The Case for Mindless Economics[†]

Faruk Gul

and

Wolfgang Pesendorfer

Princeton University

November 2005

Abstract

Neuroeconomics proposes radical changes in the methods of economics. This essay discusses the proposed changes in methodology, together with the *the neuroeconomic critique* of standard economics. We do not assess the contributions or promise of neuroeconomic research. Rather, we offer a response to the neuroeconomic critique of standard economics.

[†] This research was supported by grants from the National Science Foundation. We thank Drew Fudenberg and Philipp Sadowski for helpful comments and suggestions.

1. Introduction

Neuroeconomics proposes radical changes in the methods of economics. This essay discusses the proposed changes in methodology, together with the *the neuroeconomic critique* of standard economics. Our definition of neuroeconomics includes research that makes no specific reference to neuroscience and is traditionally referred to as psychology and economics. We identify neuroeconomics as research that implicitly or explicitly makes either of the following two claims:

Assertion I: *Psychological and physiological evidence (such as descriptions of hedonic states and brain processes) are directly relevant to economic theories. In particular, they 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 ('choice utility').*

Neuroeconomics goes beyond the common practice of economists to use psychological insights as inspiration for economic modeling or to take into account experimental evidence that challenges behavioral assumptions of economic models. Neuroeconomics appeals directly to the neuroscience evidence to reject standard economic models or to question economic constructs. Camerer, Loewenstein and Prelec (2005) (henceforth CLP (2005)) express the neuroeconomics critique as follows:

> *"First, we show that neuroscience findings raise questions about the usefulness of some of the most common constructs that economists commonly use, such as risk aversion, time preference, and altruism." (p. 31-32)*

In section 5 of this essay, we argue that Assertion I of the neuroeconomic critique misunderstands economic methodology and underestimates the flexibility of standard models. Economics and psychology address different questions, utilize different abstractions, and address different types of empirical evidence. Neuroscience evidence cannot refute economic models because the latter make no assumptions and draw no conclusions about the physiology of the brain. Conversely, brain science cannot revolutionize economics because

the latter has no vehicle for addressing the concerns of economics. We also argue that the methods of standard economics are much more flexible than it is assumed in the neuroeconomics critique and illustrate this with examples of how standard economics deals with inconsistent preferences, mistakes, and biases.

Neuroeconomists import the questions and abstractions of psychology and re-interpret economic models as if their purpose were to address those questions. The standard economic model of choice is treated as a model of the brain and found to be inadequate. Either economics is treated as amateur brain science and rejected as such or brain evidence is treated as economic evidence to reject economic models.

Kahneman (1994) asserts that subjective states and hedonic utility are *"legitimate topics of study"*. This may be true, but such states and utilities are not useful for calibrating and testing standard economic models. Discussions of hedonic experiences play no role in standard economic analysis because economics makes no predictions about them and has no data to test such prediction. Economists also lack the means for integrating measurement of hedonic utility with standard economic data. Therefore, they have found it useful to confine themselves to the analysis of the latter.

The neuroeconomics program for change in economics ignores the fact that economists, even when dealing with questions related to those studied in psychology, have different objectives and address different empirical evidence. These fundamental differences are obscured by the tendency of neuroeconomists to describe both disciplines in very broad terms.

> *"Because psychology systematically explores human judgement, behavior, well-being it can teach us important facts about how humans differ from the way traditionally described by economics,"* (Rabin (1998)).

Note the presumption that across disciplines there is a single set of constructs (or facts) for describing how humans are. Rabin omits that economics and psychology study different kinds of behavior and, more importantly, focus on different variables that influence behavior. Realistic assumptions and useful abstractions when relating visceral cues to behavior may be less realistic or useful when relating behavior to market variables. Consider the following two statements:

"Much aversion to risks is driven by immediate fear responses, which are largely traceable to a small area of the brain called the amygdala;"(Camerer, Loewenstein and Prelec (2004), p. 567 (henceforth CLP (2004)).

"A decision-maker is (globally) risk averse, [...] if and only if his von Neumann-Morgenstern utility is concave at the relevant (all) wealth levels." Ingersoll (1987).

Which of these statements is (more) true? Which provides a better understanding of risk aversion? Most researchers recognize the various terms in the second statement as abstractions belonging to the specialized vocabulary of economics. Though less apparent, the language of the first statement is equally specialized in its use of discipline-specific abstractions. The terms 'immediate fear' and 'traceable' are abstractions of psychology and neuroscience. Moreover, the term 'risk aversion' represents *a different abstraction* in the two statements above. For Ingersoll, risk aversion is an attitude towards monetary gambles. For CLP (2004), risk aversion seems to be a much broader term that is readily applied to decisions involving plane travel. It makes little sense to insist that the economic notion of risk aversion is false while the psychological notion is true.

We discuss Assertion (II) of the neuroeconomic critique in section 6. We argue that the assertion misunderstands the role of welfare analysis in economics. Standard economics identifies welfare with choice, i.e., a change (in consumption) is defined to be welfare improving if and only if, given the opportunity, the individual would choose to make that change. The neuroeconomic critique of standard welfare analysis mistakes the economic definition of welfare for a theory of happiness and proceeds to find evidence against that theory. The standard definition of welfare is appropriate because standard economics has no therapeutic ambition; it does not try to improve the decision-maker but tries to evaluate how economic institutions mediate (perhaps psychologically unhealthy) behavior of agents.

Standard welfare economics functions as a part of positive economics. It provides a benchmark for the performance of economic institutions at aggregating individual preferences. Economists use welfare analysis to explain the persistence of some (efficient) institutions or to identify problems and anomalies in models of other (inefficient) institutions. For example, observing that an existing institution leads to Pareto efficient outcomes may increase the researcher's confidence in his model, while noting that the institution leads

to Pareto inefficiency may lead researchers to seek explanations for the persistence of that institution. Within this conception of welfare economics, what is relevant are the agents' interests (or preferences) as perceived by the agents themselves. An institution's effectiveness at maximizing the true happiness of its participants cannot justify the persistence of that institution if the criterion for true happiness conflicts with the participants' revealed preferences. After all, only the latter plays a role in behavior.

Neuroeconomists expect recent developments in psychology and brain science to yield answers to age-old philosophical questions such as *"what is happiness?"*; *"should we be willing to take actions contrary to a person's wishes if we happen to know that such actions will make them happier?"* and insist on a new notion of welfare based on these answers.

Perhaps a therapist or a medical professional is guided by his answers to the two questions above; he may fashion his advice to advance the perceived objectives of the patient or to increase the patient's true happiness, as defined by the therapist himself.[1] Neuroeconomic welfare analysis assumes a relationship between the economist and economic agents similar to the therapist-patient relationship. Normative economics is therefore identified with effective therapy. The economist/therapist can influence individuals' happiness by dispensing compelling advice or by influencing the decisions of powerful (and perhaps paternalistic) intermediaries. For example, Kahneman (1994) suggests that there is

> *"...a case in favour of some paternalistic interventions, when it is plausible that the state knows more about an individual's future tastes than the individual knows presently."*

Hence, the goal of welfare economics and perhaps the goal of all economics is to affect changes that result in greater happiness to all. In this endeavor neuroeconomists plan to enlist the support of the state – a stand-in for a benign therapist – who may, on occasion, conceal facts and make decisions on behalf of the individual's future selves.

Neuroeconomists seek a welfare criterion that is appropriate for an economist who is part social scientist and part advocate/therapist; someone who not only analyzes economic

[1] This description might over-state the therapist discretion. Either a professional code or market forces may limit the extent to which he can pursue the patient's true happiness. Hence, the two philosophical questions above may or may not have some relevance to the therapist. Our contention is that they have none for economists.

phenomena but also plays a role in shaping them. Neuroeconomists assert that the standard economic welfare criterion is not adequate for this task. Our response to this criticism is simple: the standard welfare criterion is not intended to facilitate advocacy for therapeutic interventions. The standard approach assumes a separation between the economist's role as social scientist and the role that some economists may play as advisors or advocates. This separation is valuable because it enables economists to analyze and compare different institutions without having to agree on the answers to difficult philosophical questions.

Besides the two assertions stated above, neuroeconomists pose an additional challenge to standard economics: they argue that economics should take advantage of recent improvements in neuroscience, in particular, improvements in measurements. They claim that these improvements may facilitate the unification of economics and brain science:

> *"This 'rational choice' approach has been enormously successful. But now advances in genetics and brain imaging (and other techniques) have made it possible to observe detailed processes in the brain better than ever before. Brain scanning (ongoing at the new Broad Imaging Center at Caltech) shows which parts of the brain are active when people make economic decisions. This means that we will eventually be able to replace the simple mathematical ideas that have been used in economics with more neurally-detailed descriptions."* Camerer (2005).

We discuss the unification argument in section 7. Our main point is that the separation of economics and brain science is a consequence of specialization around different questions and different data; it has little to do with technological limitations in measuring brain activity. Therefore, there is no reason to expect improvements in such technologies to lead to a unification.

In this essay, we do not assess the contributions or promise of neuroeconomic research. Instead, we offer a response to the neuroeconomic critique of standard economics. Our conclusion is that the neuroeconomic critique fails to refute any particular (standard) economic model and offers no challenge to standard economic methodology.

In the next section, we define the standard approach (or standard economics) and the neuroeconomics approach. In section 3, we discuss how the different goals of psychology and of economics necessitate different abstractions. As an example, we contrast the economic concepts of "complements" and "externalities" with the psychological concept of a

"cue." In section 4, we present an example of each approach to illustrate our classification and highlight the differences in the concerns and abstractions of standard economics and neuroeconomics. In sections 5, 6, and 7 we discuss the three main arguments of the neuroeconomics critique. Section 8 contains our closing remarks.

2. The Two Approaches: Definitions and Objectives

2.1 Standard Economics

The standard approach to behavioral economics extends standard choice theoretic methods to analyze variables that are often ignored. Some of these extensions are modest and entail little more than specifying a richer set of preferences over the same economic consequences. Others necessitate novel descriptions of the relevant economic outcomes. Yet, in most cases, the subsequent analysis is very similar to what can be found in a standard graduate textbook.

In the standard approach, the term *utility maximization* and *choice* are synonymous. A utility function is always an ordinal index that describes how the individual ranks various outcomes and how he behaves (chooses) given his constraints (available options). The relevant data are revealed preference data; that is, consumption choices given the individual's constraints. These data are used to calibrate the model (i.e., to identify the particular parameters) and the resulting calibrated models are used to predict future choices and perhaps equilibrium variables such as prices. Hence, standard (positive) theory identifies choice parameters from past behavior and relates these parameters to future behavior and equilibrium variables.

Standard economics focuses on revealed preference because economic data come in this form. Economic data can – at best – reveal what the agent wants (or has chosen) in a particular situation. Such data do not enable the economist to distinguish between what the agent intended to choose and what he ended up choosing; what he chose and what he ought to have chosen. The standard approach provides no methods for utilizing non-choice data to calibrate preference parameters. The individual's coefficient of risk aversion, for example, cannot be identified through a physiological examination; it can only be revealed through choice behavior. If an economist proposes a new theory based on

non-choice evidence then either the new theory leads to novel behavioral predictions, in which case it can be tested with revealed preference evidence, or it does not, in which case the modification is vacuous. In standard economics, the testable implications of a theory are its content; once they are identified, the non-choice evidence that motivated a novel theory becomes irrelevant.

As its welfare criterion, standard economics uses the individuals' choice behavior, that is, revealed preferences. Alternative x is deemed to be better than alternative y if and only if, given the opportunity, the individual would choose x over y.[2] Hence, welfare is *defined* to be synonymous with choice behavior.

In standard economics, an individual's decisions may improve when a constraint is relaxed. For example, an agent may make better decisions if he is given better information, more resources, or more time to make his decision. However, standard economics has no therapeutic ambition, i.e., it does not try to evaluate or improve the individual's objectives. Economics cannot distinguish between choices that maximize happiness, choices that reflect a sense of duty, or choices that are the response to some impulse. Moreover, standard economics takes no position on the question of which of those objectives the agent should pursue.

The purpose of economics is to analyze institutions, such as trading mechanisms and organization structures, and to ask how those institutions mediate the interests of different economic agents. This analysis is useful irrespective of the causes of individuals' preferences. Standard economics ignores the therapeutic potential of economic policies and leaves it to therapists, medical professionals, and financial advisors to help individuals refine their goals.

2.2 Neuroeconomics

"'This new approach, which I consider a revolution, should provide a theory of how people decide in economic and strategic situations,' said Dr. Aldo Rustichini, an

[2] The welfare statement is made relative to the constraints the agent faces. For example, the agent may be imperfectly informed of the consequences of his actions. In that case, the choice of x is welfare maximizing given the agent's information. If the agent had better information, he might choose y and hence y is the welfare maximizing choice for a better informed agent. See our discussion of mistakes in Section 5.1.

economics professor at the University of Minnesota. 'So far, the decision process has been for economists a black box.'"[3]

Later, in the same article, the author explains that

"In a study published in the current issue of the journal Science, Dr. Cohen and his colleagues, including Dr. Alan G. Sanfey of Princeton, took images of people's brains as they played the ultimatum game, a test of fairness between two people. In the ultimatum game, the first player is given, say, £10 in cash. He must then decide how much to give to a second player. It could be £5, the fairest offer, or a lesser amount depending on what he thinks he can get away with. If Player 2 accepts the offer, the money is shared accordingly. But if he rejects it, both players go away empty-handed. It is a one-shot game, and the players never meet again. Most people in the shoes of Player 2 refuse to take amounts under £2 or £3, Dr. Cohen said. They would rather punish the first player than feel cheated. 'But this makes no economic sense,' he said. 'You're better off with something than nothing.'"

As the quotes above illustrate, neuroeconomics emphasizes the physiological and psychological processes underlying decision-making. The objective is to relate the decision-making process to physiological processes in the brain or to descriptions of emotional experiences. From its predecessor, psychology and economics,[4] neuroeconomics inherits the idea of modeling the decision-maker as a collection of biases and heuristics susceptible to systematic errors (effects) and inconsistencies (reversals). Hedonic utilities (true utilities) are primitives, defined independently of behavior, while behavior is determined by biases and heuristics. The focus is on showing how factors that have no effect on these true utilities–or at least affect these utilities in a manner that is ignored by standard economics–influence behavior.

Neuroeconomics is therapeutic in its ambitions: it tries to improve an individual's objectives. The central questions of neuroeconomists are: How do individuals make their choices? How effective are they at making the choices that increase their own wellbeing? By contrast, economists analyze how the choices of different individuals interact within a particular institutional setting, given their differing objectives.

[3] "Brain Experts Now Follow the Money," by Sandra Blakeslee, New York Times, June 17, 2003.

[4] This line of inquiry is often referred to as behavioral economics. We have avoided using this term, in order to distinguish it from standard economics models that deal with similar behavioral issues.

3. Different Objectives Demand Different Abstractions

Neuroeconomists argue that the time is ripe for the methodology of economics to be brought in line with the methods and ideas of psychology and neuroscience. The neuroeconomic critique begins with the implicit or explicit assumption that economics, psychology and possibly other social sciences all address the same set of questions and differ only with respect to the answers they provide:

> *"More ambitiously, students are often bewildered that the models of human nature offered in different social sciences are so different, and often contradictory. Economists emphasize rationality; psychologists emphasize cognitive limits and sensitivity of choices to contexts; anthropologists emphasize acculturation; and sociologists emphasize norms and social constraint. An identical question on a final exam in each of the fields about trust, for example, would have different "correct" answers in each of the fields. It is possible that a biological basis for behavior in neuroscience, perhaps combined with all-purpose tools like learning models or game theory, could provide some unification across the social sciences (cf. Gintis, 2003)."* CLP (2004) p. 572-3.

Contrary to the view expressed in the quoted paragraph, economics and psychology do not offer competing, all-purpose models of human nature. Nor do they offer all-purpose tools. Rather, each discipline uses specialized abstractions that have proven useful for that discipline. Not only is the word trust much less likely to come up in an economics exam than in a psychology exam, but when it does appear in an economics exam, it means something different and is associated with a different question, not just a different answer. Far from being an all-purpose tool, game theory is a formalism for stripping away all strategically irrelevant details of the context, details that Gintis describes as central for psychologists. Similarly, a learning model in economics is different than a learning model in psychology. For an economist, a theory of learning might be a process of Bayesian inference in a multi-armed bandit model. This theory of learning is useful for addressing economic phenomena such as patent races but may be inappropriate for cognitive psychologists.

Once the goals of economics and psychology are stated in a manner that makes it seem as if the two disciplines address the same questions and deal with the same empirical

evidence, it becomes reasonable for neuroeconomists to inquire which discipline has the better answers and the better tools for providing answers.

CLP assert that

> *"neuroscience findings raise questions about the usefulness of some of the most common constructs economists commonly use, such as risk aversion, time preference, and altruism,"*

Risk aversion and time preference are indispensable concepts for modern economics. The authors really intend to question the validity of these concepts; in essence, they are asserting that there is no such thing as risk aversion or time preference. 'Time preference' and 'risk aversion' are useful economic abstractions just as 'cue-conditioned cognitive process' or 'hedonic forecasting mechanisms' are abstractions useful in neuroscience and psychology. The truth (or falsehood) of an abstraction cannot be evaluated independently; the only way to assess these abstractions by assessing – within each discipline – the theories that use them.

Consider the reverse procedure of using evidence from economics in brain science. Suppose that we find that drug addicts generally satisfy the strong axiom of revealed preference in their demand behavior. Can we argue that since addicts maximize some utility function, there are no separate brain functions and conclude then that the "limbic system" does not exist? This line of reasoning is, of course, absurd because brain science takes no position on whether choices satisfy the strong axiom of revealed preference or not. The argument that evidence from brain science can falsify economic theories is equally absurd. Hsu and Camerer write,

> *"For neuroeconomists, knowing more about functional specialization, and how regions collaborate in different tasks, could substitute familiar distinctions between categories of economic behavior (sometimes established arbitrarily by suggestions which become modeling conventions) with new ones grounded in neural detail. For example, the insula activity noted by Sanfey et al. in bargaining is also present when subjects choose between gambles with ambiguous odds of winning, relative to 'risky' gambles with known odds (Ming Hsu and Camerer, 2004)."*

Economists who are not interested in the physiological mechanism behind economic decisions will not find the level of insula activity useful for classifying behavior. What Hsu

and Camerer consider "distinctions based on arbitrary modeling conventions" are likely to be much more useful to economists, given their own objectives and given the type of data that is available to them.

The presumption that economics and psychology have the same goals and rely on the same data facilitates three types of critiques of standard economics:

1. *Failure of Rationality:* Economic models of choice fail to take account of psychological or physiological phenomena or evidence.
2. *Inadequacy of Rationality:* Rationality – defined to mean some sort of consistency in the behavior and preferences of individuals – is not an adequate starting point for economics because consistency of behaviors does not mean that these behaviors will lead to good outcomes.
3. *Unification:* Recent advances in neuroscience provide rich new sources of data. Economics must take advantage of these developments.

We address these arguments in sections 5, 6, and 7 respectively. We illustrate in the remainder of this section how the different goals psychology and economics and the different data available to these two disciplines necessitate different abstractions.

3.1 A Cue or a Complement?

The concept of a "cue" offers a good illustration of how abstractions from psychology are inappropriate for economics and, conversely, how the corresponding economic concepts are inappropriate for psychology and neuroscience. Psychologists call a stimulus that triggers a desire or a craving for a particular consumption or activity a "cue" or a "cue-elicited craving."[5] For example, eating a hamburger may be a cue that triggers a craving for French fries. Drinking coffee may trigger a craving for cigarettes. Visiting the location of previous drug consumption may trigger a craving for drugs. As the example of drug consumption illustrates, cues may be determined endogenously through a process of conditioning.[6] Psychologists find the concept of a cue useful because they think of cues as

[5] See Laibson (2001) for an economic model that describes psychological cues.

[6] The agent frequently consumed the drug at a particular location and - as a result of this consumption history - being in that location triggers a craving for drugs. Similarly, the agent frequently smoked a cigarette while drinking coffee in the past. This - perhaps incidental - pairing of consumption goods in the past implies that coffee consumption triggers a craving for cigarettes.

exogenous variables in experimental settings. They investigate the physiological mechanisms behind the development of and the reaction to cues. For economists, the notion of a cue is not useful because it lumps together two distinct economic phenomena: *complements* and *externalities*.

Hamburgers and fries are complementary goods just like forks and knives. Forks do not generate a craving for knives and therefore psychologists would not consider the fork/knife complementarity to be *the same phenomenon* as the hamburger/fries complementarity. For economists the physiological distinction between the two examples is unimportant. What matters is that demand for those goods responds in a similar way to price changes.

Another form of complementarities is the one associated with non-separable preferences over consumption streams. For example, consider an individual who enjoys building matchstick models and, as a result of this hobby, develops a complementary demand for matches and glue. The complementary demand for matches and glue is acquired through learning a hobby while the complementary demand for coffee and cigarettes is acquired through a process of conditioning. For a psychologist, who is interested in the underlying causes of preferences, the coffee/cigarette and glue/matchsticks complementarities represent distinct phenomena. The first is an example of conditioning while the second is an example of learning. However, both examples are similar in terms of the variables that economists observe and care about (prices, demand).

In the cue-response pairs above, the individual controls both the cue and the response. However, some cues are not under the control of the individual. For example, a former drug addict may experience a craving for drugs as he observes drug dealers in his neighborhood. In economics, this effect is captured by the notion of an externality. For economists, the neighborhood effect on drug addicts is similar to the effect of an improved network of roads on car buyers. Both are examples of an externality that causes a shift in the demand for a good. For psychologists, the craving for drugs by seeing drug-dealers in the neighborhood is similar to the craving for cigarettes caused by drinking coffee. On the other hand, they would consider it absurd to describe the car buying example and drug addiction example as being the same phenomenon because the underlying psychological mechanisms are very different. It would be equally absurd to insist that economists treat the neighborhood

effect on drug demand as the same phenomenon as the cigarette/coffee complementarity. In economics, there are important reasons for distinguishing between complementarities and externalities. For example, externalities often suggest market failures while complementarities do not.

Economists and psychologists use different abstractions because they are interested in different phenomena and must confront different data. 'Cue-triggered responses' is not a useful abstraction in economics because it lumps together distinct economic phenomena. Conversely, the economic abstraction of a complement is not useful in psychology because it lumps together phenomena with different psychological mechanisms.

4. The Two Approaches: Examples

In this section, we illustrate the standard approach to novel behavioral phenomena with a discussion of the paper "Temporal Resolution of Uncertainty and Dynamic Choice Theory," by Kreps and Porteus (1978). We illustrate the neuroeconomics approach with a recent paper by Köszegi and Rabin (2005) entitled "Reference-Dependent Utility."

4.1 The Standard Approach: Resolution of Uncertainty

An individual goes to the hospital on Friday to have a biopsy of a suspicious mass. In case the biopsy detects cancer, surgery will be scheduled for the following Monday. When given a choice between waiting a few hours to learn the result or going home and learning the result on Monday, the individual chooses to wait. The decision to incur the cost of waiting seems plausible but is inconsistent with standard theory. Standard expected utility maximizers are indifferent to the timing of resolution of uncertainty.

In "Temporal Resolution of Uncertainty and Dynamic Choice" Kreps and Porteus (1978) (henceforth Kreps-Porteus) expand the standard model of decision making under uncertainty to include anxious individuals such as the patient in the example above.[7] Suppose there are two dates $t = 1, 2$ and a finite set of prizes Z that will be consumed at date 2 ("surgery" or "no surgery" in the example above). Standard decision theory under uncertainty defines lotteries over Z as the choice objects. But this description does not

[7] The relationship between anxiety and preference for early or late resolution of uncertainty is explored and further developed in the work of Caplin and Leahy (2001).

differentiate between lotteries that resolve at date 1 and lotteries that resolve at date 2 - and therefore cannot capture the anxious patient described above.

Let D_2 be the lotteries over Z and let D_1 be lotteries over D_2. Hence, D_1 is the set of lotteries over lotteries over Z. We refer to elements of D_1 as date-1 lotteries and elements of D_2 as date-2 lotteries. We can describe the problem of the anxious patient as a choice between two lotteries in D_1. Suppose the probability of surgery is α. Waiting for the results until Monday corresponds to a date-1 lottery where, with probability 1, the individual will face the date 2 lottery that yields surgery with probability α and no surgery with probability $1-\alpha$. Learning the result on Friday corresponds to the date-1 lottery where, with probability α, the individual faces a date-2 lottery that yields surgery with probability 1 and, with probability $1-\alpha$, the individual faces a date-2 lottery that yields surgery with probability 0.

Let p, q denote elements in D_2 and μ, ν denote elements in D_1. For simplicity, we only consider lotteries with finite supports. Let $\mu(p)$ be the probability that μ chooses the lottery $p \in D_2$. Standard expected utility theory identifies μ with the implied probability distribution over prizes, i.e., the probability distribution $q \in D_2$ that assigns probability

$$q(z) = \sum_{D_2} \mu(p)p(z) \qquad (*)$$

to prize $z \in Z$. Therefore, standard expected utility theory cannot accommodate the cancer patient's strict preference for learning the test results on Friday.

The Kreps-Porteus model takes as a primitive an individual's preferences $\succeq$ (choices) over the date-1 lotteries, D_1. Some date-1 lotteries yield a particular date-2 lottery with probability 1. We call such lotteries degenerate date-1 lotteries. In the example above, learning the test results on Monday corresponds to such a lottery. Restricting the preference $\succeq$ to degenerate date-1 lotteries, *induces* a preference on D_2, the date-2 lotteries. Let δ_p denote the date-1 lottery that yields the date-2 lottery p with probability 1. The induced preference $\succeq_2$ (on D_2) is defined as follows:

$$p \succeq_2 q \text{ if and only if } \delta_p \succeq \delta_q$$

Kreps-Porteus assume that $\succeq$ and $\succeq_2$ satisfy the standard von Neumann-Morgenstern axioms: hence, the preferences are complete, transitive, satisfy the independence axiom, and satisfy an appropriate continuity assumption. Kreps-Porteus show that the preferences on D_1 satisfy those assumptions if and only if there are utility functions u and W such that $\mu \succeq \nu$ if and only if

$$\sum_{D_2} W\left(\sum_{z\in Z} u(z)p(z)\right)\mu(p) \geq \sum_{D_2} W\left(\sum_{z\in Z} u(z)p(z)\right)\nu(p)$$

The formula above applies the standard expected utility formula twice. The term in brackets is the expected utility formula for lotteries that resolve at date 2 whereas the outer term is the expected utility formula for lotteries that resolve at date 1.

The Kreps-Porteus formalism yields a precise definition of a new phenomenon: *preference for early (or late) resolution of uncertainty.* Let μ, ν be two elements of D_1 that imply the same distribution over prizes. The lottery μ resolves all uncertainty at date 1 while the lottery ν resolves all uncertainty at date 2. In the example above, μ corresponds to the situation where the patient learns the test result on Friday and ν corresponds to the situation where the patient learns the test result on Monday. The individual has a preference for early resolution of uncertainty if he prefers μ over ν. Kreps-Porteus show that a preference for early resolution of uncertainty implies (and is implied by) the convexity of W.

Note the key steps in the modeling exercise: Kreps-Porteus start with a novel psychological phenomenon and identify the *economically relevant consequences* of that phenomenon. Once the economically meaningful consequences are identified, the psychological causes become irrelevant. For the patient above, the source of the preference for early resolution of uncertainty is anxiety. But there could be many other reasons for a preference for early resolution of uncertainty. Suppose, for example, the agent owns a lottery ticket that will either yield a large reward (with small probability) or nothing. Prior to the lottery drawing, the agent must decide which car to purchase. The outcome of the lottery will typically affect the optimal car buying decision and, therefore, the agent would be better off if the lottery drawing was held earlier. Hence, the induced preferences over lotteries imply a preference for early resolution of uncertainty. In this case, the agent has

perfectly standard preferences. The preference for early resolution of uncertainty comes about because the agent has a second payoff-relevant decision to make after choosing a lottery.

In the two examples, the causes of the decision-maker's preference for early resolution of uncertainty are different. In the first example the patient is trying to avoid anxiety while in the second decision problem he is trying to make a better informed decision. For a standard economist this distinction is irrelevant because standard economics does not study the causes of preferences. For standard theory, the only relevant distinctions between the two examples are the ones that can be identified through the decision-makers' preferences.[8]

The Kreps-Porteus theorem identifies a formula that resembles standard expected utility applied separately at each decision date. While the formula is suggestive of a mental process, this suggestiveness is an expositional device not meant to be taken literally.[9] The formula encapsulates the behavioral assumptions of the theory in a user-friendly way and thereby facilitates applications of the theory to (more complicated) economic problems.

The theory is successful if preference for early resolution of uncertainty turns out to be an empirically important phenomenon; that is, if models that incorporate it are successful at addressing economic behavior. The role of the axioms is to summarize the empirical content of the theory independently of the specific application. The generality of the representation theorem, the usefulness of the key parameters, the ease with which the parameters can be measured and, most importantly, the empirical success of the model at dealing with economic evidence determine the extent to which the theory succeeds.

Kreps-Porteus's model has been generalized and applied to Macroeconomics and Finance (see Epstein and Zin (1991a, 1991b)). These fields analyze dynamic consumption choice under uncertainty. The primitives of Kreps-Porteus's model (dated lotteries) are

[8] For example, the Kreps-Porteus independence axiom may not be appropriate in the case where the agent has a second decision to make whereas the anxious patient might very well satisfy it.

[9] A teacher in an intermediate micro class might say something like, "the consumer equates the marginal utility of consuming the good to the marginal utility of the last dollar spent on the good," while explaining a first order condition in a partial equilibrium model with separable preferences. This statement is meant to provide some intuition for the first order condition, not as a description of the consumer's mental process: the marginal utilities in question depend on the particular utility function used to represent the preference and hence are, to some extent, arbitrary. There is no presumption that either these particular marginal utilities or the underlying calculus arguments are the actual currency of the consumer's reasoning.

easily adapted to match closely the objects studied in Macroeconomics and Finance. The fact that Kreps-Porteus strip all economically irrelevant details from their model is essential for the success of this adaptation.

4.2 Neuroeconomics: Reference Dependent Utility

In a well-known experiment (Thaler 1980)), a random subset of the subjects are assigned one unit of some object and then all subjects' reservation prices for this object are elicited. The price at which subjects who were assigned a unit are willing to sell it typically exceeds the price at which the remaining subjects are willing to buy a unit. This phenomenon is referred to as the endowment effect and has motivated models that add a *reference point* to the utility function.

Köszegi and Rabin (2005) (henceforth Köszegi-Rabin) propose a novel reference-dependent preference theory. To understand the Köszegi-Rabin theory, consider a finite set of choice objects X.[10] A reference-dependent utility function U, associates a utility with each reference point $z \in X$ and each choice object $x \in X$. Hence, $U : X \times X \to I\!R$, where $U(x, z)$ is the utility of x given the reference z. This formulation of utility is not new; the novelty is in the adoption of Köszegi (2004)'s notion of a personal equilibrium to determine the reference point. In this setting, a personal equilibrium for an decision-maker facing the choice set A is any $x \in A$ such that

$$U(x, x) \geq U(y, x) \tag{2}$$

for all $y \in A$. Hence, Köszegi-Rabin define the reference point as the x that ultimately gets chosen. It follows that an alternative $x \in A$ is optimal (i.e., a possible choice) for a Köszegi-Rabin decision-maker if (and only if) condition (2) above is satisfied. Köszegi-Rabin assume that U has the form

$$U(x, y) = \sum_{k \in K} u_k(x) + \sum_{k \in K} \mu(u_k(x) - u_k(y)) \tag{3}$$

where μ is an increasing function with $\mu(0) = 0$ and K is some finite set indexing the relevant hedonic dimensions of consumption. Köszegi-Rabin note that these consumption dimensions "should be specified based on psychological principles."

[10] An element $x \in X$ may be uncertain (i.e., may be a lottery).

Köszegi-Rabin also require that

$$U(x,y) \geq U(y,y) \text{ implies } U(x,x) > U(y,x) \tag{4}$$

for all $x, y \in X$.

There are certain striking differences between the approaches of Kreps-Porteus and Köszegi-Rabin. In Kreps-Porteus, the formula is an "as if" statement and the assumed restrictions on choice behavior (axioms) are the content of the theory. In contrast, Köszegi-Rabin interpret the procedure associated with computing a personal equilibrium (i.e., finding x that satisfy equation (2)) as a description of the underlying psychological process. Köszegi-Rabin focus on psychological evidence supporting this procedure and the various assumptions on the function U.

To facilitate the comparison of the difference in the two approaches, we provide a revealed preference analysis of the Köszegi-Rabin model for the case of no uncertainty.[11] Let X be finite and let Y be the set of all nonempty subsets of X. A function $c : Y \to Y$ is a choice function if $c(A) \subset A$ for all $A \in Y$. In revealed preference terms, the Köszegi-Rabin model is an investigation of a special class of choice functions. Given any state dependent utility function U, define $C(\cdot, U)$ as follows:

$$C(A,U) = \{x \in A \,|\, U(x,x) \geq U(y,x) \forall y \in A\}$$

A choice function c is *a general Köszegi-Rabin choice function* if there exists a reference dependent utility function U such that $c = C(\cdot, U)$. If the U also satisfies (3) and (4) then c is a *a special Köszegi-Rabin choice function*. For any binary relation $\succeq$, define the function $C_{\succeq}$ as follows:

$$C_{\succeq}(A) = \{x \in A \,|\, x \succeq z \forall z \in A\}$$

It is easy to construct examples where $C_{\succeq}(A) = \emptyset$ unless certain assumptions are made on $\succeq$. We say that the choice function c is induced by the binary relation $\succeq$, if $c(A) = C_{\succeq}(A)$ for all $A \in Y$. It is well-known that $C_{\succeq}$ is a choice function whenever $\succeq$ is complete ($x \succeq y$

[11] Köszegi-Rabin emphasize applications to decision making under uncertainty. Since we limit our analysis to a setting without uncertainty, our revealed preference "version" only captures the Köszegi-Rabin model for a limited set of applications.

or $y \succeq x$ for all $x, y \in X$) and transitive ($x \succeq y$ and $y \succeq z$ implies $x \succeq z$ for all $x, y, z \in X$). However, transitivity is not necessary for $C_{\succeq}$ to be a choice function. The proposition characterizes Köszegi-Rabin choice functions:

Proposition: *The following three conditions are equivalent:*

(i) *c is a general Köszegi-Rabin choice function*

(ii) *c is a choice function induced by some complete binary relation*

(iii) *c is a special Köszegi-Rabin choice function*

Proof: See Appendix

Note that $c = C_{\succeq}$ is a choice function implies $\succeq$ is complete. Hence, we may omit the word complete in the above proposition. The equivalence of (i) and (ii) establishes that abandoning transitivity is the only revealed preference implication of the Köszegi-Rabin theory. The equivalence of (ii) and (iii) implies that the particular functional form (3) and condition (4) are without loss of generality.

The revealed preference analysis answers the following question: suppose the modeler could not determine the individual ingredients that go into the representation, how can he check whether or not the decision-maker behaves in a manner consistent with such a representation? Or to put it differently, how is the behavior of a Köszegi-Rabin-decision maker different from a standard decision-maker? For the case of deterministic choice, the answer is that the Köszegi-Rabin decision-maker may fail transitivity.

In contrast, Köszegi-Rabin treat the relevant dimension of hedonic utility and the values of the various options along these dimensions as observable and quantifiable. They emphasize that this quantification requires craft and an understanding of psychological principles.

> *"Several aspects of our theory, however, render it short of fully general and formulaically applicable. Many of our specific assumptions are based on intuition rather than direct evidence."* (p. 31).

The assumptions of many theoretical models are based on intuition rather than direct evidence. But in standard models, any future test of the assumptions and the underlying

intuitions requires direct (revealed-preference) evidence. Where Köszegi-Rabin differ from standard economics is that psychological principles and (non-choice) evidence is viewed as an alternative form of evidence and it is this type of evidence that is the focus of their attention.[12]

In Köszegi-Rabin, utility indices (u_k's) and attachment disutilities (measured by μ) are hedonic utilities and are distinct from choice utilities. The Köszegi-Rabin representation is not only a theory of choice but also a description of the underlying psychological process:

> *"By all intuition and evidence, the feeling of loss when giving up a mug is a real hedonic experience, and making choices reflecting that real hedonic experience is partly rational. But as interpreted by Kahneman (2001) and Loewenstein, O'Donoghue, and Rabin (2003), people seem to over-attend to this experience because they ignore that the sensation of loss will pass very quickly – behaving as if they would spend much time longing for the mug they once had."*

Hence, measured feelings are inputs in the Köszegi-Rabin analysis. The authors believe that these measurements will enable the analyst to identify hedonic utilities that capture the intrinsic satisfaction of consuming the good (i.e., the u_k's) and hedonic utilities that capture the real loss associated with giving up the good. Moreover, they expect hedonic measurements to distinguish behavior that results from rational assessment of utilities from behavior that results from over-attending to utilities.

Köszegi-Rabin plan to calibrate the model using psychological insights and evidence. They view the Kreps-Porteus-type insistence on calibrating through revealed preferences as an unnecessary demand for *"formulaic applicability."* The model's success is judged by the extent to which the psychological process suggested by their formula matches psychological evidence.

[12] "In Köszegi and Rabin (2004), the previous version of this paper, we argue at length (as we do briefly in the conclusion of this paper) that the consumption dimensions used in our framework should be specified based on psychological principles, and not necessarily correspond directly to quantities of different products."

5. The Failure of Rationality

Neuroeconomists share with many other critics of economics the view that individual rationality is an empirically invalid assumption. Over the years, critics of rationality have identified various economic assumptions as 'rationality.' The independence axiom, probabilistic sophistication, monotonicity of payoffs in the agent's own consumption, or the independence of payoffs from the consumption of others have all been viewed as implications of rationality before the emergence of economic models that relax these assumptions.

More recent criticisms of rationality focus on the fact that individuals make systematic mistakes even in situations where the right choice is clear. The most ambitious critics of rationality argue that the idea of utility maximization is flawed because individuals do not maximize any preference relation. In section 5.2 we argue that these criticisms typically underestimate the flexibility revealed preference methodology. In particular, we illustrate how standard economics deals with 'mistakes.' In section 5.1, we focus on the evidence reported by neuroeconomists in support of their criticism. We observe that much of this evidence misses its target because economic models make no predictions about physiological processes that underly decision making.

5.1 The Neuroeconomic Case Against Preference Maximization:

CLP (2004) offer a short-list of neuroeconomic evidence against the "standard economic concept of preference." The list begins with the following item:

> *"Feelings of pleasure and pain originate in homeostatic mechanisms that detect departures from a "set-point" or ideal level, and attempt to restore equilibrium. In some cases, these attempts do not require additional voluntary actions, e.g., when monitors for body temperature trigger sweating to cool you off and shivering to warm you up. In other cases, the homeostatic processes operate by changing momentary preferences, a process called "alliesthesia" (Cabanac, 1979). When the core body temperature falls below the 98.6F set-point, almost anything that raises body temperature (such as placing one's hand in warm water) feels good, and the opposite is true when body temperature is too high. Similarly, monitors for blood sugar levels, intestinal distention and many other variables trigger hunger. Homeostasis means preferences are "state-dependent"*

in a special way: The states are internal to the body and both affect preferences and act as information signals which provoke equilibration...." (CLP (2004), p. 562)

No observation in the above cited paragraph contradicts any principle of preference maximization. Economic models make no predictions or assumptions about body temperature, blood sugar levels, or other physiological data and therefore such data cannot refute economic models. Standard economics is not committed to a particular theory of what makes people feel good. Nor does it assume that feeling good is what people care about.

The second item challenges the adequacy of revealed preference data:

"Inferring preferences from a choice does not tell us everything we need to know, and may tell us very little. Consider the hypothetical case of two people, Al and Naucia, who both refuse to buy peanuts at a reasonable price (cf. Romer, 2000). The refusal to buy reveals a common disutility for peanuts. But Al turned down the peanuts because he is allergic: consuming peanuts causes a prickly rash, shortens his breath, and could even be fatal. Naucia turned down the peanuts because she ate a huge bag of peanuts at a circus years ago, and subsequently got sick from eating too much candy at the same time. Since then, her gustatory system associates peanuts with illness and she refuses them at reasonable prices. While Al and Naucia both revealed an identical disutility, a neurally-detailed account tells us more. Al has an inelastic demand for peanuts-you can't pay him enough to eat them!-while Naucia would try a fistful for the right price. ....." (CLP (2004), p. 563)

It is often impossible to infer preferences from a single decision. In fact, finding a small class of such experiments to identify the individual's utility function is the central concern of revealed preference theory. Hence, not buying peanuts at a single price does not imply *"...Al and Naucia both revealed an identical disutility"* and while *"a neurally-detailed account"* could *"tell us more,"* the economically meaningful information can only be elicited with a change in prices. In standard economics, the reasons for a particular ranking of alternatives is irrelevant. That Al might die from consuming peanuts and Naucia simply doesn't like consuming them matters only if at some price Naucia is willing to do so and Al is not; and even then, it is the latter fact and not the underlying reasons that are relevant.

We delay the discussion of the third item to the next section where we discuss welfare analysis. The fourth item discusses what standard economics would consider a form of

money illusion: decision-makers may derive "direct" utility from money, beyond the utility they derive from the goods purchased with money.

> *"A fourth problem with preference is that people are assumed to value money for what it can purchase – that is, the utility of income is indirect, and should be derived from direct utilities for goods that will be purchased with money. But roughly speaking, it appears that similar brain circuitry – dopaminergic neurons in the midbrain – is active for a wide variety of rewarding experiences – drugs, food, attractive faces (cite), humor (cite) – and money rewards. This means money may be directly rewarding, and it's loss painful...."* (CLP (2004), p. 565.)

There are straightforward economic tests for identifying money illusion. Such a test would entail changing prices and nominal wages in a manner that leaves the set of feasible consumption, labor supply pairs unchanged. Then, we could check if this change has shifted the labor supply curve. But the issue cannot be addressed by investigating the brain circuitry and the midbrain, since economic models are silent on the brain activity associated with decision making.

The final item deals with addiction:

> *"Addiction is an important topic for economics because it seems to resist rational explanation. .... It is relevant to rational models of addiction that every substance to which humans may become biologically addicted is also potentially addictive for rats. Addictive substances appear therefore to be "hijacking" primitive reward circuitry in the "old" part of the human brain. Although this fact does not disprove the rational model (since the recently-evolved cortex may override rat-brain circuitry), it does show that rational intertemporal planning is not necessary to create the addictive phenomena of tolerance, craving, and withdrawal. It also highlights the need for economic models of the primitive reward circuitry, which would apply equally to man and rat. ....."* (CLP (2004) p. 565-566).

That substances addictive for rats are also addictive in humans is not relevant for economics because (standard) economics does not study rats.[13] It also does not study the causes of

[13] Presumably, psychologists interested in human physiology find it worthwhile to study rats because of the similarities in the neurological make-up of the two species. Apparently, the similarities between the economic institutions of the two species are not sufficient to generate interests in rats among economists.

preferences. To say that a decision-maker prefers x to y is to say that he never chooses y when x is also available, nothing more. Hence, addiction can be identified as a distinct economic phenomenon only through its distinct choice implications not through the underlying brain processes. The fact that addictive substances appear to be *"hijacking primitive reward circuitry,"* fails to disprove the rational model not because the cortex may override rat-brain circuitry but because the rational model addresses neither the brain-circuitry nor the cortex.

What the authors describe as evidence is in fact a statement of a their philosophical position. They have decided that the cortex represents planned action (rational choice), while certain processes in other parts (presumably in the midbrain) represent overwhelming physiological influences (i.e., the hijacking of the primitive reward circuitry).

> *"Many of the processes that occur in these systems are affective rather than cognitive; they are directly concerned with motivation. This might not matter for economics were it not for the principles that guide the affective system – the way that it operates – is so much at variance with the standard economics account of behavior."* (CLP (2005) p. 25-26).

Hence, every decision that is associated with the latter types of processes is interpreted as evidence that rational choice theory is wrong. This critique fails because standard economics takes no position on whether a particular decision represents a manifestation of free will or a succumbing to biological necessity. Rationality in economics is not tied to physiological causes of behavior and therefore the physiological mechanisms cannot shed light on whether a choice is rational or not in the sense economists use the term. Brain mechanisms by themselves cannot offer evidence against transitivity of preferences or any other choice-theoretic assumption. Therefore, evidence that utility maximization is not a good model of the brain cannot refute economic models.

Discussing decision making under uncertainty, Camerer (2005) writes:

> *"For example, when economists think about gambling they assume that people combine the chance of winning (probability) with an expectation of how they will value winning and losing ("utilities"). If this theory is correct, neuroeconomics will find two processes in the brain – one for guessing how likely one is to win and lose, and another for*

evaluating the hedonic pleasure and pain of winning and losing-and another brain region which combines probability and hedonic sensations. More likely, neuroeconomics will show that the desire or aversion to gamble is more complicated than that simple model."

Camerer assumes that there is one set of correct abstractions for both economics and neuroscience and tries to identify whether the ones currently used in economics belong to that set. The conceptual separation between probabilities and utilities is very important for expected utility *theory*. This separation need not have a physiological counterpart. Even if it did, mapping that process into the physiology of the brain and seeing if it amounts to *"one [process] for guessing how likely one is to win and lose, and another for evaluating the hedonic pleasure and pain of winning and losing-and another brain region which combines probability and hedonic sensations"* is a problem for neuroscience, not economics. Since expected utility theory makes predictions only about choice behavior, its validity can be assessed only through choice evidence. If economic evidence leads us to the conclusion that expected utility theory is appropriate in a particular set of applications, then the inability to match this theory to the physiology of the brain might be considered puzzling. But this puzzle is a concern for neuroscientists, not economists.

Standard economics does not address mental processes and, as a result, economic abstractions are typically not appropriate for describing them. In his (1998) survey, Rabin criticizes standard economics for failing to be a good model of the mind, even though standard economics never had such ambitions:

"Economists have traditionally assumed that, when faced with uncertainty, people correctly form their subjective probabilistic assessments according to the laws of probability. But researchers have documented many systematic departures from rationality in judgment under uncertainty."

Many economists (including the authors of many introductory economic textbooks) are aware that most people do not think in terms of probabilities, subjective or otherwise. Nor does standard economics assume that consumers know Bayes' law in the sense that a graduate student in economics would be expected to know it. Economic models connect to reality through economic variables, prices, quantities etc. and not through their modeling

of the individual's decision-making process. Evidence of the sort cited in neuroeconomics may inspire economists to write different models but it cannot reject economic models.

Our central argument is simple: neuroscience evidence cannot refute economic models because the latter make no assumptions or draw no conclusions about physiology of the brain. Conversely, brain science cannot revolutionize economics because it has no vehicle for addressing the concerns of the latter. Economics and psychology differ in the question they ask. Therefore, abstractions that are useful for one discipline will typically be not very useful for the other. The concepts of a preference, a choice function, demand function, GDP, utility, etc. have proven to be useful abstraction in economics. The fact that they are less useful for the analysis of the brain does not mean that they are bad abstractions in economics.

5.2 Mistakes

Individuals sometimes make obviously bad decisions. Neuroeconomists use this fact as proof of the failure revealed preference theory. Bernheim and Rangel (2005) provide the following example:

> *"(...) American visitors to the UK suffer numerous injuries and fatalities because they often look only to the left before stepping into streets, even though they know traffic approaches from the right. One cannot reasonably attribute this to the pleasure of looking left or to masochistic preferences. The pedestrian's objectives - to cross the street safely - are clear, and the decision is plainly a mistake."*

Standard economics has long recognized that there are situations where an outsider could improve an individual's decisions. Such situations come up routinely when agents are asymmetrically informed. Hence, standard economics deals with 'mistakes' by employing the tool of *information economics*.

Consider the following thought experiment. A prize ($100) is placed either in a red or in a blue box and the agent knows that there is a 60% chance that the money is in the red box. Confronted with a choice between the two boxes, the agent chooses the red box. An observer who has seen that the money was placed in the blue box may think that the agent prefers choosing red to getting $100. This inference is obviously incorrect because "choose $100" is a strategy that is not available to the agent. The observer who thinks

the agent prefers red to $100 has not understood the agent's constraints. Given agent's constraints, his choice of the red box is optimal.

Many situations in which agents systematically make mistakes can be interpreted as situations where agents face subjective constraints on the feasible strategies that are not apparent from the description of the decision problem. The strategy *"only cross the street when no car is approaching"* may be unavailable in the sense that it violates a subjective constraint on the set of feasible strategies. Hence, a standard economic model of the street-crossing problem would add a constraint on the set of feasible strategies as part of the description of the agent.

Suppose the economist asserts that the American tourist prefers not being run over by a car but finds it more difficult to implement that outcome in the UK than in the US. As evidence for this assertion the economist could point to data showing that American tourists in London avoid unregulated intersections. That tourists incur a cost to cross at regulated intersections suggests (i) they are unable to safely cross the street without help and (ii) they are not suicidal.

Framing effects can be addressed in a similar fashion. Experimenters can often manipulate the choices of individuals by restating the decision problem in a different (but equivalent) form. Standard theory interprets a framing effect as a change in the subjective constraints (or information) faced by the decision maker. It may turn out that a sign that alerts the American tourist to 'look right' alters the decision even though such a sign does not change the set of alternatives. The standard model can incorporate this effect by assuming that the sign changes the set of feasible strategies for the tourist and thereby alters the decision. With the help of the sign, the tourist may be able to implement the strategy *"always look right then go"* while without the sign this strategy may not be feasible for the tourist.

For standard economics, the fact that individuals make mistakes is relevant only if these mistakes can be identified through economic data. That behavior would have been different under a counter-factual scenario in which the agent did not make or was prevented from making these mistakes, is irrelevant.

6. The Inadequacy of Rationality

Neuroeconomists criticize both standard positive economics and standard normative analysis. In the previous section, we described and responded to the neuroeconomic critique of positive economics. Here, we address the neuroeconomic critique of normative economics.

Kahneman (1994) notes that *"[t]he term 'utility' can be anchored in the hedonic experience of outcomes, or in the preference or desire for that outcome."* Because agents make mistakes, neuroeconomists conclude that a person's choices do not maximize the hedonic consequences of these choices. More generally, neuroeconomists argue that choices do not maximizing the individual's well-being or happiness.

The neuroeconomic critique of standard welfare analysis relies on two related arguments: first, what people choose often fails to make them happy. Second, proper welfare analysis should be based on what makes people happy and such measurements necessitate neuroscientific input. Even if direct measurement of happiness through brain scans is not yet feasible, neuroeconomists believe that such measurement will eventually be possible.

> *"A third problem with preferences is that there are different types of utilities which do not always coincide.(...) For example, Berridge and Robinson (1998) have found distinct brain regions for "wanting" and "liking," which correspond roughly to choice utility and experienced utility. The fact that these areas are dissociated allows a wedge between those two kinds of utility... If the different types of utility are produced by different regions, they will not always match up. Examples are easy to find. Infants reveal a choice utility by putting dirt in their mouths, but they don't rationally anticipate liking it. Addicts often report drug craving (wanting) which leads to consumption (choosing) that they say is not particularly pleasurable (experiencing). Compulsive shoppers buy goods (revealing choice utility) which they never use (no experienced utility)(...)"* CLP (2004, p. 564).

Neuroeconomists use such evidence and related (thought) experiments to suggest that the concept of a preference that simultaneously determines behavior and "what is good for the agent" can be wide off the mark. Hence, neuroeconomists distinguish between *"decision*

utilities", which generate behavior, and *"experienced utilities"* which indicate what makes the agent happy.

In section 6.2, we discuss and respond to this neuroeconomic critique of standard welfare analysis. In sections 6.3 and 6.4, we consider two examples of substantive rationality in the literature: recent proposals for paternalism (section 6.3) and welfare analysis in multi-self models (section 6.4). First, we provide a brief summary of standard welfare analysis.

6.1 Standard Welfare Analysis

Economists use welfare analysis to examine how institutions mediate the interests of the participating individuals. Welfare improving changes to an economic institution are *defined* to be changes to which the individual(s) would agree. The policy x is deemed better than the policy y for an individual if and only if, given the opportunity, the individual would choose x over y. The choice of x over y may be motivated by the pursuit of happiness, a sense of duty or religious obligation, or reflect an impulse. In all cases, it constitutes an improvement of economic welfare.

Economic welfare analysis is a tool for analyzing economic institutions and models. For example, economic analysis of a trading institution may establish that the institution yields Pareto efficient outcomes and, therefore, there is no institutional change that will improve the economic welfare of all participants. Economists view such results as *successes* of their theories because the results demonstrate that the *economic model* of the institution is "stable"; there are no changes that are mutually agreeable to all participants. Conversely, models of economic institutions will raise suspicion if there are obvious welfare improving changes (changes that all individuals would agree to) because the availability of such changes suggests that the model misses important aspects of the underlying reality.

Economists use the revealed preference of individuals as a welfare criterion because it is the only criterion that can be integrated with positive economic analysis. For example, consider the economic analysis of farm subsidies. Economists have found that US farm subsidies are inefficient, i.e., farm subsidies could be eliminated and farmers could be compensated in a way that would increase the economic welfare of all US households. The most interesting aspect of this observation is that farms subsidies persist despite their

inefficiency. Motivated by this and related observations, economists have examined the mechanisms (political and economic) that lead to the persistence of inefficient policies.

The example of farm subsidies is typical for the use of welfare analysis in economics. Normative statements (farm subsidies are inefficient) are used to define new positive questions (what makes farm subsidies persist?) that lead to better models of the underlying institution. Economists use welfare analysis to identify the interests of economic agents and to ask whether existing policies can be interpreted as an expression of those interests or whether the understanding of the institutional constraints on policies remains incomplete. This use of welfare analysis *requires* the standard definition of economic welfare. There is no reason for economic agents to gravitate towards policies and institutions that yield higher welfare if the underlying notion of welfare does not reflect the interests of agents as the agents themselves perceive these interests.

6.2 Neuroeconomic Welfare Analysis

Neuroeconomists treat the economists definition of welfare *as if* it were a theory of happiness and proceed to find evidence against this theory. CLP write,

> *"Economics proceeds on the assumption that satisfying people's wants is a good thing. This assumption depends on knowing that people will like what they want. If likes and wants diverge, this would pose a fundamental challenge to standard welfare economics. Presumably welfare should be based on 'liking.' But if we cannot infer what people like from what they want and choose, then an alternative method for measuring liking is needed, while avoiding an oppressive paternalism."* (p. 36)

Welfare in economics is a *definition* and *not a theory* (of happiness). Therefore, the divergence of "liking and wanting" does not pose any challenge to the standard definition of welfare, no matter how the former is defined. Standard economics offers no substantive criterion for rationality because it has no therapeutic ambition; it does not attempt to cure decision-makers who make choices that do not generate the most pleasure. The more modest economic definition of welfare is mandated by the role of welfare analysis in economics.

To compare this role with the role envisaged by neuroeconomists, suppose that a trading institution is found to be (economically) inefficient. Typically, this will imply that

someone can set up an alternative institution and make a profit. Hence, we can expect this change to take place without a benevolent dictator, simply as a result of self-interested entrepreneurship. Suppose a psychologist argues that an inefficient trading institution leads to higher 'experienced' utility than an efficient one and agents are mistaken in their preference for the economically efficient institution. Whether or not this assertion is true, the economic analysis of the trading institution is valid. The economically efficient trading institution is still the one we can expect to prevail. Moreover, *since agents perceive their own interests to coincide with the economic welfare criterion* there is no obvious mechanism (economic or political) by which the psychologically superior institution could emerge.

Neuroeconomists would argue that even though a welfare criterion based on the individuals own "preferences or desires" may be relevant for positive analysis, a substantive criterion is needed for normative theory. For neuroeconomists, the goal of welfare analysis is to advocate changes that improve decision-maker's well-being. To achieve their goal, neuroeconomists can either try to convince people to want what is good for them (*therapy*) or make the right choice on their behalf (*paternalism*). Kahneman (1994) summarizes both these positions as follows:

> *"However, truly informed consent is only possible if patients have a reasonable conception of expected long-term developments in their hedonic responses,... A more controversial issues arises if we admit that an outsider can sometimes predict an individual's future utility far better than the individual can. Does this superior knowledge carry a warrant, or even a duty, for paternalistic intervention? It appears right for Ulysses' sailors to tie him to the mast against his will, if they believe that he is deluded about his ability to resist the fatal call of the sirens."*

The neuroeconomic view of welfare analysis builds on an inappropriate analogy between an economist and a therapist. It may be the case that sometimes outsiders know more about the future utility of an individual than the individual himself. But the goal of economics is not to prepare the economist for service at times when he finds himself in the role of that outsider.

If economists were in the business of investment counselling, it might make sense for neuroeconomists to focus on the conflict between what the typical consumer/investor wants

to do now and what will make him happy in the future. But economists do not deal with patients (or even clients). Therefore it is not clear who the recipient of their counselling would be. The neuroeconomic view of the economist as a therapist is inappropriate both as a description of what economists do, and as a description of what they could be doing.

Of course, one could argue that economists should identify a substantive criterion for rationality (i.e., a criterion for measuring what really makes individuals happy) and advocate changes that increase welfare according to this criterion regardless of whether or not they have the means to convince the potential beneficiaries to follow this advice. The hope being that someone other than the potential beneficiary might be convinced to implement the policies. This view is apparent in Kahneman's search for a benevolent paternalistic figure in his examples:

> *"...the physician could probably ensure that the patient will retain a more favourable memory of the procedure by adding to it a medically superfluous period of diminishing pain. Of course, the patient would probably reject the physician's offer to provide an improved memory at the cost of more actual pain. Should the physician go ahead anyway, on behalf of the patient's future remembering self?"* (Kahneman (1994))

In the same article, Kahneman suggests that there is

> *"a case in favour of some paternalistic interventions, when it is plausible that the state knows more about an individual's future tastes than the individual knows presently."*

When economists or political scientists model the government, they do so either by endowing the government with certain objectives or by modeling government as an institution where conflicting incentives of various agents interact. In Kahneman's analysis, the government is a benign and disinterested agent whose only role is to serve as the object of the modeler's lobbying efforts.

Welfare analysis for neuroeconomics is a form of social activism; it is a recommendation for someone to change his preferences or for someone in a position of authority to intervene on behalf of someone else. In contrast, welfare economics in the standard economic model is integrated with the model's positive analysis; it takes agents' preferences as given and evaluates the performance of economic institutions.

Regardless of one's views on the importance and efficacy of social activism, there are advantages to separating the role of the economist as a researcher from the role a particular

economist might play as an advocate. This separation enables the positive analysis to proceed without having to resolve difficult philosophical problems such as figuring out what makes people happy or who is more deserving of happiness. It also enables other researchers to assess and critique a particular piece of analysis without having to evaluate the merits of the underlying moral philosophy or the effectiveness of the researcher's activism.

6.3 Proposals for Paternalistic Welfare Criteria

Two recent articles outline plans for welfare economics based on paternalistic principles. In both papers, the authors are motivated by evidence showing that the specification of the default option affects individual choices of retirement plans. Rates of enrollment in 401(k) plans are significantly higher when the default option is to enroll than when the default option is not to enroll.

A standard interpretation of the 401(k) problem would argue that the default matters for the decision problem as perceived by the individual. The employee's set of feasible strategies changes with the default just as the feasible strategies of the American tourist in London change when a sign is placed at the side of the road alerting the tourist to look right. The welfare maximizing default option is the one that agents would choose when asked to choose among defaults.

Thaler and Sunstein (2003) (henceforth TS) seek paternalistic principles for choosing a default option. TS advocate *libertarian paternalism* and suggest the following three guiding principles:

> *"First, the libertarian paternalist might select the approach that majority would choose if explicit choices were required and revealed."*

Hence, the libertarian paternalist is to substitute the predicted preferences of the majority for the preferences of the individual.

> *"Second, the libertarian paternalist might select the approach that would force people to make their choices explicit."* Finally, *"the libertarian paternalist might select the approach that minimizes the number of opt-outs."*

TS offer no arguments for why their principles are likely to lead to greater happiness. In fact, they offer no defense of these principles. They simply say that the libertarian paternalist might choose to use them.

The fact that the TS principles are not particularly compelling as moral philosophy is a side issue. The real issue is that it is difficult to see what question their proposal addresses. To put it differently, it is unclear who they have in mind as the potential beneficiary of their philosophical argument. The TS motivation for paternalism seems to be that it is inevitable:

> *"The first misconception is that there are viable alternatives to paternalism. In many situations, some organization or agent must make a choice that will affect the choices of some other people."*

Clearly, the decisions of one agent may affect the utility of others. Economic analysis suggests that the interests of the agent in control are a good place to start when analyzing such situations. For example, in order to maximize profits, firms may wish to make their benefit plans as attractive as possible to their future employees. In that case, firms will choose plans (and their default options) in accordance with how the employees would choose them. It may be impractical to ask prospective employees about their preferred default option on the retirement plan and therefore the firm will use its best subjective assessment of the employees preferences.

Of course, the employer may have different objectives and may choose a plan that differs from his best guess of the employees preferred plan. Presumably, he would do so to increase his own welfare. In this situation, as in the situation of the pure profit maximizing employer, there is no role for the TS principles. The TS argument amounts to telling employers that when they face incomplete information they should adopt a different objective. Standard economics would predict that employers will take the best action given their own objectives and given what they know about the preferences of the employees.

In a recent paper, Camerer, Issacharoff, Loewenstein, O'Donoghue and Rabin (2003, henceforth CILOR) introduce and advocate the notion of "asymmetric paternalism:"

> *"A regulation is asymmetrically paternalistic if it creates large benefits for those who make errors, while imposing little or no harm on those who are fully rational."*

CILOR do not explain which preferences reflect bounded rationality and which reflect full rationality, when benefits are large and when there is little or no harm. Nevertheless, their implicit welfare criterion is familiar. As we described in section 5.2, the mistakes of

boundedly rational agents can be modeled as a subjective informational constraint facing these agents. With this re-interpretation, the CILOR principle amounts to an (epsilon) version of the Pareto principle: help the uninformed without hurting the informed (too much). However, there is an important difference between the CILOR version of the Pareto principle and the Pareto principle in standard economics: CILOR view their principle as a framework for activism. They urge their readers to adopt their modified libertarian philosophy in place of the purely libertarian philosophy that they perceive as guiding many economists (and perhaps some lawyers) or the unabashed paternalism favored by behavioral economists.

> *"Our paper seeks to engage two different audiences with two different sets of concerns: For those (particularly economists) prone to rigid antipaternalism, the paper describes a possibly attractive rationale for paternalism as well as a careful, cautious, and disciplined approach. For those prone to give unabashed support for paternalistic policies based on behavioral economics, this paper argues that more discipline is needed and proposes a possible criterion"* (CILOR p. 1212).

Of course, it is legitimate for TS and CILOR to engage employers or the legal and economics professions in a moral debate. But this has little to do with welfare economics which is not concerned with moral philosophy or with providing a disciplined guide for social action.

Standard economists spend little time or effort advocating normative criteria even when they feel that the right normative criterion is unambiguous. For example, many economists and decision-theorists believe in the importance of making decisions under uncertainty consistent with some subjective probability assessments. Moreover, hardly anyone would question the normative appeal of using Bayes' law when updating probabilities. There are many research papers where agents are endowed with subjective probabilities and use Bayes law. The purpose of these papers is not to advocate the use of subjective probabilities or Bayesian revision; rather, the normative appeal of the Savage model serves as a starting point for the positive analysis. The ultimate value of Savage's contribution depends not on the ability of his followers to convince individual economic agents or benign planners to adopt his view of probability but on the success his followers have at developing models that address economic data.

6.4 Preference Reversals and Multiselves

There is evidence that individuals resolve the same intertemporal trade-off differently depending on when the decision is made.[14] Researchers, starting with the work of Strotz (1955), have argued that this phenomenon requires modeling the individual as a collection of distinct selves with conflicting interests. Such models represent a major departure from standard economics conception of the individual as the unit of agency. For example, if the individual cannot be identified as a coherent set of interests, then the economists' welfare criterion is not well-defined. Hence, for neuroeconomists, preference reversals constitute an empirical validation of the psychologist's – as opposed to the economist's – view of the individual.

Consider the following example: in period 1, the agent chooses the consumption stream $(0,0,9)$ over $(1,0,0)$ and chooses $(1,0,0)$ over $(0,3,0)$. In period 2 the agent chooses $(0,3,0)$ over $(0,0,9)$. Suppose the agent faces the following decision problem: he can either choose $(1,0,0)$ in period 1 or leave the choice between $(0,0,9)$ and $(0,3,0)$ for period 2. Confronted with this choice, the agent picks $(1,0,0)$.

In Gul and Pesendorfer (2001), (2004) and (2005), we propose a standard, single-self model that accounts for this behavior. To illustrate the approach, define $\mathcal{C}$ to be the set of second period choice problems for the individual; that is, an element $C \in \mathcal{C}$ consists of consumption streams with identical first period consumption levels: $(c_1, c_2, c_3), (c_1', c_2', c_3') \in C \in \mathcal{C}$ implies $c_1 = c_1'$. In period 2, the individual chooses a consumption stream from some C. In period 1, the individual chooses a choice problem C for period 2. Choosing $(1,0,0)$ in period 1 corresponds to $\{(1,0,0)\}$ while the option of leaving it to period 2 to choose between $(0,3,0)$ and $(0,0,9)$ is described as

$$C = \{(0,3,0),(0,0,9)\}$$

With this notation, we can summarize the (period 1) behavior as

$$\{(0,0,9)\} \succ \{(1,0,0)\} \succ C = \{(0,3,0),(0,0,9)\} \sim \{(0,3,0)\}$$

[14] See Loewenstein, et al. for a recent survey of the experimental evidence. In the typical experiment, subjects choose between a smaller, date 2 reward and a larger, date 3 reward. If the choice is made at date 2, then the smaller-earlier reward is chosen. If the choice is made earlier (i.e., at date 1) then the larger-later reward is chosen. This phenomenon is sometimes referred to as dynamic inconsistency or a preference reversal.

Note that choosing between $\{(0,3,0)\}$ and $\{(0,0,9)\}$ is not the same as choosing from the set $C = \{(0,3,0),(0,0,9)\}$. In the former case, the consumer *commits* to a consumption path in period 1 while in the latter he chooses in period 2. The preference statements above indicate that the individual prefers a situation where he is committed to $(0,0,1)$ to a situation where he chooses from C in period 2. When such a commitment is unavailable, and the agent is confronted with C in period 2, he chooses $(0,3,0)$.

Standard economic models identify choice with welfare. Therefore, the choice of $(0,3,0)$ from C in period 2 is welfare maximizing as is the choice of $\{(0,0,9)\}$ over $\{(0,3,0)\}$ in period 1. The interpretation is that, in period 2, the agent struggles with the temptation to consume 3 units. Temptation is costly to resist and therefore consuming (rather than holding out for 9 in period 3) is the optimal (and welfare maximizing) choice in period 2. In period 1, higher period 2 consumption is not tempting and therefore the agent prefers $\{(0,0,9)\}$ over $\{(0,3,0)\}$. Period 1 behavior reveals that the individual's welfare is higher *in all periods* when he is committed to $(0,0,9)$ than when he must choose from C in period 2.[15]

The multi-self model abandons the revealed preference approach to welfare and constructs paternalistic welfare criteria. Consider again the three period model. In each period, the individual's preference are described by a utility function, U_t. For concreteness, assume:

$$\begin{aligned} U_1(c_1,c_2,c_3) &= c_1 + \beta\delta c_2 + \beta\delta^2 c_3 \\ U_2(c_1,c_2,c_3) &= c_2 + \beta\delta c_3 \\ U_3(c_1,c_2,c_3) &= \delta c_3 \end{aligned} \tag{9}$$

where $\delta = \beta = 1/2$. While different papers postulate different welfare criteria for such situations, the common argument is that preference reversals necessitate a criterion for trading-off the utility of the various selves. The most common practice in this literature is to treat the U_0 below as the welfare criterion.

$$U_0(c_1,c_2,c_3) = c_1 + \delta c_2 + \delta^2 c_3$$

[15] Note that choosing between $\{(0,0,9)\}$ and C is not a feasible option in period 2. Therefore, revealed preference experiments cannot uncover whether or not in period 2, the individual has a preference (or distaste) for commitment.

This particular welfare criterion may seem odd. After all, U_0 quite arbitrarily sets $\beta = 1$ and assigns a higher welfare to $(1, 0, 11)$ than to $(2, 3, 0)$ even though selves 1 and 2 prefer $(2, 3, 0)$. The multiple-selves literature interprets U_0 as the preferences with the"present bias" removed.[16] In other words, $\beta < 1$ is diagnosed as a defect and the role of policy intervention is to cure this defect.

Note that hyperbolic discounting (or time inconsistency) is not necessary for generating conflict among the various selves of the individual: Consider again the three period example above but now let $\beta = 1$. The resulting utility functions describe standard preferences with exponential discounting. Consider the two consumption streams: $(1, 0, 0)$ and $(0, 0, 4)$ and note that $U_1(1, 0, 0) = U_1(0, 0, 4)$ but $U_2(1, 0, 0) < U_2(0, 0, 4)$ and $U_3(1, 0, 0) < U_3(0, 0, 4)$; that is, the allocation $(1, 0, 0)$ is *Pareto dominated* by the allocation $(0, 0, 4)$ even though the usual welfare criterion of the multiselves literature (U_0) would deem the two alternatives welfare equivalent.[17]

Economists often note the arbitrariness of using U_0 as a welfare criterion in the multiselves model. It is not clear what hedonic utility calculations have led neuroeconomists to decide that U_0 represents the right trade-off among the hedonic utilities of the various selves. Our point is different: standard economics has neither need nor use for a welfare criterion that trades off utility among the various selves of a single individual. Such trade-offs can never play a role in explaining or understanding economic institutions. By definition, only behavior can influence economic data or institutions. Hence, beyond their effect on behavior the various 'selves' are irrelevant for the analysis. By contrast, neuroeconomists view the existence of multiple selves as both an opportunity and a rationale for activism. They wish to urge the individual to do a better job at accommodating the welfare of their future selves (i.e., resist $\beta < 1$ and other biases). Failing that, they would like to convince third parties to intervene on behalf of the agent's future selves. This therapeutic/paternalistic stance is similar to the position of medical professionals who attempt to cure a patient's addiction. By proposing a welfare criterion, the modeler is either urging

[16] See, for example, Rabin and O'Donoghue (2003).

[17] In standard analysis, this issue does not arise because the same utility function (U_1) is used to describe behavior (and welfare) at each decision date. In period 2, period 1 consumption cannot be altered and therefore the additively separable form of the utility function allows us to drop the first term *as a simplification* without affecting optimal choices.

the individual to reform his behavior, or urging someone in a position of authority to force the individual to do so.

Identifying what makes people happy, defining criteria for trading-off one person's (or selves) happiness against the happiness of another, and advocating social change in a manner that advances overall happiness by this criterion is a task many neuroeconomists find more worthy than dealing with the more pedestrian questions of standard economics. However, the expression of this preference constitutes neither an empirical nor a methodological criticism of standard economics.

7. The Unification of Economics and Neuroscience

Neuroeconomists often cite improvements in neuroscience, in particular, improvements in measurements, as a central reason for unifying the disciplines of economics, psychology, and brain science:

> *"Since feelings were meant to predict behavior, but could only be assessed from behavior, economists realized that without direct measurement, feelings were useless intervening constructs. In the 1940s, the concepts of ordinal utility and revealed preference eliminated the superfluous intermediate step of positing immeasurable feelings. Revealed preference theory simply equates unobserved preferences with observed choices. Circularity is avoided by assuming that people behave consistently, which makes the theory falsifiable; once they have revealed that they prefer A to B, people should not subsequently choose B over A..... The 'as if' approach made good sense, as long as the brain remained substantially a black box. The development of economics could not be held hostage to progress in other human sciences. But now neuroscience has proved Jevons' pessimistic prediction wrong; the study of the brain and nervous system is beginning to allow direct measurement of thoughts and feelings."* (CLP (2005), p. 10).

Thus, neuroeconomists view the revealed preference approach to be an outdated concession to technological limitations of the past.[18] Since the technology for distinguishing between "liking" (i.e., a criterion of substantive rationality) and "wanting" (i.e., choice) may soon

[18] For Kahneman, the rejection of hedonic utility as the basis for economic analysis of decisions has less to do with technology than the adherence to an outdated philosophy of science. Rabin (1996) seems to view a doctrinaire obstinacy as the only explanation for the persistence of economists' "habitual" assumptions.

be available, economics (and presumably other social sciences) should abandon the revealed preference methodology and adopt the methodology of psychology and neuroscience.

The dominant role of revealed preference analysis in economics has little to do with technology. Economic phenomena consist of individual choices and their aggregates and do not include hedonic values of utilities or feelings. Therefore, it is not relevant for an economic model to explore the feelings associated with economic choices. The point of revealed preference theory is to separate the theory of decision making from the analysis of emotional consequences of decisions. This separation is useful whether or not emotions can be measured simply because it facilitates specialization. Note that the more detailed and sophisticated the measurement the greater is the potential benefit of specialization.

Brain imaging data are of a radically different form than typical economic data. If the prediction of great advances in brain science turn out to be correct, they will certainly be accompanied by theoretical advances that address the particular data in that field. It is unreasonable to require those theories to be successful at addressing economic data as well. By the same token, the requirement that economic theories simultaneously account for economic data and brain imaging data places an unreasonable burden on economic theories.

Note that the above does not say that psychological factors are irrelevant for economic decision making, nor does it say that economists should ignore psychological insights. Economists routinely take their inspiration from psychological data or theories. However, economic models are evaluated by their success at explaining economic phenomena. Since hedonic utility values or brain imaging data are not economic phenomena, economists should not feel constrained to choose models that succeed as models of the brain.

The arguments advanced by neuroeconomists in favor of unification often fail to distinguish between a novel philosophical position and a scientific breakthrough. Often, what neuroeconomists present as an empirical challenge to economics is best viewed as an invitation to an ethical debate. For example, Kahneman (1994) writes:

> *"The history of an individual through time can be described as a succession of separate selves... Which one of these selves should be granted authority over outcomes in the future?"*

Hence, neuroeconomics interprets the individual as a flawed and inconsistent sequence of "pleasure machines," that need therapeutic and paternalistic assistance for assessing the right intertemporal trade-offs and making the right choices.

It is not clear what evidence neuroeconomics can offer to answer questions like "should physicians increase the actual pain experienced by the patient in order to facilitate his memory and improve his decision making for the future?" (Kahneman (1994)). What is clear is that finding out how to trade-off the welfare of one self against another or deciding "[w]hich one of these selves should be granted authority over outcomes in the future," is not an economic problem.

CLP (2004) suggest a more modest goal: that neuroscience may facilitate direct measurement of preference parameters by "asking the brain, not the person," (p. 573). The authors have no example of observing a choice parameter – such as the coefficient of relative risk aversion or the discount factor – through brain imaging, no suggestions as to how such inference could be done. They offer no criteria for distinguishing a brain where $\delta = .97$ versus one where $\delta = .7$. They do not explain what language to use when 'asking the brain, rather than the person,' which language the brain will use to respond, or what to do when the brain's answer conflicts with the answer of the person.

In the end, scientific developments play a small role in the arguments of neuroeconomists: when it comes to substantiating the central philosophical position that there is a difference between what people want and what is good for them, subjective readings of the facial expressions of mice do just as well as anything that might be learned from fMRI readings.

8. Conclusion: Why the Neuroeconomics Critique Fails

Kahneman (1994) notes the following two problems facing "a critic of the rationality assumption":

> *"... (ii) a willingness of choice theorists to make the theory even more permissive, as needed to accommodate apparent violations of its requirements; (iii) a methodological position that treats rationality as a maintained hypothesis making it very difficult to disprove...."*

Kahneman's observations make it clear that rationality is not an assumption in economics but a methodological stance. This stance reflects economists' decision to view the individual as the unit of agency and investigate the interaction of the purposeful behaviors of different individuals within various economic institutions. One can question the usefulness of this methodological stance by challenging individual economic models or the combined output of economics but one cannot disprove it.

The difficulties that Kahneman observes for critics of the rationality assumption are no different than the difficulties that one would encounter when challenging the assumption that laboratory experiments on individual choice are useful for understanding real-life behavior. For example, a critic of such experiments may complain that real-life choice problems do not come with explicit probabilities. If successful, such a criticism will lead to a new class of experiments, ones that do not make explicit references to probabilities.[19] However, a critic cannot expect to disprove the usefulness of experimental methods for understanding choice behavior. Criticisms that aim to disprove a broad and flexible methodology as if it were a single falsifiable assumption are best viewed as demands for a shift in emphasis from questions that the critic considers uninteresting to ones that he finds more interesting.

This latter description fits our view of what CLP (2005), Rabin (1998), and Kahneman (1994) describe as the radical challenge to economics:

> *"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. Neuroscience, we will argue, points to an entirely new set of constructs to underlie economic decision making. The standard economic theory of constrained utility maximization is most naturally interpreted either as the result of learning based on consumption experiences (which is of little help when prices, income and opportunity sets change), or careful deliberation- a balancing of the costs and benefits of different options – as might characterize complex decisions like planning for retirement, buying a house, or hammering out a contract. Although economists may privately acknowledge that actual flesh-and-blood human beings often*

[19] Compare for example, earlier experiments on the Allais' Paradox and the common ratio effect with later experiments on framing and reference points.

choose without much deliberation, the economic models as written invariably represent decisions in a 'deliberative equilibrium,' . . . " (CLP (2005), p. 10).

Populating economic models with "flesh-and-blood human beings," was never the objective of economists. Constrained optimization, Bayes law, and other economic abstractions do not represent the state-of-the art psychology of an earlier era. Therefore, there is no reason to believe that making the state-of-the art psychology of our time available earlier would have had such a profound effect on the development of economics.

Rabin (1998) argues that

"it is sometimes misleading to conceptualize people as attempting to maximize a coherent, stable, and accurately perceived $U(x)$."

Economists have at their disposal numerous devices to incorporate instability (or change) into individual preferences. They can assume that the decision-maker's preferences depend on an exogenous state variable, on the information of his opponents, or his own consumption history. The decision-maker may be learning about a relevant preference parameter, over time. All this flexibility or permissiveness notwithstanding, it is likely that the economists' model of the individual is not suitable for psychologists' goals. It does not follow from this that economists should adopt both the goals and methods of psychology.

Regardless of the source of their inspiration, economic models can only be evaluated on their own terms, with respect to their own objectives and evidence. A revolution in economics has to yield great economic insights. The CLP and Rabin agendas seem far reaching only because they define the task of economics as continually importing psychology-neuroscience ideas. Both papers offer very little in the way of novel economic analysis or implications. Rabin observes that *". . . fairness and reference-level effects (reviewed in Section 2) and framing effects (reviewed in Section 4) are likely to contribute to downward stickiness in wages"* but leaves *"it for other forums to explore these implications."* Similarly, all the challenges CLP (2005) identify for the emerging discipline of neuroeconomics resemble the current questions of psychology more than the current questions of economics.

A choice theory paper in economics must identify the revealed preference implications of the model presented and describe how revealed preference methods can be used to identify its parameters. Revealed preference earns such a central role in economics because this

is the form of evidence that is available to economists - and not because of a philosophical stance against other forms of evidence.

Greater psychological realism is not an appropriate modeling criterion for economics and therapeutic social activism is not its goal. Welfare analysis helps economists understand how things are by comparing the existing situation to how things might have been in a plausible alternative institutional setting; welfare theory is not a blueprint for a social movement.

We may be sceptical of neuroscientists' ability to come up with a universal, physiologically grounded criteria for measuring happiness. We may also have doubts about the potential effectiveness of neuroeconomists at convincing individuals or society as a whole, to adopt policies that increase "total happiness" by their measure. Our response to the neuroeconomics welfare theory is simpler: such a combination of moral philosophy and activism has never been the goal of economics; grounding this combination in biology is unlikely to make it so.

9. Appendix

Proof of the Proposition

First, we will show that (i) implies (ii): Suppose c is a general Köszegi-Rabin choice function. Then, there exists a reference dependent utility function U such that $c = C(\cdot, U)$. Define $\succeq$ as follows: $x \succeq y$ if $U(x,x) \geq U(y,x)$. Then, for all $A \in Y$,

$$c(A) = C(A,U) = \{x \in A \,|\, U(x,x) \geq U(y,x)\} = \{x \in A \,|\, x \succeq y\} = C_{\succeq}(A)$$

as desired.

To prove that (ii) implies (iii), assume that $c = C_{\succeq}$ and let n be the cardinality of X. Recall that $\succeq$ is a complete, reflexive, binary relation. We write $x \succ y$ for $x \succeq y$ and $y \not\succeq x$. Let $K = X \times X$. For $k = (w,z) \in K$, we define the function $u_{(w,z)} : X \to \{-2, 0, 2, 3\}$ as follows:

$$u_{(w,z)}(x) = \begin{cases} 3 & \text{if } x = w = z \\ 2 & \text{if } x = w \text{ and } w \succ z \\ -2 & \text{if } x = z \text{ and } w \succ z \\ 0 & \text{otherwise.} \end{cases}$$

Define the function μ as follows:

$$\mu(t) = \begin{cases} 16nt & \text{if } t \in \{-4, -3, 4\} \\ t & \text{if } t \in \{-2, 0, 2, 3\} \end{cases}$$

Clearly, μ is strictly increasing and $\mu(0) = 0$. Let

$$U(x,y) = \sum_{k \in K} u_k(x) + \sum_{k \in K} \mu(u_k(x) - u_k(y))$$

To complete the proof, we will show that $C_{\succeq} = C(\cdot, U)$; that is $x \succeq y$ iff $U(x,x) \geq U(y,x)$ for all $x, y \in X$, and

$$U(x,y) \geq U(y,y) \text{ implies } U(x,x) > U(y,x) \tag{4}$$

Note that

$$2n \geq \sum_{k \neq (x,x)} u_k(x) \geq -2n \tag{$*$}$$

Let $K_{x,y} = K\backslash\{(y,y),(x,x),(x,y),(y,x)\}$ and note that for $k \in K_{x,y}$

$$2 \geq u_k(x) - u_k(y) \geq -2 \qquad (**)$$

Equations (*) and (**) and the definition of μ imply that

$$4n \geq \sum_{K_{x,y}} (u_k(x) - u_k(y)) = \sum_{K_{x,y}} \mu(u_k(x) - u_k(y)) \geq -4n$$

Let $x \succeq y$. Note that $\mu(u_{(x,y)}(y) - u_{(x,y)}(x)) \leq 0$ and, since $x \succeq y$, we also have $\mu(u_{(y,x)}(y) - u_{(y,x)}(x)) \leq 0$. It follows that

$$\begin{aligned} U(x,x) - U(y,x) &= \sum_{k\in K} (u_k(x) - u_k(y)) - \sum_{k\in K} \mu(u_k(y) - u_k(x)) \\ &\geq -4n - \sum_{K_{x,y}} \mu(u_k(x) - u_k(y)) - \\ &\quad - \mu(u_{(x,x)}(y) - u_{(x,x)}(x)) - \mu(u_{(y,y)}(y) - u_{(y,y)}(x)) \\ &\geq -8n + 48n - 3 > 0 \end{aligned}$$

Conversely, let $y \succ x$. Then, $\mu(u_{(x,y)}(y) - u_{(x,y)}(x)) = 0$ and $\mu(u_{(y,x)}(y) - u_{(y,x)}(x)) = 64n$. Therefore,

$$\begin{aligned} U(x,x) - U(y,x) = &\sum_{k\in K} (u_k(x) - u_k(y)) - \sum_{k\in K} \mu(u_k(y) - u_k(x)) \\ &\leq 4n - \sum_{K_{x,y}} \mu(u_k(y) - u_k(x)) - \mu(u_{(x,x)}(y) - u_{(x,x)}(x)) \\ &\quad - \mu(u_{(y,y)}(y) - u_{(y,y)}(x)) - \mu(u_{(y,x)}(y) - u_{(y,x)}(x)) \\ &\leq 8n - 3 + 48n - 64n < 0 \end{aligned}$$

Finally, suppose $U(x,y) - U(y,y) \geq 0$. Then,

$$\begin{aligned} U(x,x) - U(y,x) &\geq U(x,x) - U(y,x) - U(x,y) + U(y,y) \\ &= -\sum_{k\in K} [\mu(u_k(y) - u_k(x)) + \mu(u_k(x) - u_k(y))] \\ &= -2(\mu(-3) + \mu(3)) = 2(48n - 3) > 0 \end{aligned}$$

completing the proof that (ii) implies (iii). That (iii) implies (i) is immediate. □

References

1. Bernheim, B. D. and A. Rangel, 2004. "Addiction and Cue-Conditioned Cognitive Processes," *American Economic Review*, 94(5): 1558–90.

2. Caplin A, and J. Leahy, 2001. "Psychological Expected Utility Theory and Anticipatory Feelings," *Quarterly Journal of Economics*, 2001, 55–80.

3. Camerer, Colin F., George Loewenstein and Drazen Prelec, 2005. "Neuroeconomics: How neuroscience can inform economics." *Journal of Economic Literature*, Vol. 34, No. 1.

4. Camerer, Colin F.; George Loewenstein; and Drazen Prelec, 2004. "Neuroeconomics: Why economics needs brains." *Scandinavian Journal of Economics*, 2004, Vol. 106, no. 3, 555–79.

5. Camerer, Colin; Samuel Issacharoff; George Loewenstein; Ted O'Donoghue; and Matthew Rabin, 2003. "Regulation for conservatives: Behavioral economics and the case for 'asymmetric paternalism'." *Univ. Penn. Law Review*, Vol. 151, 2111–1254.

6. Epstein, L. and S. Zin, 1989. "Substitution, Risk Aversion, and the temporal Behavior of Consumption and Asset Returns: A Theoretical Framework," *Econometrica* 57, 937-69.

6. Epstein, L. and S. Zin, 1991. "Substitution, Risk Aversion, and the temporal Behavior of Consumption and Asset Returns: An Empirical Analysis," *Journal of Political Economy* 99, 2, 263-86.

7. Frederick, S., G. Loewenstein, T. O'Donoghue, "Time Discounting and Time Preference: A Critical Review" *Journal of Economic Literature*, 2002, XL (2): 351–402.

8. Gruber, J. and B. Köszegi, "Is Addiction "Rational"? Theory and Evidence", *Quarterly Journal of Economics*, 2001: 1261–1303.

9. Gul, F. and W. Pesendorfer, "Temptation and Self-Control," *Econometrica*, 2001, 69(6): 1403–1436.

10. Gul, F. and W. Pesendorfer,2004. "Self-Control and the Theory of Consumption," *Econometrica*.

11. Gul, F. and W. Pesendorfer, 2005. "The Revealed Preference Theory of Changing Tastes", *Review of Economic Studies*.

12. Hsu, Ming and Camerer, Colin F., 2004, Ambiguity-Aversion in the Brain. Caltech, Working Paper.

13. Kahneman, D., 1994. "New challenges to the rationality assumption." *Journal of Institutional and Theoretical Economics*, 150, 18–36.

14. Kreps, D, 1979. "A Preference for Flexibility", *Econometrica*, 47: 565–576.

15. Kreps, D. and E. L. Porteus, 1978. "Temporal Resolution of Uncertainty and Dynamic Choice Theory," *Econometrica*.

16. Koszegi, B. and M. Rabin, 2004. "A Model of Reference-Dependent Preferences," mimeo.

17. Laibson, D., 1997. "Golden Eggs and Hyperbolic Discounting," *Quarterly Journal of Economics*, 112: 443–477.

18. Laibson, David. 2001. A Cue-Theory of Consumption. The Quarterly Journal of Economics, 116(1): 81–119.

19. Loewenstein, G., 1996. "Out of Control: Visceral Influences on Behavior," Organizational Behavior and Human Decision Processes, 65, 272–292.

20. Loewenstein, G. and D. Prelec, "Anomalies in Intertemporal Choice," *Quarterly Journal of Economics*, 1992, 107, 573–597.

21. O'Donoghue and Rabin, "Doing it Now or Later," American Economic Review, 1999, 89(1): 103–124.

22. O'Donoghue and Rabin, "Studying Optimal Paternalism, Illustrated with a Model of Sin Taxes," American Economic Review Papers and Proceedings, May 2003, 93(2), 186–191.

23. Rabin M., "A Perspective on Psychology and Economics," European Economic Review, forthcoming, 2002.

24. Rabin M., "Psychology and Economics," *Journal of Economic Literature*, 36, pp. 11–46, March 1998.

25. Strotz, R. H., 1956. "Myopia and Inconsistency in Dynamic Utility Maximization," *Review of Economic Studies*, 23(3): 165–180.

26. Thaler, R., 1980. Towards a positive theory of consumer choice. *Journal of Economic Behavior and Organization*, 1, 39–60.