA*, P. Asquith and T. Nickles (eds.). East Lansing: Philosophy of Science Association, p.669-688.

Kuhnen, C. and Knutson, B. 2005. The Neural Basis of Financial Risk Taking. *Neuron*, 47: 763-70.

Kuorikoski, J., Lehtinen, A. and Marchionni, C. 2010. Economic modelling as robustness analysis. *British Journal for the Philosophy of Science*, 61(3), p.541-567.

Kuorikoski, J. and Ylikoski, P. 2010. Explanatory relevance across disciplinary boundaries: the case of neuroeconomics. *Journal of Economic Methodology*, vol. 17, no.2, p.219-228.

Kutas, M. and Dale, A. 1997. Electrical and magnetic readings of mental functions. In Rugg, M.D. (Ed.), *Cognitive neuroscience*, p.197-237. London: University College Press.

Lakatos, I. 1970. Falsification and the methodology of scientific research programmes. In I. Lakatos and A. Musgrave, eds. *Criticism and the growth of knowledge*. Cambridge, p.91-195.

Lakatos, I. 1971. History of Science and Its Rational Reconstruction. In *Boston Studies in Philosophy of Science*, VIII, edited by R. s. Cohen, and C. R. Buck.

Lakatos, I. and Zahar, E.G. 1975. Why Did Copernicus' Research Program Supersede Ptolemy's?. In R. Westman (ed.): *The Copernican Achievement*. Berkeley: University of California Press, p.354-83.

Landa, J. T. and Ghiselin, M.T. 1999. The emerging discipline of bioeconomics: aims and scope of the journal of bioeconomics *J. Bioecon*. 1, 5–12.

Landreth, A. and Bickle, J. 2008. Neuroeconomics, neurophysiology and the common currency hypohtesis. *Economics and Philosophy*, 24, p.419–429.

Latsis, S.J. 1976. A research programme in economics. In S. Latsis, ed., *Method and appraisal in economics*. Cambridge: Cambridge University Press, p.1-41.

Laudan, L. 1981. A Confutation of Convergent Realism. *Philosophy of Science*, 48, 1, p.19-49.

Laudan, L. 1984. *Science and Values*. Berkeley: University of California Press.

Laudan, L. 1990. Demystifying Underdetermination. In W.Savage, ed., *Scientific Theories. Minnesota Studies in the Philosophy of Science*, p.267-97. Minneapolis: University of Minnesota Press.

Le Bihan, D. 2007. The 'wet mind': water and functional neuroimaging, *Physics in Medicine and Biology*, 52(7), R57-R90.

LeDoux, J.E. 1996. *The Emotional Brain: The Mysterious Underpinnings of Emotional Life*. New York: Simon and Schuster.

Levi, I. 1986. The Paradoxes of Allais and Ellsberg. *Economics and Philosophy*, 2, p.23-53.

Levins, R. 1966. The strategy of model building in population biology. *American Scientist*, 54: 421-431.

Levins, R. 1968. *Evolution in changing environments: some theoretical explorations*. Princeton: Princeton University Press.

Libet, B. 1965. Cortical activation in conscious and unconscious experience. *Perspect. Biol. Med*. 9, p.77-86.

Libet, B. 1983. Time of conscious intention to act in relation to onset of cerebral activity. *Brain*, 106: 23-42.

Libet, B. 1985. Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action. *Behavior and Brain Sciences*, 8(4): 529–66.

Libet, B. 1996. Commentary on free will in the light of neuropsychiatry. In *Philosophy, Psychiatry & Psychology*, 3.

Libet, B., Wright, E., Feinstein, B. and Pearl, D. 1979. Subjective referral of the timing for a conscious experience: A functional role for the somatosensory specific projection system in man. *Brain*, 102, p.191-222.

Lichtenstein, S. and Slovic, P. 1971. Reversals of Preference Between Bids and Choices in Gambling Situations. *Journal of Experimental Psychology*, 89, p.46-55.

Lieberman, M.D., Berkman, E.T. and Wager, T.D. 2009. Correlations in Social Neuroscience Aren't Voodoo. Commentary on Vul et al. (2009). *Perspectives on Psychological Science*, 4, p.299–307.

Liepert, J., H. Bauder, W. H. R. Miltner, E. Taub and C. Weiller. 2000. Treatment-induced massive cortical reorganization after stroke in humans. *Stroke*, 31, p.1210-16.

Lloyd, D. 2002. Studying the Mind from the Inside Out. *Brain and Mind*, 3, p.243-259.

Loewenstein, G. 1996. Out of control: visceral influences on behaviour. *Organizational behaviour and human decision processes*, 65(3): 272-92.

Loewenstein, G. and Haisley, E. 2008. The Economist as Therapist: Methodological Ramifications of 'Light' Paternalism. In A. Caplin and A. Schotter (Eds.), *Perspectives on the Future of Economics: Positive and Normative Foundations*. Available at: http://papers.ssrn.com/Sol3/papers.cfm?abstract_id=962472.

Loewenstein, G. and O'Donoghue, T. 2004. *Animal spirits: affective and deliberative processes in economic behavior*. Working Paper, Carnegie Mellon University.

Loewenstein, G., Rick, S. and Cohen, J.D. 2008. Neuroeconomics. *Annual Review of Psychology*, 59, p.647-72.

Logothetis, N.K. 2002. The neural basis of the blood-oxygen-level-dependent functional magnetic resonance imaging signal. *Philosophical Transactions: Biological Sciences*, 357, p.1003-1037.

Logothetis, N.K. 2003. The underpinnings of the BOLD functional magnetic resonance imaging signal. *The Journal of Neuroscience*, 23(10), p.3963-71.

Logothetis, N.K. 2008a. What we can do and what we cannot do with fMRI. *Nature*, 453, p.869-878.

Logothetis, N.K. 2008b. What we can do and what we cannot do with fMRI. Supplementary information. *Nature*, 453, p.1-14.

Logothetis, N.K., Pauls, J., Augath, M., Trinath, T. and Oeltermann, A. 2001. Neurophysiological investigation of the basis of the fMRI signal. *Nature*, 412: 150-7.

Logothetis, N.K. and Pfeuffer, J. 2004. On the nature of the BOLD fMRI contrast mechanism. *Magnetic Resonance Imaging*, 22(10): 1517-31.

Logothetis, N.K. and Wandell, B.A. 2004. Interpreting the BOLD signal, *Annual Review of Physiology*, 66, p.735-769.

Loomes, G. and Sugden, R. 1982. Regret Theory: An Alternative Theory of Rational Choice under Uncertainty, *Econ. J*. 92, p.805-24.

Loomes, G. and Sugden, R. 1983. A Rationale for Preference Reversal, *American Economic Review*, 73, p.428-32.

Loomes, G. and Sugden, R. 1986. Disappointment and Dynamic Consistency in Choice under Uncertainty, *Review of Economic Studies.* 53:2, pp. 271-82.

Loomes, G. and Sugden, R. 1987. Some Implications of a More General Form of Regret Theory, *Journal of Economic Theory*, 41:2, p.270-87.

Loomes, G. and Sugden, R. 1995. Incorporating a Stochastic Element into Decision Theories. *European Economic Review*, 39, p.641-8.

Lucas, R.E. Jr. 1980. Methods and problems of business cycle theory. *Journal of Money, Credit and Banking*, 12, p.696-715.

Ludvig, N., Botero, J.M., Tang, H.M., Gohil, B. and Kral, J.G. 2001. Single-cell recording from the brain of freely moving monkeys. *Journal of Neuroscience Methods*, vol.106, no.2, p.179-187.

MacCrimmon, K. and Larsson, S. 1979. Utility Theory: Axioms versus Paradoxes. In *Expected Utility Hypotheses and the Allais Paradox*. M. Allais and O. Hagen, eds. Dordrecht: Reidel.

Machamer P., Darden, L. and Craver, C. 2000. Thinking about mechanisms. *Philosophy of Science*, 67, p.1-25.

Machina, M. 1982. Expected Utility Analysis without the Independence Axiom. *Econometrica*, 50, p.277-323.

Machina, M. 1987. Choice Under Uncertainty: Problems Solved and Unsolved. *The Journal of Economic Perspectives*, vol.1, no.1, p.121-154.

Machlup, F. 1955. The Problem of Verification in Economics. *Southern Economic Journal*, vol.22, p.1-21.

MacLean, P.D. 1990. *The Triune Brain in Evolution: Role in Paleocerebral Function*. New York: Plenum.

Mäki, U. 1988. On the problem of realism in Economics. *Fundamenta Scientiae*, 9, p.353-373.

Mäki, U. 1990. Friedman and Realism. *Research in the History of Economic Thought and Methodology*, 10, p.171-198.

Mäki, U. 1992. On the method of isolation in Economics. In Dilworth C. (ed) *Intelligibility in Science IV*, Rodophi, p.317-351.

Mäki, U. 1993. Two Philosophies of the Rhetoric of Economics. In Henderson, W., Dudley-Evans, T. and Blackhouse, R. (eds.). *Economics and Language*. Routledge: London.

Mäki, U. 1994. Reorienting the assumptions issue. In *New Directions in Economic Methodology*, Roger Backhouse Eds. London: Routledge, p.236-256.

Mäki, U. 1996. Two Portraits of Economics. *Journal of Economic Methodology*, 3 (1), p.1-38.

Mäki, U. 2001. Realisms and their opponents. In *International Encyclopedia of the Social and Behavioral Sciences*, Vol.19. Elsevier, Amsterdam, p.12815-12821.

Mäki, U. 2002. Some nonreasons for nonrealism about economics, in Mäki, U. (ed.) *Fact and Fiction in Economics: Models, Realism and Social Construction*, Cambridge, Cambridge University Press, p.90-104.

Mäki, U. 2005. Models are experiments, experiments are models. *Journal of Economic Methodology*, vol. 12, no.2, p.303-315.

Mäki, U. 2007. *Realism and Economic Methodology*. London: Routledge.

Mäki, U. 2009. MISSing the World. Models as Isolations and Credible Surrogate Systems. *Erkenntnis*, 70, p.29–43.

Mäki, U. 2010. When economics meets neuroscience: hype and hope. *Journal of Economic Methodology*. Vol. 17, No. 2, p.107-117.

Mameli, M. 2001. Modules and Mindreaders. *Biology and Philosophy*, 16, p.377-393.

Markowitz, H. 1952. The Utility of Wealth. *Journal of Political Economy*, 60, 151-158.

Marr, D. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York: W.H. Freeman & Company.

Marshall, A. 1890. *Principles of Economics*. Book III. London: Macmillan and Co., Ltd. 1961.

Massey, D.S. 2002. A brief history of human society: the origin and the role of emotion in social Life. *American Sociological Review*, 67: 1-29.

Matthewson, J. and Calcott, B. 2011. Mechanistic models of population-level phenomena. *Biology and Philosophy*, 26, p.737–756.

Matthewson, J. and Weisberg, M. 2009. The structure of tradeoffs in model building. *Synthese*, 170, p.169-90.

May, R.M. 2001. *Stability and Complexity In Model Ecosystems*. Princeton: Princeton University Press.

McCabe, K. 2003a. Neuroeconomics Explained. Center for the Study of Neuroeconomics. Posted on September 15, 2003 at http://neuroeconomics.typepad.com/neuroeconomics/2003/09/neuroeconomics_.html

McCabe, K. 2003b. Neuroeconomics. *Encyclopedia of Cognitive Science*, Nature Publishing Group, Macmillan Publishing, New York: 294-8.

McCabe, K. 2008. Neuroeconomics and the Economic Sciences. *Economics and Philosophy*, 24, p.345-368.

McCabe, K., Houser, D., Ryan, L., Smith, V. and Trouard, T. 2001. A Functional Imaging Study of 'Theory of Mind' in Two-Person Reciprocal Exchange. *Proceedings of the National Academy of Sciences,* 98: 11832-5.

McCauley, R.N. 1986. Intertheoretic Relations and the Future of Psychology. *Philosophy of Science*, 53, p.179-199.

McCauley, R.N. 1996. Explanatory Pluralism and the Coevolution of Theories in Science. In *Philosophy and the Neurosciences*. 2001. W. Bechtel, P. Mandik, J. Mundale, and R. Stufflebeam (eds.), Oxford: Blackwell Publishers.

McCauley, R.N. 2007. Reduction: Models of Cross-Scientific Relations and Their Implications for the Psychology-Neuroscience Interface. In *Handbook of the Philosophy of Science: Philosophy of Psychology and Cognitive Science*. P. Thagard (ed.). Amsterdam: Elsevier, 105-158.

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

McCloskey, D.N. 1983. The rhetoric of economics. *Journal of Economic Literature*, 21, p.481-517.

McCloskey, D.N. 1990. Storytelling in economics. In D. Lavoie (ed.). *Economics and Hermeneutics*, London: Routledge, p.61–75.

McClure, S.M., Ericson, K.M., Laibson, D.I., Loewenstein, G. and Cohen, J.D. 2007. Time Discounting for Primary Rewards. *Journal of Neuroscience*, 27(21): 5796-5804.

McClure, S.M., Laibson, D., Lowenstein, G. and Cohen, J. 2004a. Separate Neural Systems Value Immediate and Delayed Monetary Rewards. *Science*, 306: 503-7.

McClure, S.M., York, M.K. and Montague, P.R. 2004b. The Neural Substrates of Reward Processing in Humans: The Modern Role of fMRI. *Neuroscientist*, 10 (3): 260-68.

McCormick, D.A., Shu, Y.S. and Hasenstaub, A. 2003. Balanced recurrent excitation and inhibition in local cortical networks. *Excitatory-Inhibitory Balance: Synapses, Circuits, Systems*. Hensch, T. (Ed). Kluver Academic Press, New York.

McMullin, E. 1985. Galilean Idealization. *Studies in History and Philosophy of Science*, XVI, p.247-73.

McMullin, E. 1993. Rationality and paradigm change in science. In P. Horwich (Ed.), *World Changes: Thomas Kuhn and the Nature of Science*, p.55-78. Cambridge: The MIT Press.

McMullin, E. 2005. *The Church and Galileo*. McMullin, E. (eds.). University of Notre Dame Press.

Medler, D.A., Dawson, M.R.W. and Kingstone, A. 2005. Functional localization and double dissociations: the relationship between internal structure and behavior. *Brain and Cognition*, 57, p.146-150.

Mendelson, J. 1967. Lateral hypothalamic stimulation in satiated rats: the rewarding effects of selfinduced drinking. *Science*, 157: 1077-9.

Metzinger, T. 2000. *Neural Correlates of Consciousnes: Empirical and Conceptual Questions*. MIT Press.

Mill, J.S. 1843. *A System of Logic*. London: Longmans, Green & Co., 1949.

Mill, J.S. 1844. On the Definition and Method of Political Economy. In Hausman, D. (ed.). 1994. *The Philosophy of Economics*. Cambridge: Cambridge University Press.

Montague, P.R. 2007a. Neuroeconomics: a view from neuroscience. *Functional Neurology*, 22 (4), p.219-34.

Montague, P.R. 2007b. The First Wave. *Trends in Cognitive Sciences*, 11(10), p.407-409.

Montague, P.R. and Berns, G.S. 2002. Neural Economics and the Biological Substrates of Valuation. *Neuron*, 36(2): 265–84.

Montague, P.R., Berns, G.S., Cohen, J.D., McClure, S.M., Pagnoni, G. 2002. Hyperscanning: simultaneous fMRI during linked social interactions. *Neuroimage*, 16: 1159-64.

Montague, P.R. and Lohrenz, T. 2007. To detect and correct: Norm violations and their enforcement. *Neuron*, 56, p.14-18.

Morgan, M.S. 2001. Models, stories and the economic world. *Journal of Economic Methodology*, 8:3, p.361–384.

Morgan, M.S. 2002. Models, stories, and the economic world. In Maki, U. (ed.) *Fact and Fiction in Economics. Models, Realism, and Social Construction*. Cambridge: Cambridge University Press, p.178–201.

Morris, G., Nevet, A., Arkadir, D., Vaadia, E. and Bergman, H. 2006. Midbrain dopamine neurons encode decisions for future action. *Nature Neuroscience*, 9: 1057-63.

Morrison, M. 2000. *Unifying Scientific Theories*. Cambridge: Cambridge University Press.

Morrison, M. and Morgan, M.S. 1999. Models as mediating instruments. In *Models as Mediators*, M.S. Morgan, and M. Morrison, eds., Cambridge University Press, Cambridge, p.10-37.

Morton, J. 1984. Brain-based and non-brain based models of language. In D. Caplan, A. R. Lecours and A. Smith (Eds.), *Biological perspectives in language* (pp. 40–64). Cambridge, MA: MIT Press.

Moscati, I. 2006. Epistemic Virtues and Theory Choice in Economics. London School of Economics-Centre for Philosophy of Natural and Social Science, Discussion Paper 79/06.

Moscati, I. 2008. Review of *The Foundations of Positive and Normative Economics: A Handbook*, ed. Andrew Caplin and Andrew Schotter. Oxford University Press. *Economics and Philosophy*, 26 (1), p.101-108.

Mundale, J. 1998. Brain mapping and cognitive science. In Bechtel, W. and Graham, G. (eds.). *A Companion to Cognitive Science.* Oxford: Basil Blackwell.

Mundale, J. 2001. Neuroanatomical Foundations of Cognition: Connecting the Neuronal Level with the Study of Higher Brain Areas. In W.P. Bechtel, P. Mandik, J. Mundale & R.S. Stufflebeam (eds.), *Philosophy and the Neurosciences: A Reader*. Blackwell.

Mundale, J. and Bechtel, W. 1996. Integrating Neuroscience, Psychology, and Evolutionary Biology through a Teleological Conception of Function. *Minds and Machines*, 6, p.481-505.

Musgrave, A. 1981 Unreal Assumptions in Economic Theory: the F-twist untwisted. *Kyklos*, Vol.34, 3, p.377-387.

Nagel, E. 1961. *The structure of science: Problems in the logic of scientific explanation*. New York: Harcourt, Brace and World.

Nagel, E. 1963. Assumptions in Economic Theory. *The American Economic Review*, Vol.53, No.2, p.211-9.

Nagel, E. 1974. Issues in the Logic of Reductive Explanations. Reprinted in *Philosophy of Science*: *The Central Issues*. 1998. Curd, M. and Cover, J.A. (eds.), New York: W.V. Norton & Company, p.905-921.

Nagel, T. 1974. What is it like to be a bat? *Philosophical Review* 83, 4, p.435–50.

Nagel, T. 1986. *The View from Nowhere*. Oxford University Press.

Newman, S.D., Twieg, D.B. and Carpenter, P.A. 2001. Baseline conditions and subtractive logic in neuroimaging. *Human Brain Mapping*, Vol.14, 4, p.228-235.

Newton-Smith, W.H. 1978. The underdetermination of theory by data. *Aristotelian Society, Suppl*. Vol. 52, p.71-91.

Nickles, T. 2006. Scientific Revolutions. *The Philosophy of Science: An Encyclopedia*, vol. 2, S. Sarkar and J. Pfeifer (eds.), New York: Routledge, p.754-765.

Nobel Press Release. 2002. Psychological and experimental economics. Daniel Kahneman and Vernon Smith. http://nobelprize.org/nobel_prizes/economics/laureates/2002/press.html

Nozick, R. 1974. *Anarchy, State and Utopia*. New York: Basic Books.

O'Connor, T. 2009. Conscious Willing and the Emerging Sciences of Brain and Behavior. In G.F. Ellis, N. Murphy, and T. O'Connor (Eds.), *Downward Causation and the Neurobiology of Free Will*. New York: Springer Publications, p.173-186.

Odenbaugh, J. 2003. Complex Systems, Trade-Offs and Mathematical Modeling: A Response to Sober and Orzack. *Philosophy of Science* , LXX, p.1496–1507.

Odenbaugh, J. 2005. Idealized, inaccurate, and successful: a pragmatic approach to evaluating models in theoretical ecology. *Biology and Philosophy*, 20, p.231-255.

Odenbaugh, J. and Alexandrova, A. 2011. Buyer beware: robustness analyses in economics and biology. *Biology and Philosophy*, 26, p.757-771.

O'Doherty, J.P., Deichmann, R., Critchley, H.D. and Dolan, R.J. 2002. Neural responses during anticipation of a primary taste reward. *Neuron*, 33: 815–26.

Ogawa, S. and Lee, T.M. 1990. Magnetic resonance imaging of blood vessels at high fields: in vivo and in vitro measurements and image simulation. *Magnetic Resonance in Medicine*, 16(1), p.9-18.

Ogawa, S., Lee, T.M., Kay, A.R. and Tank, D.W. 1990. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Science USA*, 87, p.9868–9872.

Okasha, S. 2000. The underdetermination of theory by data and the 'strong programme' in the sociology of knowledge. *International Studies in the Philosophy of Science*, Vol.14, no.3, p.283-297.

Olds, J. and Milner, P. 1954. Positive reinforcement produced by electrical stimulation of septal area and other regions of rat brain. *Journal of Comparative and Physiological Psychology*, 47: 419–27.

Ollinger, J.M. 1994. Positron emission tomography: physical models and reconstruction issues. *Image Processing*, IEEE International Conference**,** 3, p.543-547.

Ollinger, J.M. and Fessler, J.A. 1997. Positron-emission tomography. *Signal Processing Magazine*, IEEE, 14 (1), p.43-55.

Oppenheim, P. and H. Putnam. 1958. The unity of science as a working hypothesis. In H. Feigl et al. (eds.), *Minnesota Studies in the Philosophy of Science*, vol. 2, Minneapolis: Minnesota University Press.

Ortmann, A. 2008. Prospecting Neuroeconomics. *Economics and Philosophy*, 24, p.431-448.

Padoa-Schioppa, C. 2008. The syllogism of neuro-economics. *Economics and Philosophy*, 24, p.449–457.

Padoa-Schioppa, C. and Assad, J.A. 2006. Neurons in the Orbitofrontal Cortex Encode Economic Value. *Nature*, 441(4): 223-26.

Padoa-Schioppa, C. and Assad, J.A. 2008. The Representation of Economic Value in the Orbitofrontal Cortex is Invariant for Changes of Menu. *Nature Neuroscience*, 11(12): 95-102.

Panksepp, J. 1998. *Affective Neuroscience*. Oxford: Oxford University Press.

Papineau, D. 1996. Theory-dependent Terms. *Philosophy of Science*, 63, p.1-20.

Papineau, D. 2002. *Thinking about Consciousness*. Oxford University Press.

Papineau, D. 2003. Could there be a science of consciousness? *Nous*, 13, p.205-220.

Pareto, V. 1909 [1971]. *Manual of Political Economy*. Edited by Ann S. Schwier and Alfred N. Page. Translated by Ann S. Schwier. New York: A.M. Kelley.

Parfit, D. 1982. Personal identity and rationality. *Synthese*, 53, p.227-41.

Parfit, D. 1984. *Reasons and Persons*. Oxford University Press.

Park, J.W. and Zak, P.J. 2007. Neuroeconomic Studies. *Analyse & Kritik*, 29, p.47-59.

Parker, G. and Smith, J.M. 1990. Optimality Theory in Evolutionary Biology. *Nature*, 348, p. 27-33.

Payzan, E. and Bourgeois-Gironde, S. 2005. Epistemological foundations for Neuroeconomics. August 2005 draft. Downloadable at http://en.scientificcommons.org/16624738

Pemberton, J. 2005. Why Idealized Models in Economics have Limited Use. In Jones, M.R. and Cartwright, N. (eds.). *Idealization XII: Correcting the Model. Idealization and Abstraction in the Sciences*, vol.86, p.35-46. New York: Rodopi.

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. and Frith, C. 2006. Dopamine-Dependent Prediction Errors Underpin Reward-Seeking Behavior in Humans. *Nature Online Letters*, p.1-4.

Petersen, S.E. and Fiez, J.A. 1993. The processing of single words studied with positron emission tomography. *Annual Review of Neuroscience, 16*, p.509-530.

Petersen, S.E., Fox, P.T., Posner, M.I., Mintun, M. and Raichle, M.E. 1989. Positron emission tomographic studies of the processing single words. *Journal of Cognitive Neuroscience, 1*(2), p.153-170.

Pigou, A.C. 1920 [2002]. *The Economics of Welfare London.* New Brunswick, NJ: Transactions Press.

Platt, M. L. and Glimcher, P. W. 1998. Neurons in area LIP carry information correlated with movement probability and reward magnitude. *Invest. Ophthalmol. Vis. Sci*. 39.

Platt, M.L. and Glimcher, P.W. 1999. Neural correlates of decision variables in parietal cortex. *Nature* 400(6741): 233-238.

Plott, C.R. 1996. Rational Individual Behavior in Markets and Social Choice Processes: The Discovered Preference Hypothesis. In *Rational Foundations of Economic Behavior*. K. Arrow, E. Colombatto, M. Perleman, and C. Schmidt, eds. London: Macmillan and NY: St. Martin's, pp. 225-50.

Poincaré, H. 1905. *Science and Hypothesis.* London: Walter Scott Publishing.

Poldrack, R.A. 2006. Can Cognitive Processes by Inferred from Neuroimaging Data? *Trends in Cognitive Science*, 10: 59-63.

Poldrack, R.A. and Wagner, A.D. 2004. What Can Neuroimaging Tell Us About the Mind? Insights from Prefontal Cortex. *Current Directions in Psychological Science*, 13(5): 177-81.

Popper, K.R. 1959. *The Logic of Scientific Discovery*. London: Hutchinson.

Popper, K.R. 1963. *Conjectures and Refutations.* New York: Harper and Row.

Prelec, D. and Bodner, R. 2003. Self-signaling and self-control. In G. Loewenstein, D. Read and R. Baumeister (eds.), *Time and Decision*, p.277-298. New York: Russell Sage Foundation.

Preuschoff, K., Bossaerts, P. and Quartz, S.R. 2006. Neural Differentiation of Expected Reward and Risk in Human Subcortical Structures, *Neuron*, 51: 381-90.

Price, C.J. and Friston, K.J. 2005. Functional ontologies for cognition: The systematic definition of structure and function. *Cognitive Neuropsychology*, no.3&4, p.262-275.

Psillos, S. 1999. *Scientific realism: How science tracks truth*. Routledge.

Putnam, H. 1975. Philosophy and our mental life. In *The Philosophy of Mind*, Beakley B. And Ludlow, P. Eds. 1992. ch.13, p.91-99. The MIT Press.

Quartz, S.R. 2008. From cognitive science to cognitive neuroscience to neuroeconomics. *Economics and Philosophy*, 24, p.459–471.

Quine, W.V.O. 1953. Two Dogmas of Empiricism. In *From a Logical Point of View*. Cambridge, MA: Harvard University Press, p.20-46.

Quine, W.V.O. 1975. On empirically equivalent systems of the world, *Erkenntnis*, 9, p.313-328.

Rabin, M. 1998. Psychology and Economics. *Journal of Economic Literature*, 36(1), p.11-46.

Rabin, M. 2002. A perspective on Psychology and Economics. *European Economic Review*, 46(4-5), p.657-685.

Raichle, M.E. 1998. Behind the scenes of functional brain imaging: A historical and physiological perspective. *Proceedings of the National Academy of Science USA*, vol. 95, p.765-772.

Ramsey, F. 1931. *The Foundations of Mathematics and Other Logical Essays*. New York: Harcourt Brace.

Ratzsch, D. 2009. Science and Religion. In Flint, T. and Rea, M. (Eds.), *The Oxford Handbook of Philosophical Theology*, Ch.3, p.54-77. NY: Oxford.

Raz, J. 1986. *The Morality of Freedom*, Oxford: Clarendon Press.

Read, D. 2007. Experienced utility: Utility theory from Jeremy Bentham to Daniel Kahneman. Working paper No: Lse OR 04-64. Later published in *Thinking & Reasoning*, 13(1): 45-61.

Redelmeier D.A., Katz, J. and Kahneman, D. 2003. Memories of colonoscopy: A randomized trial. *Pain, 104* , 187-194.

Redgrave, P. and Gurney, K.N. 2006. The Short-Latency Dopamine Signal: A Role in Discovering Novel Actions? *Nature Reviews Neuroscience,* 7, p.967-975.

Richardson, R.C. 2009. Multiple realization and methodological pluralism. *Synthese*, 167, p.473-492.

Rilling, J.K., Gutman, D.A., Zeh, T.R., Pagnoni, G., Berns, G.S. and Kilts, C.D. 2002. A neural basis for social cooperation. *Neuron*, 35: 395-405.

Rilling, J.K., Sanfey, A.G., Aronson, J.A., Nystrom, L.E. and Cohen, J.D. 2004a. Opposing BOLD responses to reciprocated and unreciprocated altruism in putative reward pathways. *Neuroreport*, 15: 2539-43.

Rilling, J.K., Sanfey, A.G., Aronson, J.A., Nystrom, L.E. and Cohen, J.D. 2004b. The neural correlates of theory of mind within interpersonal interactions. *Neuroimage*, 22:1694-1703.

Robertson, L.C., Knight, R.T., Rafed, R. and Shimamura, A.P. 1993. Cognitive neuropsychology is more than single case studies. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*, p.710-717.

Robbins, L. 1935. An Essay on the Nature and Significance of Economic Science, Ch. VI. Reproduced in *Philosophy of Economics*. Hausman, D. 1994.

Rockel, A.J., Hiorns, R.W. and Powell, T.P. 1980. The basic uniformity in structure of the neocortex. *Brain*, 103, p.221–244.

Rorty, R. 1965. Mind-Body Identity, Privacy, and Categories. *Review of Metaphysics*, 19 (1), 73, p.24-54.

Rosen, B.R., Buckner, R.L. and Dale, A.M. 1998. Event-related fMRI: past, present, and future. *Proceedings National Academy of Science, Usa*, 95: 773–80.

Rosenberg, A. 1992. *Economics. Mathematical Politics or Science of Diminishing Returns?* Chicago. University of Chicago Press.

Roskies, A.L. 2006. Neuroscientific Challenges to Free Will and Responsibility. *Trends in Cognitive Sciences*, 10 (9), p.419-423.

Roskies, A.L. 2007. Are Neuroimages like Photographs of the Brain? *Philosophy of Science*, 74, p.860–72.

Roskies, A.L. 2008. Neuroimaging and Inferential Distance. *Neuroethics*, 1, p.19-30.

Ross, D. 2002. Why people are atypical agents. *Philosophical Papers* 31: 87-116.

Ross, D. 2005. *Economic Theory and Cognitive Science: Microexplanation*. Cambridge, MA: MIT Press.

Ross, D. 2008a. Two Styles of Neuroeconomics. *Economics and Philosophy*, 24, p.473-483.

Ross, D. 2008b. Economics, Cognitive Science and Social Cognition. *Journal of Cognitive Systems Research*, 9: 125-135.

Ross, D. 2008c. Ontic Structural Realism and Economics. *Philosophy of Science*, 75 (5), p.732-743.

Ross, D. 2010. The Economics Agent: Not Human, but Important. In *Handbook of the Philosophy of Science*, Vol. 13. *Economics*, ed. U. Maki. Amsterdam: Elsevier, p.627-671.

Ross, D., Kincaid, H., Spurrett, D. and Collins, P. 2010. *What Is Addiction?* MIT Press.

Ross, D., Sharp, C., Vuchinich, R. and Spurrett, D. 2008. *Midbrain Mutiny: The Picoeconomics and Neuroeconomics of Disordered Gambling*. MIT Press.

Roughgarden, J. 1979. *Theory of Population Genetics and Evolutionary Ecology: An Introduction*. New York: Macmillan Publishing.

Rubinstein, A. 2007. Instinctive and cognitive reasoning: response times study. *Economic Journal*, 117, p.1243-59.

Rubinstein, A. 2008. Comments on Neuroeconomics. *Economics and Philosophy*, 24, p.485-494.

Rugg, M.D. and Coles, M.G.H. 1995. *Electrophysiology of mind*. Oxford: Oxford University Press.

Rustichini, A. 2003. Brain Experts Now Follow the Money. Interview by Sandra Blakeslee, *New York Times*, June 17.

Rustichini, A. 2005. Neuroeconomics: Present and Future. *Games and Economic Behavior*, 52: 201-212.

Rustichini, A. 2009. Is There a Method of Neuroeconomics? *American Economic Journal:Microeconomics*, 1(2), p.48-59.

Salmon, W.C. 1971. *Statistical Explanation and Satistical Relevance*. Pittsburgh: University of Pittsburgh Press.

Salmon, W.C. 1984. *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press.

Samuelson, P.A. 1938. A Note on the Pure Theory of Consumer's Behavior. *Economica*, New Series, 5(17): 61-71.

Samuelson, P.A. 1947. *Foundations of Economic Analysis*. Cambridge, MA: Harvard University Press.

Sanfey, A.G., Rilling, J.K., Aronson, J.A., Nystrom, L.E. and Cohen, J.D. 2003. The neural basis of economic decision-making in the Ultimatum Game. *Science*, 300, p.1755-58.

Sankey, H. 1993. Kuhn's Changing Concept of Incommensurability. *British Journal for the Philosophy of Science*, 44, p.759-774.

Sankey, H. 1994. *The Incommensurability Thesis*. Aldershot: Avebury Press.

Sankey, H. 1998. Taxonomic Incommensurability. *International Studies in Philosophy of Science*, 12, p.7-16.

Sartori, G. and Umiltà, C. 2000. How to Avoid the Fallacies of Cognitive Subtraction in Brain Imaging. *Brain and Language*, 74, p.191-212.

Sarver, T. 2008. Anticipating regret: Why fewer options may be better. *Econometrica*, 76, p.263-305.

Savage, L.J. 1954. *The Foundations of Statistics*. New York, NY: Wiley.

Savoy, R.L. 2001. History and Future Directions of Human Brain Mapping and Functional Imaging. *Acta Psychologica*, 107, p.9-42.

Saxe, R., Brett, M. and Kanwisher, N. 2006. Divide and conquer: A defense of functional localizers. *NeuroImage*, 30, p.1088-1096.

Saypol, J.M., Roth, B.J., Cohen, L.G. and Hallett, M. 1991. A Theoretical Comparison of Electric and Magnetic Stimulation of the Brain. *Annals of Biomedical Engineering*, vol. 19, p.317-328.

Schaffner, K.F. 1967. Approaches to Reduction. *Philosophy of Science*, 34, p.137-147.

Schaffner, K.F. 1976. Reductionism in biology: Prospects and problems. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 1974, p.613-632.

Schaffner, K.F. 1977. Reduction, reductionism, values, and progress in the biomedical sciences. In R. Colodny (Ed.), *Logic, laws, and life*, p.143-171. Pittsburgh: University of Pittsburgh Press.

Schelling, T. 1978. Egonomics, or the art of self-management. *American economic review*, 68(2), p.290-294.

Schelling, T. 1980. The intimate contest for self-command. *Public Interest*, 60, p.94-118.

Schneider, W. and Shiffrin, R.M. 1977. Controlled and Automatic Human Information Processing: I. Detection, Search and Attention. *Psychological Review*, 84(1): 1–66.

Schotter, A. 2008. What's so informative about choice? In *The Foundations of Positive and Normative Economics: A Handbook*, ed. Andrew Caplin and Andrew Schotter. Oxford University Press, Ch.3, p.70-94.

Schreiber, C. A., and D. Kahneman, 1996. Beyond the Peak and End Hypothesis: Exploring the Relation between Real-Time Displeasure and Retrospective Evaluation. Working Paper, Princeton University.

Schultz, W. 1998. Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, 80: 1-27

Schultz, W. 2000. Multiple Reward Signals in the Brain. *Nature Reviews Neuroscience*, 1: 189-207.

Schultz, W., Dayan, P. and Montague, P.R. 1997. A neural substrate of prediction and reward. *Science*, 275: 1593-99.

Schultz, W. and Dickinson, A. 2000. Neuronal coding of prediction errors. *Annual Reviews in Neuroscience*, 23: 473-500.

Schwarz, N. and Strack, F. 1991. Evaluating one's life: A judgment model of subjective well-being. In F. Strack, M. Argyle, & N. Schwarz (Eds.), *Subjective well-being: An interdisciplinary perspective*. Oxford: Pergamon Press.

Searle, J. 1980. Minds, Brains and Programs. *Behavioral and Brain Sciences*, 3 (3), p.417–457.

Searle, J. 1990. Is the Brain's Mind a Computer Program? *Scientific American*, 262 (1), p.26–31.

Sen, A. 1987. *On Ethics and Economics*. Oxford: Blackwells.

Shallice, T. 1979. Case study approach in neuropsychological research. *Journal of Clinical Neuropsychology*, 1, p.183-211.

Shallice, T. 1988. *From Neuropsychology to Mental Structure*, Cambridge University Press.

Shiffrin, R.M. and Schneider, W. 1977. Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychol. Rev.*, 84, p.127–90.

Shizgal, P. 1997. Neural basis of utility estimation. *Curr. Opin. Neurobiology*, 7: 198–208.

Shizgal, P. and Conover, K. 1996. On the neural computation of utility. *Current Directions in Psychological Science*, 5(2), 37-43.

Simon, H.A. 1955. A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*, Vol. 69, 1, p.99-118.

Simon, H.A. 1957. *Models of Man: Social and Rational*. New York: John Wiley & Sons, Inc.

Simonson, I. and Tversky, A. 1992. Choice in Context: Tradeoff Contrast and Extremeness Aversion. *Journal of Marketing Research*, 29(3): 281-95.

Singer, T., Seymour, B., O'Doherty, J.P., Stephan, K.E., Dolan, R.J. and Frith, C.D. 2006. Empathic neural responses are modulated by the perceived fairness of others. *Nature*, 439: 466-9.

Sklar, L. 1967. Types of inter-theoretic reduction. *British Journal for Philosophy of Science*, 18, p.109-124.

Smith, V.L. 1991. Rational Choice: The Contrast Between Economics and Psychology. *Journal of Political Economy*, 99, 4: 877-97.

Smith, V.L. 2007. *Rationality in Economics: Constructivist and Ecological Forms* (New York: Cambridge University Press.

Snell, J., Gibbs, B.J. and Varey, C. 1995. Intuitive Hedonics: Consumer Beliefs about the Dynamics of Liking. *Journal of Consumer Psychology*, IV, 33–60.

Spiegler, R. 2008. Comments on the potential significance of neuroeconomics for economic theory. *Economics and Philosophy*, 24, p.515–521.

Spitzer, M., Fischbacher, U., Herrnberger, B., Grön, G. and Fehr, E. 2007. The neural signature of social norm compliance. *Neuron*, 56, p.185-196.

Stanford, K. 2010. *Exceeding Our Grasp: Science, History, and the Problem of Unconceived Alternatives*. Oxford University Press.

Starmer, C. 2000. Developments in nonexpected utility theory: the hunt for a descriptive theory of choice under risk. *Journal of Economic Literature* 3, 38, p.332-82.

Sternberg, S. 1969. Discovery of Processing Stages: Extensions of Donders Method, *Acta Psychologica*, 30, 276-315.

Stigler, G. 1950. The Development of Utility Theory, I. *The Journal of Political Economy*, Vol. 58, No. 4, p.307-327.

Stone, A., Shiffman, S. and DeVries, M. 1999. Rethinking self-report assessment methodologies. In Kahneman, D., Diener, E., & Scharwz, N. (Eds.) *Well-being: The foundations of hedonic psychology* (p.26-39). New York, NY: Cambridge University Press.

Sugden, R. 1991. Rational Choice: A Survey of Contributions from Economics and Philosophy. *Econ. J*. 101, p.751-85.

Sugden, R. 1992. How People Choose. In *The Theory of Choice*. Ch. 3. Heap, Hollis, Lyons, Sugden and Weale. Ed. Blackwell.

Sugden, R. 2000. Credible worlds: the status of theoretical models in economics. *Journal of Economic Methodology*, 7(1): 1-31.

Sugden, R. 2004. The Opportunity Criterion: Consumer Sovereignty without the Assumption of Coherent Preferences. *American Economic Review*, 94: 1014-33.

Sugden, R. 2006. Taking unconsidered preferences seriously. In *Preferences and well-being*. Olsaretti, S. (eds.) Paperback: p.209-32.

Sugden, R. 2008. Why incoherent preferences do not justify paternalism. *Constitutional Political Economy*, 19 (3), p.226-248.

Sugrue L.P., Corrado, G.S., and Newsome, W.T. 2005. Choosing the greater of two goods: neural currencies for valuation and decision making. *Nature Review Neuroscience*, 6, p.363-375.

Sullivan, J.A. 2009. The multiplicity of experimental protocols: a challenge to reductionist and non-reductionist models of the unity of neuroscience. *Synthese*, 167, p.511-539.

Sunder, S. 2006. Economic Theory: Structural Abstraction or Behavioral Reduction? *History of Political Economy (Annual Supplement)*, Vol.38, p.322-342.

Sunstein, C.R. and Thaler, R.H. 2003. Libertarian paternalism is not an oxymoron. *The University of Chicago Law Review*, 70(4), p.1159-1202.

Suppe, F. 1977. *The Structure of Scientific Theories*. Urbana: University of Illinois Press.

Suppe, F. 1989. *The Semantic View of Theories and Scientific Realism*. Urbana: University of Illinois Press.

Suppes, P. 1960. A Comparison of the Meaning and Uses of Models in Mathematics and the Empirical Sciences. *Synthèse*, 12, p.287-301.

Suppes, P. 1967. What is a scientific theory? In Morgenbesser, S. (ed.), *Philosophy of Science Today*. New York: Basic Books, p.55-67.

TenHouten. W.D. 1991. Into the wild blue yonder: On the emergence of the ethnoneurologies - the social-science based neurologies and the philosophy-based neurologies. *Journal of Social and Biological Structures*, 14(4), p.381-408.

Ter-Pogossian, M.M., Raichle, M.E. and Sobel, B.E. 1980. Positron Emission Tomography. *Scientific American*, 243 (4), p.170-181.

Thaler, R.H. and Sunstein, C.R. 2008. *Nudge: improving decisions about health, wealth, and happiness*. Yale University Press.

Thirion, B., P. Pinel, S. Meriaux, A. Roche, S. Dehaene and J-B. Poline. 2007. Analysis of a large FMRI cohort: statistical and methodological issues for group analyses. *NeuroImage* 35: 105–20.

Toga, A.W. and Mazziotta, J.C., 2002. *Brain mapping: the methods*. 2nd ed. Academic Press.

Tomlin, D., Kayali, M.A., King-Casas, B., Anen, C., Camerer, C.F., Quartz, S.R. and Montague, P.R. 2006. Agent-specific responses in the cingulate cortex during economic exchanges. *Science*, 312, p.1047-50.

Tooby, J. and Cosmides, L. 1994. Better than Rational: Evolutionary Psychology and the Invisible Hand. *The American Economic Review*, 84(2): 327-32.

Tooby, J. and Cosmides, L. 2005. Evolutionary psychology: Conceptual foundations. *Handbook of Evolutionary Psychology*. Eds. Buss, D.

Toulmin, S. 1972. *Human Understanding*, vol.1, *The collective use and evolution of concepts*. Princeton: Princeton University Press.

Tullock, G. 1979. Sociobiology and economics. *Atlantic Economic Journal*, Vol.7, no.3, p.1-10.

Tversky, A. 1969. Intransitivity of Preferences, *Psychological Review*, 76, p.31-48.

Tversky, A. and Kahneman, D. 1974. Judgement Under Uncertainty: Heuristics and Biases. *Science*, 185 (4157), p.1124-31.

Tversky, A. and Kahneman, D. 1981. The Framing of Decisions and the Psychology of Choice *Science*, 211(4481): 453–8.

Tversky, A. and Kahneman, D. 1986. Rational Choice and the Framing of Decisions. *The Journal of Business*, 59(4, part 2): 251-78.

Tversky, A. and Kahneman, D. 1987. Rational Choice and the Framing of Decisions. In Hogarth and Reder eds. *Rational Choice: The Contrast between Economics and Psychology*. Chicago: Univ. Chicago Press.

Tversky, A. and Kahneman, D. 1992. Advances in Prospect Theory: Cumulative Representation of Uncertainty. *J. Risk Uncertainty* 5:4, pp. 297-323.

Tversky, A. and Thaler, R.H. 1990. Anomalies: Preference Reversals. *Journal of Economic Perspectives*, 4(2): 201-11.

Ubel, P.A., Loewenstein, G., Schwarz, N. and Smith, D. 2005. Misimagining the unimaginable: The disability paradox and healthcare decision making. *Health Psychology*, 24, S57-S62.

Uttal, W.R. 2001. *The New Phrenology: The Limits of Localizing Cognitive Processes in the Brain*, MIT Press.

Van Fraassen, B. 1980. *The Scientific Image*. Oxford: Oxford University Press.

Van Fraassen, B. 2000. The semantic approach to scientific theories. In *The Nature of Scientific Theory*, ed. Lawrence Sklar, Vol. 2 of *Philosophy of Science*, p.175-194. New York: Garland Publishing Inc.

Van Orden, G.C. and Paap, K.R. 1997. Functional neuroimages fail to discover pieces of mind in the parts of the brain. *Philosophy of Science*, 64, S85–S94.

Van Orden, G.C., Pennington, B.F and Stone, G.O. 2001. What do double dissociations prove? *Cognitive Science*. 25, 111-72.

Van‘t Wout, M., Kahn, R.S., Sanfey, A.G. and Aleman, A. 2005. Repetitive transcranial magnetic stimulation over the right dorsolateral prefrontal cortex affects strategic decision-making. *Neuroreport*, 16, p.1849-1852.

Van 't Wout, M., Kahn, R. S., Sanfey, A. G. and Aleman, A. 2006. Affective State and Decision-Making in the Ultimatum Game. *Experimental Brain Research*, 169(4), 564-568.

Vardi, Y., Shepp, L.A. and Kaufman, L. 1985. A Statistical Model for Positron Emission Tomography. *Journal of the American Statistical Association*, vol. 80, no.389, p.8-20.

Vercoe, M. and Zak, P.J. 2010. Inductive modeling using causal studies in neuroeconomics: brains on drugs. *Journal of Economic Methodology*, Vol. 17, No. 2, p.133-146.

Von Neumann, J. and Morgenstern, O. 1944. *Theory of Games and Economic Behavior*, Princeton University Press.

Vromen, J. 2007. Neuroeconomics as a Natural Extension of Bioeconomics: The Shifting Scope of Standard Economic Theory. *Journal of Bioeconomics*, 9, p.145–167.

Vromen, J. 2010a. On the surprising finding that expected utility is literally computed in the brain, *Journal of Economic Methodology*, Vol. 17, No. 1, p.17-36.

Vromen, J. 2010b. Where economics and neuroscience might meet. *Journal of Economic Methodology*, Vol. 17, No. 2, p.171-183.

Vul, E., Harris, C., Winkielman, P. and Pashler, H. 2009. Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition. *Perspectives on Psychological Science*, Vol.4, No.3, p.274-290.

Weber, M. 1904. Objectivity in Social Science and Social Policy. In E. Shils and H. Finch, eds. *The Methodology of the Social Sciences*. New York, Free Press, 1949, p.49-112.

Weintraub, E.R. 1979. *Microfoundations.* Cambridge: Cambridge University Press.

Weisberg, D.S., F.C. Keil, J. Goodstein, E. Rawson, and J.R. Gray. 2008. The seductive allure of neuroscience explanations. *Journal of Cognitive Neuroscience*, 20(3), p.470-477.

Weisberg, M. 2004. Qualitative theory and chemical explanation. *Philosophy of Science, 71*, p.1071–81.

Weisberg, M. 2006. Robustness analysis. *Philosophy of Science*, 73, p.730–742.

Weisberg, M. 2007a. Three Kinds of Idealization. *The Journal of Philosophy*, 104 (12), p.639-659.

Weisberg, M. 2007b. Who is a Modeler? *British Journal for Philosophy of Science*, 58, p.207–233.

Weslake, B. 2010. Explanatory Depth. *Philosophy of Science*, 77, p.273-294.

Wilcox, N.T. 2008. Against Simplicity and Cognitive Individualism. *Economics and Philosophy*, 24, p.523-532.

Wilkinson, D. and Halligan, P. 2004. The relevance of behavioural measures for functional-imaging studies of cognition. *Nature Reviews Neuroscience, 5*, p.67–73.

Wilson, E.O. 1998. *Consilience: the Unity of Knowledge*. Knopf, New York.

Wilson, T. and LaFleur, S. 1995. Knowing what you'll do: Effects of analyzing reasons on self-predictions. *Journal of Personality and social psychology*, 68(1): 21-35.

Wimsatt, W.C. 1976. Reductionism, levels of organization and the mind-body problem. In G. Globus, I. Savodnik and G. Maxwell (eds.), *Consciousness and the Brain*. New York: Plenum.

Wimsatt, W.C. 1987. False Models as a Means to Truer Theories. In M. Nitecki and A. Hoffmann (Eds.), *Neutral models in biology*. Oxford: Oxford University Press.

Woodward, J. 2003. *Making things happen*. New York: Oxford University Press.

Woodward, J. 2006. Some varieties of robustness. *Journal of Economic Methodoly*, 13(2), p.219–240.

Woodward, J. and Hitchcock, C. 2003a. Explanatory Generalizations, pt.1, 'A Counterfactual Account'. *Nous* 37 (1), p.1–24.

Woodward, J., and C. Hitchcock. 2003b. Explanatory Generalizations, pt.2, 'Plumbing Explanatory Depth'. *Nous* 37 (2), p.181-99.

Worrall, J. 1978. The ways in which the methodology of scientific research programmes improves on Popper's methodology. In G. Radnitzky and G. Anderson (eds.). *Progress and rationaliry in science*. Dordrecht (Holland), p.45-70.

Worrall, J. 1989. Structural Realism: The Best of Both Worlds? *Dialectica*, 43, 1-2, p.99-124.

Ylikoski, P. and Kuorikoski, J. 2010. Dissecting Explanatory Power. *Philosophical Studies*, 148, p.201-219.

Zahar, E.G. 1973. Why did Einstein's Programme Supersede Lorentz's. *British Journal for the Philosophy of Science*, 24, p.95-123, 223-62.

Zajonc, R.B. 1980. Feeling and Thinking: Preferences Need No Inferences. *American Psychologist*, 35(2): 151–75.

Zajonc, R.B. 1984. On the Primacy of Affect. *American Psychologist*, 39(2): 117–23.

Zajonc, R.B. and McIntosh, D.N. 1992. Emotions Research: Some Promising Questions and Some Questionable Promises. *Psychological Science*, 3(1): 70–74.

Zak, P.J. 2004. Neuroeconomics. In *Philosophical Transactions of the Royal Society Biology*, London, 359: 1737–48.

Zak, P.J. and Denzau, A. 2001. Economics is an evolutionary science. In *Evolutionary approaches in the behavioral sciences: toward a better understanding of human nature*. Somit&Peterson:31-65.

Zak, P.J. and Fakhar, A. 2006. Neuroactive Hormones and Interpersonal Trust: International Evidence. *Economics and Human Biology*, 4, p.412-429.

Zak, P.J., Kurzban, R. and Matzner, W.T. 2004 The neurobiology of trust. *Annuals of the New York Academy of Science*, 1032, p.224-227.

Zak, P.J., Kurzban, R. and Matzner, W.T. 2005. Oxytocin Is Associated with Human Trustworthiness. *Hormones and Behavior*, 48(5): 522-7.

Zak, P.J., Stanton, A. and Ahmadi, S. 2007. Oxytocin Increases Generosity in Humans. *PloS ONE*, 2(11): e1128, 1-5.