NBER WORKING PAPER SERIES

NEUROECONOMICS: A SOBER (BUT HOPEFUL) APPRAISAL

B. Douglas Bernheim

Working Paper 13954
http://www.nber.org/papers/w13954

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
April 2008

I am grateful to Antonio Rangel and Colin Camerer for stimulating discussions and comments. I also acknowledge financial support from the National Science Foundation through grant number SES-0452300. The views expressed herein are those of the author(s) and do not necessarily reflect the views of the National Bureau of Economic Research.

Neuroeconomics: A Sober (but Hopeful) Appraisal
B. Douglas Bernheim
NBER Working Paper No. 13954
April 2008
JEL No. D01,D60,D87

ABSTRACT

This paper evaluates the prospects for the emerging field of neuroeconomics to shed light on traditional positive and normative economic questions. It argues that the potential for meaningful contributions, though often misunderstood and frequently overstated, is nevertheless present.

B. Douglas Bernheim
Department of Economics
Stanford University
Stanford, CA 94305-6072
and NBER
bernheim@stanford.edu

The last few years have witnessed impressive progress toward understanding the neurobiology of decision making. This progress reflects the individual and collaborative efforts of scholars from a variety of intersecting disciplines. The pace of discovery plainly establishes the viability of neuroeconomics as an independent, self-sustaining field, one that addresses a new set of fascinating and scientifically meritorious questions. Many participants in this growing field, as well as interested observers, hope that neuroeconomics will also eventually make foundational contributions to the various traditional fields from which it emerged, including economics, psychiatry, and artificial intelligence. My purpose here is to evaluate its potential contributions to economics.

Some would argue that any aspect of economic decision making is definitionally an aspect of economics. According to that view, neuroeconomics necessarily contributes to economics by expanding the set of empirical questions that economists can address. I will avoid such semantic quibbles. My interest here is in assessing whether, in time, neuroeconomics is likely to shed useful light on traditional economic questions. I recognize of course that the scope of traditional economics may eventually expand to include portions of neuroeconomics, even if neuroeconomics never addresses any economic question currently regarded as standard. However, regardless of whether economists eventually broaden their interests, it is still both legitimate and important to ask whether neuroeconomics can illuminate the issues that economists have historically addressed. While the scope of traditional economics is difficult to define with precision, I am content with an operational definition, based on the collection of questions and issues currently discussed in standard economic textbooks and leading professional journals.

The potential importance of neuroeconomics for economic inquiry has already been the subject of much debate. For example, an optimistic assessment appeared in a paper titled "Neuroeconomics: Why Economics Needs Brains," by Colin Camerer, George Loewenstein, and Drazen Prelec [2004].[1] Subsequently, Faruk Gul and Wolfgang Pesendorfer [2005]

[1]See also Glimcher and Rustichini [2004], Camerer, Loewenstein, and Prelec [2005], Rustichini [2005], Glimcher, Dorris, and Bayer [2005], and Camerer [2007]

penned a broad critique of neuroeconomics, titled "The Case for Mindless Economics," which expressed deeply rooted skepticism. My assessment lies between those extremes. I caution against dismissing the entire field merely because current technology is limited, or because some of the early claims concerning its potential contributions to standard economics were excessive and/or poorly articulated. However, because I share many of the conceptual concerns raised by Gul and Pesendorfer, I also see a pressing need for a sober and systematic articulation of the field's relevance. Such an articulation would ideally identify standard economic questions of broad interest (e.g., how taxes affect saving), and outline conceivable research agendas based on actual or potential technologies that could lead to specific, useful insights of direct relevance to those questions. Vague assertions that a deeper understanding of decision-making processes will lead to better models of choice will not suffice to convince the skeptics.

This paper represents my attempt to identify and articulate the specific ways in which neuroeconomics might contribute to mainstream economics, as well as the limitations of those potential contributions. It sets forth both my reservations and my reasons for guarded optimism. As will be evident, my evaluation is based in large part on the contemplation of research agendas that may or may not become technologically or practically feasible. My contention is only that there are conceivable paths to relevant and significant achievements, not that success is guaranteed. At this early stage in the evolution of neuroeconomics, the speculative visualization of such achievements is critical, both because it justifies the continuing interest and patience of mainstream economists, and because it helps neuroeconomists to hone more useful and relevant agendas.

The paper is organized as follows. I discuss potential contributions to positive economics in Section 1, potential contributions to normative economics in Section 2, and draw overall conclusions in Section 3.

1 Positive economics

While neuroeconomists are convinced that a better understanding of *how* decisions are made will lead to better predictions concerning *which* alternatives are chosen, many traditional economists greet that proposition with skepticism. In this section, I discuss and elaborate upon the basis for their skepticism, and then attempt to identify specific ways in which neuroeconomics could in principle contribute to traditional positive economics.

1.1 A framework for discussion

Advocates and critics of neuroeconomics (as it pertains to standard economics) often appear to speak at cross-purposes, using similar language to discuss divergent matters, thereby rendering many exchanges largely unresponsive on both sides. In the earnest hope of avoiding such difficulties, I will first provide a framework for my discussion, so that I can articulate and address particular issues with precision.

Suppose our objective is to determine the causal effects of a set of environmental conditions, x, on a decision vector, y.[2] For the time being, we will take x to include only the types of variables normally considered by economists, such as income and taxes. We recognize nevertheless that y depends not only on x, but also on a set of unobservable conditions, ω, which may include variables of the type studied by neuroeconomists. We hypothesize that the causal relationship between y and the environmental conditions, (x, ω), is governed by some function f:

$$y = f(x, \omega) \tag{1}$$

It is important to emphasize that the function f could be either a simple reduced form (e.g., a demand function expressing purchases of a good as a function of its own price, the prices of other goods, and income), or a more elaborate structural economic model. For

[2] Sometimes, the objective of traditional positive economics is simply to forecast y given a set of observed conditions x, without interpreting the forecasting relation as causal. In some contexts, it may be helpful to condition such forecasts on neuroeconomic variables; see the discussion in Sections 1.4 and 1.5, below.

instance, f could identify choices that maximize some objective function given the available alternatives when the conditions x and ω prevail.[3]

Economists typically treat the unobserved conditions, ω, as noise and attempt to determine the causal effects of the observed environmental conditions, x, on the distribution of decisions, y. If the distribution of ω is governed by a probability measure μ, then the distribution of y will correspond to a probability measure $\eta(\cdot \mid x)$, where for any Borel set A, $\eta(A \mid x) = \mu(\{\omega \mid f(x, \omega) \in A\})$. For example, the standard linear model assumes that

$$f(x, \omega) = x\beta + \varepsilon(\omega),$$

where ε is an unspecified function. It follows that $\eta(A \mid x) = \mu(\{\omega \mid x\beta + \varepsilon(\omega) \in A\})$.

Generally, economists attempt to estimate η directly from data on observable conditions, x, and decisions, y. In the case of the linear model, they estimate the parameter vector β along with parameters governing the distribution of $\varepsilon(\omega)$. There is no opportunity to recover the form of the function ε or the distribution of ω. Nor is there an obvious need. For example, when studying the behavioral effect of a sales tax on consumption, a traditional economist would not be concerned with quantifying the variation in that effect attributable to specific genetic traits; rather, she would focus on the distribution of responses (most notably the average) without conditioning on genetics. Accordingly, the identification of the causal relation $\eta(A \mid x)$, where x consists of standard economic variables such as income and taxes, is arguably the primary objective of traditional positive economics.

In contrast, the objective of positive neuroeconomics is, in effect, to get inside the function f by studying brain processes. To illustrate, let's suppose that neural activity, z (a vector), depends on observed and unobserved environmental conditions, through some function Z:

$$z = Z(x, \omega)$$

[3] In the latter case, an economist would typically interpret the free parameters of the objective function as aspects of preferences. However, modern choice theory teaches us that preferences and utility functions are merely constructs that economists invent to summarize systematic behavioral patterns. We are of course concerned with the accurate estimation of those parameters, but only because they allow us to recover the behavioral relation f.

Choices result from the interplay between cognitive activity the environmental conditions:[4]

$$y = Y(z, x, \omega)$$

It follows that

$$f(x, \omega) = Y(Z(x, \omega), x, \omega)$$

Positive neuroeconomics attempts to uncover the structure of the functions Z (the process that determines of neural activity) and Y (the neural process that determines decisions). Neuroeconomics necessarily treats the function f as a reduced form, even if it represents a structural economic model.[5] Neuroeconomic research can also potentially shed light on the distribution of ω (the measure μ), which is the other component of η, the object of primary interest from the perspective of traditional positive economics.

The tasks of traditional positive economics and positive neuroeconomics are therefore plainly related. The question at hand is whether their interrelationships provide traditional positive economists with useful and significant opportunities to learn from neuroeconomics.

1.2 Is the relevance of neuroeconomics self-evident?

Most members of the neuroeconomics community believe that the relevance of their field to economics is practically self-evident; consequently, they are puzzled by the persistent skepticism among mainstream economists. To motivate their agenda, they sometimes draw analogies to other subfields that have successfully opened "black boxes." For example, Camerer, Loewenstein, and Prelec [2004] write (see also Camerer, Loewenstein, and Prelec, 2005, and Camerer, 2007):

[4] The arguments of Y include x and ω in addition to z because the same neural activity could lead to different outcomes depending on the environmental conditions.

[5] As an example, let's suppose that the individual has a discrete number of choices, $y_1, ..., y_K$. Let z_i be a vector of affective responses when y_i is chosen; let $z = (z_1, ..., z_K)$. Since affective responses may depend on the environment (x, ω), we write $z = Z(x, \omega)$. Imagine that the individual actually chooses i (and hence y_i) to maximize some function $U(z_i)$. Then $Y(z) = y_{\arg\max_i U(z_i)}$ (which depends only on z and not directly on x or ω). In that case, $f(x, \omega) = y_{\arg\max_i \widehat{U}(y_i)}$, where $\widehat{U}(y_i) \equiv U(Z_i(x, \omega))$.

> "Traditional models treated the firm as a black box which produces output based on inputs of capital and labor and a production function. This simplification is useful but modern views open the black box and study the contracting practices inside the firm—viz., how capital owners hire and control labor. Likewise, neuroeconomics could model the details of what goes on inside the consumer mind just as organizational economics models what goes on inside firms." (Camerer, Loewenstein, and Prelec, 2004, p. 556)

From the perspective of a mainstream economist, analogies between neuroeconomics and the theory of the firm are misleading. In developing the theory of the firm, economists were not motivated by the desire to improve the measurement of reduced form production functions relating output to labor and capital. Rather, questions pertaining to the internal workings of the firm (unlike those pertaining to the internal workings of the mind) fall squarely within the historical boundaries of mainstream economics, because they concern organized exchange between individuals. The literature on the theory of the firm reflects a recognition that such exchange takes place not only within markets, but also within other types of institutions, including firms. It embraces the premise that resource allocation depends on the nature and scope of each exchange-facilitating institution. An economist who seeks to understand prices, wages, risk sharing, and other traditional aspects of resource allocation has an undeniable stake in understanding how trade plays out within a range of institutions, including markets and firms, and how different types of exchange come to be governed by different types of institution. In contrast, the mind is not an economic institution, and exchange between individuals does not take place within it.[6]

Notably, economists have not materially benefited from a long-standing ability to open up other black boxes. For example, we could have spent the last hundred years developing highly nuanced theories of production processes through the study of physics and engineering, but

[6] A mainstream economist might also take a prescriptive interest in the organization of firms: economic analysis can help to diagnose and fix a company that allocates resources inefficiently. In contrast, the diagnosis and treatment of poorly performing brains is traditionally the province of psychologists and psychiatrists, not economists.

did not. A skeptical mainstream economist might also note that models of neural processes are also black boxes. Indeed, the black box analogy is itself false: we are dealing not with a single black box, but rather with a Russian doll. Do we truly believe that good economics requires mastery of string theory?

In giving voice to these responses, I do not in any way wish to suggest that we should let all black boxes (or Russian dolls) remain closed. After all, the field of macroeconomics appears to have progressed through the systematic exploration of microfoundations. Nevertheless, it is understandable that so many economists are unmoved by the amorphous possibility that delving into the nuts and bolts of decision-making will lead to better and more useful economic theories. To persuade them that a particular black box merits opening, one must at least provide a speculative roadmap, outlining reasonably specific potentialities which economists would recognize as both directly relevant and within the realm of possibility. What has been offered along these lines to date is far too vague and insubstantial to convert the skeptics.

1.3 Some specific sources of skepticism

Neuroeconomists have certainly attempted to offer economists a variety of affirmative motivations for opening the black box of the human mind. Many mainstream economists find those motivations unpersuasive because they see neuroeconomic inquiry as largely orthogonal to traditional economic analysis, a view that finds its most forceful articulation in the work of Gul and Pesendorfer [2005]. To identify motivations that economists would generally find persuasive, one must first understand the logic of that view, and appreciate its appeal.

Much of the prevailing skepticism concerning the magnitude of the contribution that neuroeconomics can potentially make to standard positive economics arises from the following three considerations.

First, unless neuroeconomics helps us recover the behavioral relation η, its contributions will not advance the historical objectives of positive economics. Though the functions Y and Z are obviously interesting, the questions they address directly are not ones that mainstream

economists traditionally examine.

Second, because the behavioral relation η involves no neural variables, traditional positive economists can divine its properties from standard economic data. Distinguishing between two neural processes, (Y, Z, μ) and (Y', Z', μ'), is helpful to such an economist only if the differences between those processes lead to significant differences between the corresponding reduced form representations, η and η'. But if the latter differences are indeed significant, then an economist can (in principle) test between η and η' directly using standard economic data, without relying on neuroeconomic methods.

Third, while neuroeconomics potentially offers another route to uncovering the structure of the relation η, there is skepticism concerning the likelihood that it will actually improve upon traditional methods. The prospects for building up a *complete* model of complex economic decisions from neural foundations would appear remote at this time. Even if such a model were assembled, it might not be especially useful. Precise algorithmic models of decision making of the sort to which many neuroeconomists aspire would presumably map highly detailed descriptions of environmental and neurobiological conditions into choices. In constructing the distribution η from Y, Z, and μ, a microeconomist would treat vast amounts of this "micro-micro" information as noise. An economist might reasonably hope to apprehend the structure of η more readily by studying the relationship between y and x directly, particularly if the explanatory variables of interest (x) include a relatively small number of standard environmental conditions. As an example, suppose η is the household demand function for a good. What does a standard economist lose by subsuming all of the idiosyncratic, micro-micro factors that influence decisions, many of which change from moment to moment, within a statistical disturbance term? What can neuroeconomics teach us about the relationship between average purchases and the standard economic variables of interest (prices, income, and advertising), that we cannot discern by studying those relationships directly?

These considerations do not, however, rule out the possibility that neuroeconomics might

make significant contributions to mainstream economics. With respect to the second consideration, even the most skeptical economist must acknowledge that the standard data required to address questions of interest are sometimes unavailable, and are rarely generated under ideal conditions. Surely we should explore the possibility that new types of data and methods of analysis might help us overcome those limitations. Thus, the third consideration emerges as the most central to my appraisal, and the rest of this section is devoted to its evaluation.

In principle, even without providing a complete neural model of complex economic decision making, neuroeconomics offers several potential routes to uncovering the structure of standard behavioral relationships. First, it will lead to the measurement of new variables, which may usefully find their way into otherwise standard economic analyses. I discuss that possibility in Sections 1.4 and 1.5. Second, detailed knowledge concerning the neural processes of decision making may help economists discriminate between theories and/or choose between models. As discussed in Section 1.6, the formulation of rigorous tests may prove challenging. Standard economic theories of decision making concern choice patterns, and are therefore agnostic with respect to decision processes; hence, they may have few testable neural implications. Sections 1.7 and 1.8 examine the more modest possibility that an understanding a neural processes may provide economists with informal but nevertheless useful guidance with respect to model selection (specifically, explanatory variables and functional forms).

A skeptic might observe that the most promising routes to meaningful contributions are also the most limited. An economist whose analysis incorporates neural variables would not necessarily require extensive knowledge of neuroeconomic methods or a deep appreciation of neural processes; instead, she might simply rely on neuroeconomists to identify and collect the relevant data. Similarly, even if findings from neuroscience informally guide aspects of model selection (variables and/or functional forms), once a traditional positive economist knows the structure of the selected model, she can discard all information concerning neural

processes without loss. Consider an example. Neuroeconomics can perhaps illuminate the relationships between marketing and attention, and between attention and purchases. From those relationships, we can derive a reduced form relating marketing to purchases. But once that reduced form relationship is known, a positive economist arguably gains nothing of value from knowledge of its neural underpinnings.[7]

Many psychologists would view the positions outlined above as a form of radical behaviorism. They are surprised that economists still hew so rigidly to a perspective that psychology abandoned decades ago. Yet the different paths of psychology and economics are not so difficult to understand once we consider divergent objectives of those disciplines. I would point to two important differences. First, unlike economics, the field of psychology has traditionally subsumed questions about the mind. Thus, traditional psychological questions pertain to aspects of the functions Y and Z, whereas traditional economic questions do not. Second, questions in psychology often focus on the micro-micro determinants of behavior. A psychologist is potentially interested the particular factors that cause a single individual to behave in a certain way at a specific moment. In contrast, traditional economic analysis usually treats such idiosyncratic influences as background noise.

1.4 Are there uses for exogenous neuroeconomic variables?

The discussion in Section 1.3 takes $\eta(\cdot \mid x)$, with x defined to include only traditional economic variables, as the object of interest for traditional positive economics. It therefore ignores the possibility that neuroeconomics might redraw the boundary between the set of variables that economists treat as observable (x), and those they treat as unobservable (ω). More formally, by measuring some vector of variables $\widetilde{\omega}$, a neuroeconomist can repartition the environmental conditions (x, ω) into (x^0, ω^0), where $x^0 = (x, \widetilde{\omega})$ and $\omega = (\omega^0, \widetilde{\omega})$, and potentially allow economists to recover the causal relation $\eta^0(\cdot \mid x^0)$.

[7] The manner in which marketing influences attention is potentially of interest to a *normative* economist. For example, in situations where choice evidence is inconsistent, information concerning attention may reveal whether a particular choice reflects a full understanding of the true opportunity set. I discuss such possibilities, and the implied normative role for neuroeconomics, in Section 2.2.1.

It is important to acknowledge at the outset of this discussion that the barriers to redrawing the boundary between observable and unobservable variables may be practical and political, not merely technological. Participants in large-scale surveys may well decline to cooperate with neural or genetic "fingerprinting." Even if such information were collected, privacy concerns might preclude its release to the research community. After all, many existing data sets omit variables that are innocuous by comparison to genetics, such as the state or zip code of an individual's residence. Still, social attitudes toward privacy issues are changing (as exemplified by postings of personal information on the web), and there are various ways to protect the confidentiality of survey participants, for example by placing conditions and restrictions on the data's usage. For the purpose of this discussion, let us suspend disbelief and consider the possibilities.

Why might the distribution $\eta^0(\cdot \mid x^0)$, which subsumes the behavioral effects of neural variables, as well as the effects of standard environmental factors conditional on neural variables, be of interest to mainstream economists? The answer is not obvious. Suppose a neuroeconomist discovers a genetic trait that helps predict saving (a "patience gene"). Should mainstream economists greet that discovery with enthusiasm? Economics has not, after all, concerned itself historically with the relationship between genetics and saving. An economist might question whether that knowledge is likely to improve his understanding of the effects of, say, capital income taxes (an element of x) on asset accumulation, averaged or aggregated over the elements of ω (including genetics).

There are certainly contexts in which the information contained in $\eta(\cdot \mid x)$ completely answers a traditional economic question. However, in other contexts, $\eta^0(\cdot \mid x^0)$ also contains pertinent information. In addition, even if $\eta(\cdot \mid x)$ is the object of interest, the use of neural variables may facilitate its accurate measurement (implicitly or explicitly through the estimation of $\eta^0(\cdot \mid x^0)$). The following is a list of contexts in which neural variables may prove useful.

Detecting and mitigating bias associated with omitted variables. Economists often worry

that the explanatory variables in behavioral regressions may be correlated with unobserved aspects of preferences or talent that in turn influence behavior. Neuroeconomists hope to identify exogenous neural variables, such as genetic traits, that are associated with specific predispositions and abilities. If such data were available (admittedly a very tall order, both practically and politically), economists could use it to create neural proxies for tastes and talents. A significant partial correlation between an explanatory variable and a behaviorally pertinent neural proxy would point to an omitted variables problem, which the inclusion of such proxies would presumably mitigate.

Sometimes, correlations between explanatory variables and pertinent unobserved characteristics arise from selection effects. Consider, for example, the literature concerning the effects of 401(k) retirement saving plans on asset accumulation. It is widely recognized that those who are predisposed to save may sort themselves into jobs with 401(k) plans or press their employers to create such plans (see, e.g., the discussion in Bernheim, 2002). If that predisposition is omitted from a regression of saving on 401(k) eligibility (and other variables), the coefficient of the eligibility variable may exaggerate its causal effect. Armed with data on a "patience gene," we could determine whether an individual's underlying propensity to save predicts eligibility for a 401(k) plan, conditional on other characteristics, and thereby assess both the presence and potential severity of self-selection. The addition of sufficiently powerful neural taste proxies to the regression would presumably mitigate the resulting bias.

Curing endogeneity. In many economic settings, the decisions of distinct individuals are codetermined. To identify the causal effect of one individual's choice on another's decision, we require an instrument – specifically, a variable that directly affects the decision of one and only one individual. Neural predispositions arguably have that property. Consider, for example, the problem of estimating the size of peer effects in the context of charitable giving (e.g., Andreoni and Scholz, 1998, Carman 2003). The effect of one person's giving on another's gift is difficult to measure both because of selection effects (people may sort themselves into peer groups based on common characteristics related to giving), and because

peers are mutually influenced by each others' gifts. The discovery and measurement of an "altruism gene" could allow us to detect the presence of peer effects by studying the relationship between an individual's giving and the genetic charitable predispositions of his peers, controlling for his own predisposition.[8] From that reduced form behavioral relationship, we could recover a structural economic model relating each individual's gift to the giving of his peers. Equivalently, we could treat the endogeneity problem arising from selection and codetermination by using each peer's genetic charitable predisposition as an instrument for his or her gift.

Forecasting behavior as of a particular moment in time. Sometimes an economist is narrowly concerned with the accuracy of a behavioral forecast at a particular moment in time. If a neural condition is known to correlate with behavior, then a forecaster ought to use any available information concerning that condition. Consider the following example, originally suggested by Antonio Rangel. Suppose that an equity investor's neural state at the start of a day (a predetermined variable) predicts the nature of his trading strategy (e.g., caution versus aggression) over the course of the day better than conventional variables. Then by collecting neural measurements for a sample of traders, one might be able to forecast short-term movements of the stock market.

Extrapolating behavioral responses from one population to another. Sometimes, economists observe the effects of a policy intervention for one population (for example, participants in a pilot study), and must extrapolate its effects for a second population (for example, residents of a state). Suppose that responses differ from individual to individual according to observable characteristics (for example, age, gender, or ethnicity), and that the compositions of the two populations differ with respect to those characteristics. In that case, one can compensate for the compositional differences in two steps: (i) measure the responses in the first population conditional on the observable characteristics; (ii) aggregate based on the composition of the second population. Certain observable characteristics, such as

[8] Such a discovery could also help us detect and evaluate selection effects by directly measuring the extent to which people sort themselves into groups based on their genetic charitable predisposition.

gender and ethnicity, are of course simply aspects of genetics. To improve the accuracy of the overall forecast for the second population, one could add other pertinent genetic traits (as identified by neuroeconomists) to the list of observable characteristics upon which the analysis is conditioned.

Assessing the likely sensitivity of behavior to policy interventions. An appreciation of the role of genetics in decision making may lead to useful insights concerning the likely sensitivity of behavior to environmental conditions. Consider, for example, the intergenerational transmission of wealth. The discovery and analysis of a "patience gene" could shed light on the extent to which correlations between the wealth of parents and children reflect genetic predispositions rather than environmental factors that are presumably more amenable to policy interventions. One could in principle measure the effect of such a trait on asset accumulation by comparing the behavior of siblings with and without the trait (controlling, of course, for other pertinent factors such as birth order and gender). In combination with an estimate of the likelihood that a parent will pass the trait on to any given child, that information would permit one to infer the importance of a purely genetic (and hence fixed) component of intergenerational wealth transmission.

Keeping up with the real world. Neuroeconomics is potentially of interest to private companies. Some neuroeconomists foresee a near-term future in which employers subject job applicants to genetic tests that evaluate predispositions, and shoppers routinely encounter remote eye scans that allow advertisers to project highly tailored promotional messages (as in the science fiction film *Minority Report*). To describe and analyze resource allocation in that brave new world, economists would need to consider the roles of neural variables.

The deployment of such technologies would also be of potential interest and concern to public policy makers. Though it would also be theoretically possible to design public policies that more effectively promote social welfare by differentiating between individuals based on their neural characteristics, ethical and political concerns would likely preclude such alternatives, just as they preclude differential treatment based on gender and ethnicity.

Concerns over privacy, due process, and discrimination might also lead to limitations on the use of neural and/or genetic data by private firms. Without studying the neural correlates of behavior, economists will be unequipped to evaluate the effects of such policy choices on resource allocation.

1.5 Are there uses for endogenous neuroeconomic variables?

As I explained in Section 1.1, one of the main objectives of neuroeconomics is to uncover the structure of the function Y, which maps endogenous neural activity, z, along with the environmental conditions x and ω, to decisions. Based on existing findings concerning Y, it is already possible to predict certain choices from particular types of endogenous neural activity with a high degree of accuracy. For example, activity in the nucleus accumbens, the insula, and/or the mesial prefrontal cortex predicts purchase decisions (Knutson, 2007) and risk-taking (Kuhnen and Knutson, 2005), while right orbital frontal cortex activation in response to ambiguity predicts ambiguity aversion (Hsu et. al., 2005). Because accurate behavioral prediction is a central goal of positive economics, many neuroeconomists have offered such findings as evidence of their field's relevance (see, e.g., Camerer, 2007).

Why are mainstream economists unpersuaded by this evidence? In the context of most traditional economic questions, they see little value in predicting behavior based on its endogenous components (here, z). Consider the following stark example. Suppose our goal is to predict whether individual customers at a grocery store will purchase milk. After carefully studying a large sample of customers, a confused graduate student declares success, noting that it is possible to predict milk purchases accurately with a single variable: whether the customer reaches out to grab a carton of milk. The technology to collect this highly predictive data has long been available; economists have demurred not due to a lack of creativity, boldness, and vision, but rather because such predictions are of no value to them.

By discussing the preceding example, I do not mean to trivialize neuroeconomic research. Findings that help us understand the neurobiology of cognition and decision-making have unquestioned scientific merit. I am concerned here only with a narrow issue: whether those

findings illuminate traditional economic questions. For all its scientific merit, the ability to predict choices from endogenous brain activity is largely orthogonal to the objectives of mainstream economists.

It is useful to restate this point using the formal notation introduced in Section 1.1. The historical objective of positive economists is to improve the prediction of choice (y) from standard exogenous variables (x), such as taxes, income, prices, and so forth. The observation that one can more accurately predict choice from endogenous neural variables (z) simply does not speak to that objective.

Mainstream economists should not, however, completely dismiss the possibility that endogenous neural variables will prove useful. In some situations, information concerning some aspect of the environmental conditions, x, or the decision, y, may not be available. Data on neural activity (z) along with knowledge of the functions Y and Z can then potentially permit us to impute the missing conventional variables, and use the imputed values in otherwise standard economic analyses.

Imputations of unobserved environmental conditions. Private information plays a central role in large segments of modern economic theory. However, in many if not most cases, economists are no better positioned to observe private information than anyone else. For example, in the context of insurance markets, economists are often limited to some subset of the data available to insurance companies, which are typically unable to monitor policyholders' activities or elicit important private information.

Provided the collection of pertinent neural data is feasible, neuroeconomists could conceivably draw reliable inferences about private information without observing it directly. The technical feasibility of this agenda has already been established. Specifically, Wang et. al. (2006) conducted an experimental study of a "biased transmission game" involving cheap talk, in which a party known as the "sender" observes a state of the world and transmits a message to another party known as the as the "receiver," who then makes a decision. Payoffs are structured so that the sender wishes to mislead the receiver as to the true state. Sta-

tistical analysis reveals that it is possible to make meaningful inferences about the sender's private information by observing his pupil dilation (which tends to reflect arousal and/or stress).

In principle, neural discernment of private information could facilitate improved tests of economic theories in which such information plays a prominent role. However, the neuroeconomics community has yet to produce a useful application along these lines. This state of affairs is no accident. In a well-designed laboratory experiment, one can measure and manipulate private information directly, so neural inference is redundant. In the field, where private information is not observable, opportunities for collecting pertinent neural measurements are rare. Thus, it will be challenging to design an application that is both useful and feasible.

Imputations of choices. Empirical economists are often constrained by the quality and availability of standard choice data. One common problem is that choice data are not generally available for samples in which environmental conditions are randomly assigned. Consequently, causal interpretations of estimated behavioral relationships are often controversial. Random assignment is feasible in laboratory experiments, but the experimental replication of significant real-world choices can be extremely costly. In other contexts, such as when a firm introduces a new product, choice data are simply unavailable. Marketing specialists often estimate demand curves for new products based on answers to hypothetical questions, but such answers are notoriously suspect.

A more fundamental concern is that, even in the most favorable circumstances, we can only observe a single choice for a given individual at a particular moment in time; we never observe an individual's choice mapping. The estimation of a behavioral relationship thus requires the analyst to maintain hypotheses concerning the stability of that relationship either across distinct moments in time or across potentially diverse individuals.

How can neuroeconomics address these difficulties? When a person is presented with a set of prospects, each prospect generates neural responses. It is natural to hypothesize that

her choice from any such set bears a stable relation to those responses. If that hypothesis proves correct, appropriate neural measurements would permit us to predict accurately the choices she would make from any set of prospects even when no choice is offered. These imputed choices could in principle substitute for actual choices in otherwise standard economic analyses. For example, an economist might use such data to determine the causal effects on behavior of environmental conditions that only vary endogenously outside of the laboratory; the need for a costly controlled experiment with actual choices might be eliminated.

Neural methods may also enable economists to collect large amounts of imputed choice data from a single individual at a single point in time. To take a simple (if stark) example, suppose that when the brain is presented with a collection of prospects, each one generates an identifiable neural response that codes for value. Suppose further that, over a short time horizon, those relative values dictate any choice the individual might make from any subset of those prospects.[9] In that case, the measurement of those neural responses would allow an economist to construct an individual's entire choice mapping – that is, the choice she would make for every conceivable collection of objects drawn from the pertinent set – at a single moment in time. Using such data, an economist might, for example, be able to reconstruct a single consumer's demand curve at a particular moment in time, without maintaining any hypotheses concerning the stability of that relation across either time or diverse individuals. The potential richness of an imputed choice data set might even facilitate more powerful tests of consumer theory.

Presumably, neuroeconomic research will also identify limits on the reliability of imputed choices. For example, predictions may become less accurate when an individual is presented with a large number of prospects before being asked to choose from some subset. Under some conditions, neural responses may track hypothetical decisions more accurately than actual decisions. Imputations may prove less reliable when the choice is unfamiliar (e.g., the purchase of a new product), and/or sensitive to presentation when the alternatives are

[9] Note that this neuroeconomic hypothesis is testable. Indeed, it is consistent with the neural evidence presented by Padoa-Schioppa and Assad [2006].

complex. All of those possibilities bear careful investigation.

1.6 Do economic theories have testable implications concerning neural processes?

Perhaps the most tantalizing claim concerning the potential prospects of neuroeconomics is that an understanding of neural processes may provide economists with new opportunities to formulate direct tests of both standard and nonstandard (behavioral) theories of decision making (see, e.g., Camerer, 2007).[10] While such advances are conceivable, it is important for neuroeconomists to acknowledge the difficulty of this endeavor, and to avoid premature conceptual leaps, especially if they hope to be taken seriously by mainstream economists.

The central conceptual difficulty arises from the fact that standard economic theory (including neoclassical economics as well as much of modern behavioral economics) is agnostic with respect to the nature of decision *processes*. No explicit assumptions are made concerning the inner workings of the brain. For example, contrary to the apparent belief of many non-economists, economists do not proceed from the premise that an individual literally assigns utility values to alternatives, and from any opportunity set chooses the alternative with the highest assigned value. This disciplinary agnosticism with respect to process accounts for Gul and Pesendorfer's [2005] contention that neural evidence cannot shed light on standard economic hypotheses.

Foundational economic assumptions concern choice patterns, not processes. Neoclassical decision theory follows from a collection of choice axioms, the most critical of which is sometimes labeled *independence of irrelevant alternatives* (a generalization of the more familiar weak axiom of revealed preference). According to that axiom, if an individual chooses a particular alternative from an opportunity set, then he will also choose that alternative from any smaller set, provided the alternative remains available. When the independence axiom

[10] This issue is distinct from the possibility, to which I alluded at various points in Sections 1.4 and 1.5, that the measurement of neural variables may facilitate tests of conventional economic theories (e.g., by providing instruments or permitting reliable imputations for missing variables). The question here is whether one can test an economic theory of behavior by examining the *process* that governs decision-making.

is satisfied, there exists an ordering (interpreted as preferences) that rationalizes all of the individual's choices, in the sense that she always chooses the most highly ranked alternative according to the ordering. With some additional (but largely technical) axioms, one can also represent her choices as maximizing a continuous utility function. Within this framework, preferences and utility are merely constructs, invented by the economist to provide a convenient representation of choice patterns. The theory does not assert that these constructs have counterparts within the brain. Consequently, those who would test the theory by searching for such counterparts have misunderstood the theory's foundations.

Consider the following example. A computer has been programmed to make choices from collections of alternatives drawn from some universe of possibilities, $\mathbb{X}$. (To keep matters simple, I will assume that $\mathbb{X}$ is a finite set.) The program includes a large data set, with entries of the following form: (a, b, c), where $c \in \{a, b\}$. There is exactly one such entry for every pair $(a, b) \in \mathbb{X} \times \mathbb{X}$. When the computer is presented with an opportunity set $X \subset \mathbb{X}$, it places the elements of X in an arbitrary order, x_1, x_2, ..., x_N (where N is the number of elements in X). It then determines its choice, y, through the following procedure. To begin, provisionally set $y = x_1$. For $n = 2, ..., N$, iteratively update y as follows: look up the triplet (y, x_n, c); if $c = y$, then leave y unchanged, but if $c = x_n$, then change y to x_n. When this process is complete, choose y.

Let's assume the computer has been supplied with data that meet the following requirements. First, the triplet (a, b, a) appears in the data if and only if the triplet (b, a, a) also appears. Second, if the triplets (a, b, a) and (b, c, b) appear in the data, then so does (a, c, a). In that case, the data code for a binary relation that is reflexive, complete, and transitive. Given the decision algorithm, the computer's choices will satisfy the independence axiom. It will behave as if it chooses the best available alternative according to a preference ranking, or equivalently the alternative that delivers the highest value of some utility function. Yet even the most careful inspection of the computer code will fail to uncover a process that either consults a ranked list or assigns and maximizes utility values. Were we to leap to the

conclusion that such an inspection justifies rejection of the hypothesis that the computer is a neoclassical decision maker, we would clearly be in error.

The preceding observations do not, however, imply that neural evidence is conceptually incapable of shedding light on standard economic hypotheses. Choice axioms cannot be valid unless the neural processes that govern choice are capable of delivering decisions that conform to the axioms; thus, a mainstream economist cannot remain *entirely* agnostic as to process. To take an extreme possibility, if neuroeconomists succeed in reducing all pertinent neural decision processes to a precise computational algorithm for some reasonably large class of decision problems, they will be able to determine whether the algorithm delivers choices that satisfy the independence axiom, and thereby test neoclassical decision theory. However, that potentiality does not convincingly establish the value of neuroeconomics, for two reasons.

First, assume we have reason to believe that the brain sometimes employs a particular decision algorithm, but have not yet established the scope of that algorithm's application. Suppose the algorithm's implications for choice within some domain of decision problems, A, would be inconsistent with some economic theory;[11] moreover, there is no subset of A for which the same statement holds. We might hope to disprove the economic theory by demonstrating that the decision algorithm in fact governs choices throughout the domain A. However, a formal test of the latter hypothesis would presumably involve a comparison between the algorithm's behavioral predictions and actual choices throughout A. But if data on those decisions are available, we can test the economic theory directly, without concerning ourselves with the nuts and bolts of decision processes. Thus, the incremental contribution of neuroeconomics is not obvious.

Second, neuroeconomics is still a long way from reducing the neural processes that govern the complex decisions with which economists are conventionally concerned to precise algorithms, especially for broad classes of environments. Existing algorithmic representations of such processes pertain only to very simple tasks and functions. Much of what is

[11] Note that we can logically deduce those implications from the algorithm itself; no data analysis is required.

known has a qualitative flavor, e.g., that certain types of decisions involve elevated activity in particular regions of the brain, and that those regions tend to be associated with specific functions. While it is conceivable that we might be able to test economic theories using such information, the necessary conceptual groundwork for such a test has not yet been laid.

To describe what that groundwork would entail, I will introduce a bit of notation. Each possible neural architecture (or decision process), n, implements a particular computational algorithm, a. I will use A to denote the function that maps neural architectures to computational algorithms. In turn, every possible computational algorithm, a, implements a particular choice correspondence, c. I will use H to denote the function that maps computational algorithms into choice correspondences.

Now suppose we wish to formulate a rigorous neural test of the economic hypothesis that an individual's choice correspondence lies within some set C_x, defined by a choice axiom x (such as the independence axiom). A test along those lines would require us to identify testable features of the set of neural architectures, N_x, that generate choice correspondences in C_x (formally, $N_x = \{n \mid H(A(n)) \in C_x\}$). More specifically, as shown in Figure 1, we would need to complete the following steps:

1. Characterize $H^{-1}(C_x)$, the set of conceivable computational decision algorithms for which implied choices would satisfy the pertinent axioms (see the arrow labeled "Step 1" in Figure 1);

2. Characterize $A^{-1}(H^{-1}(C_x)) = N_x$, the set of conceivable neural decision processes that would implement the algorithms identified in step 1 (see the arrow labeled "Step 2" in Figure 1); and

3. Identify the testable features of elements of $A^{-1}(H^{-1}(C_x))$, so that it is possible to determine whether or not those features are in fact present by examining neural evidence. Formally, let E denote the set of neural architectures that are consistent with the available neural evidence; we reject the hypothesis that $c \in C_x$ if $E \cap A^{-1}(H^{-1}(C_x)) = \varnothing$,

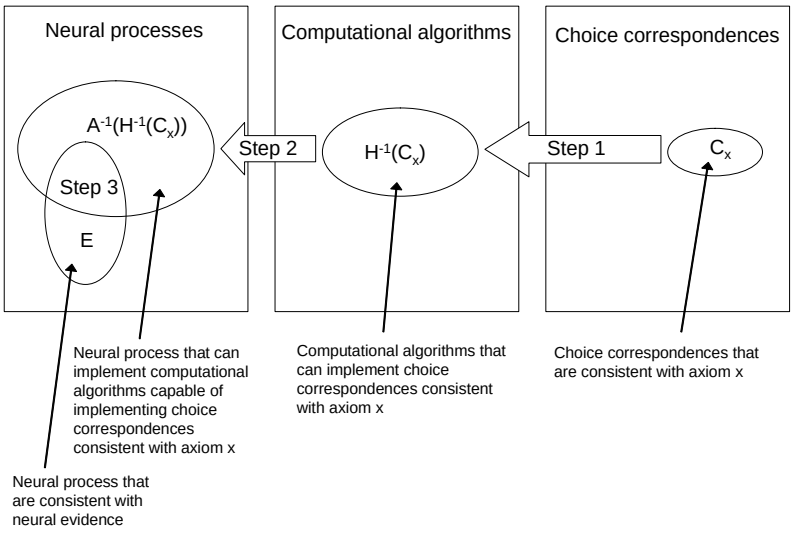

Figure 1: Necessary conceptual groundwork for a neural test of economic theory

and fail to reject it if $E \cap A^{-1}(H^{-1}(C_x))$ is non-empty. (Thus, in Figure 1, the label "Step 3" lies in the intersection of N_x and E). Naturally, this test has power only if there is some other behavioral hypothesis of interest, corresponding to an alternative choice axiom y, for which the testable features of N_x and N_y differ.

To my knowledge, no one has yet provided the characterizations referenced in steps 1, 2, and 3 for any widely invoked choice axiom. Those are extremely challenging tasks. It is likely that the characterizations will be highly complex, and there is no guarantee that any particular choice axiom will have useful testable implications in terms of neural processes. Consider the independence axiom. With respect to the characterization referenced in step 1, three types of algorithms come immediately to mind: one that codes for utility values, one that codes for preference rankings, and one that codes for complete, reflexive, and transitory binary relations (as illustrated above); there may well be many others. With respect to the characterizations referenced in steps 2 and 3, it is important to bear in mind that neural processes can have vestigial elements (evolutionary relics with minor roles in

decision making); consequently, the set of neural processes that can implement any particular computational algorithm is extremely large, and may exhibit enormous variety in terms of potentially testable features. I suspect it will be quite challenging to identify features of neural architecture that are either necessary or sufficient for the implementation of choice correspondences with particular properties.

Despite these conceptual concerns, matters are not completely hopeless. Neuroeconomics will presumably progress by formulating and testing increasingly specific theories of neural decision processes. We can associate any such theory, T, with a statement of the form $n \in M_T$, where n once again indicates the brain's neural architecture, and M_T is a theory-specific set. Even if the set $A^{-1}(H^{-1}(C_x))$ is analytically intractable, we may be able to identify a plausible neural theory, T, for which we can usefully characterize the set $M_T \cap A^{-1}(H^{-1}(C_x))$, and thereby derive testable implications of the choice axiom x within the restricted set of neural processes, M_T. In that case, we could formulate direct neural tests of x, treating T as a maintained hypothesis. Neuroeconomic research could then proceed along two complementary tracks: test T, and test x maintaining T.

As an example, consider a neural theory T which holds the following: when an individual is presented with a set of objects, his brain assesses a value for each, and encodes that value through a particular neural response; moreover, if he is then given the opportunity to choose among those objects, he selects the one with the greatest assessed value. If we maintain that hypothesis,[12] we may be able to derive testable neural implications for choice axioms. Take the independence axiom: the testable implication is that, if the individual is presented with a set of objects (at which point his neural valuations are measured) and then offered a choice among any subset, he will select the alternative with the greatest assessed value among the ones that are available. That implication is consistent with the limited results of one experiment, in which monkeys were offered choices among small numbers of alternatives (Padoa-Schioppa and Assad, 2007).

[12] Many neuroeconomists regard this theory as plausible; see, e.g., Padoa-Schioppa and Assad [2006].

Unfortunately, the neuroeconomic community has not yet generally acknowledged the conceptual challenges that one necessarily confronts when attempting to derive testable implications of economic theories for neural processes. Instead, neuroeconomists have sometimes proceeded (at times implicitly) as if those implications are obvious or easily motivated. That practice leaves many mainstream economists with the regrettable (and often inaccurate) impression that neuroeconomists do not adequately understand the economic theories upon which they hope to shed light. Examples of neuroeconomic results that have been incorrectly interpreted (sometimes by the authors but more often by others) as testing economic theories include the following.

Example #1: Dynamic inconsistency and quasihyperbolic discounting. McClure et. al. [2004] report that decisions activate distinct regions of the brain to differing degrees depending on whether they involve immediately available or delayed rewards. Moreover, the pattern of activation does not vary significantly with the amount of delay as long as rewards are not immediate. The paper is sometimes interpreted as providing a neural test of the popular β-δ model of quasihyperbolic discounting. That interpretation is inappropriate. The evidence does not establish that the forms of neural activity in question are related to valuation as opposed to some other function that plays a causal role in decision making, such as information processing.[13] Even assuming that the activity does involve valuation, the evidence does not establish that those valuations are time-inconsistent, or rule out the possibility that any inconsistencies are harmonized by other structures.

A rigorous neural test of the β-δ model would require a careful examination of the relationships between choice patterns, computational algorithms, and neural processes (as in Figure 1). Even without providing complete characterizations of those relationships, it is easy to see that the evidence in McClure et. al. [2004] cannot provide the basis for a valid test. We can frame the issue as a computer programming task. It is plainly possible to

[13] Nor does the evidence in McClure et. al. [2004] establish that those forms of neural activity are critical parts, rather than by-products, of the decision process. I acknowledge that it is in principle possible to establish a causal role for specific types of brain activity in particular decisions by microstimulating the pertinent brain regions. Such evidence would not, however, address the other concerns raised in the text.

write a program that implements time-consistent decisions, but that nevertheless evaluates immediate and delayed rewards in separate subroutines. Likewise, it is plainly possible to write a program that implements time-inconsistent decisions, but that nevertheless evaluates immediate and delayed rewards using precisely the same lines of code. Thus, evidence of this type is inherently incapable of distinguishing between the β-δ model and the conventional model of time-consistent choice.

Though the paper is not completely clear on this point,[14] a friendly reading of McClure et. al. [2004] suggests that the authors had in mind a more reasonable interpretation of the evidence. Specifically, *under the maintained hypothesis that choices conform to the β-δ model*, they test the supplemental hypothesis that such choices are generated by a dual-system process, with limbic structures governing the evaluation of immediate rewards, and the lateral prefrontal cortex and related structures governing the evaluation of delayed rewards. Their evidence is certainly consistent with that supplemental hypothesis, even though it sheds no light on the validity of the maintained hypothesis. Of course, mainstream economics concerns itself with the maintained hypothesis, and not with the supplemental hypothesis. Accordingly, the typical economist finds this paper fascinating, but not particularly relevant.

Example #2: Altruism and "warm glow" giving. Harbaugh et. al. [2005] report that tax-like transfers to charity produce neural activity in areas of the brain that have been linked to reward processing, and that voluntary transfers of the same magnitude generate higher levels of that activity (see also Harbaugh et. al., 2008). The authors interpret those findings as evidence that the motivations for giving include, respectively, pure altruism (in the case of the first finding) and the "warm glow" that flows from self-sacrifice (in the case of the second). Unfortunately, that interpretation is problematic, partly because it reflects confusion concerning the nature of pure altruism and warm glow giving as *economic* hypotheses, and partly because it implicitly relies on various unstated assumptions concerning the nature of computational and neural implementations of the pertinent choice patterns.

[14] Potential confusion arises because McClure et. al. [2004] do not explicitly state whether β-δ behavior is treated as a maintained hypothesis, or as part of a joint hypothesis.

Economists who study altruism and voluntary giving (myself included) frequently motivate, formulate, and describe their models in terms of utility and well-being. However, it is important to remember that economic theories of giving remain rooted in choice. Economists do not abandon their standard framework when studying giving; instead, they simply broaden the definitions of the objects of choice to include elements that affect other people. As a matter of fundamentals, specific hypotheses are still identified with choice patterns.

For the purpose of discussing the hypotheses at issue, I will define the typical object of choice as a triplet, (c, t, b), where c is private consumption, t is the size of the transfer from the individual to a charity, and b is the total budget of the charity. The total transfer from others, T, is implied as a residual: $T = b - t$. The hypothesis of pure altruism holds that (c', t', b') is chosen over (c'', t'', b'') iff $(c', \widehat{t}', b')$ is chosen over $(c'', \widehat{t}'', b'')$ for all t', t'', $\widehat{t}'$, and $\widehat{t}''$, so that preferences can be defined over pairs of the form (c, b). (Note that variations in t with b fixed imply variations in T.) The hypothesis of pure warm glow giving holds that (c', t', b') is chosen over (c'', t'', b'') iff $(c', t', \widehat{b}')$ is chosen over $(c'', t'', \widehat{b}'')$ for all b', b'', $\widehat{b}'$, and $\widehat{b}''$, so that preferences can be defined over pairs of the form (c, t). Economists are interested in these hypotheses because they have divergent positive implications, for example concerning the degree to which public contributions crowd out private contributions.

Harbaugh et. al. [2005] suggest that it is possible to detect the presence of pure altruism by examining neural responses to an outcome it is mandated (that is, *when no choice is involved*), and to detect the presence of warm glow motives by comparing neural responses when an outcome is voluntarily chosen and when it is mandated. However, because both hypotheses (altruism and warm glow) pertain to choice patterns, neither has any implication concerning well-being, feelings, or neural activity when choice is absent. Consequently, an examination of such activity following mandates cannot reveal which motive lies behind *behavior*.

These points require some elaboration. Formally, let U denote the utility function that rationalizes the individual's choices. To test the hypotheses of interest, we would attempt

to determine whether the arguments of U exclude b (for pure altruism) or t (for warm glow giving). That is not what Harbaugh et. al. [2005] do; indeed, in their experiment, b and t do not vary independently. Rather, they implicitly posit the existence of another function, call it W, that measures well-being when outcomes are mandated, and they attempt to test two hypotheses using neural data: first, that the arguments of W exclude b and t (by examining whether mandated contributions produce elevated activity); second, that $W = U$ (by examining whether mandated and voluntary contributions produce the same level of activity). To the extent those hypotheses are meaningful and testable,[15] they pertain to normative matters, but have no positive implications (because behavior conforms to U, not W), and therefore cannot differentiate between behavioral theories.

Taken at face value, the evidence in Harbaugh et. al. [2005] rejects the hypothesis that the arguments of W exclude b and t, in that a mandated contribution elevates neural reward-related activity, as well as the hypothesis that $W = U$, in that a voluntary contribution elevates neural reward-related activity more than a mandated contribution. The authors construe the first pattern as evidence of pure altruism and the second as evidence of a warm glow motivation. In their view, an individual would only experience a warm glow if she made a contribution voluntarily; thus, elevated activity from a mandatory contribution must be attributable to altruism. They also assume that a pure altruist would benefit equally from a contribution regardless of whether it was voluntary or mandated; hence, greater elevation with a voluntary contribution must be attributable to a warm glow. But those addendums are not part of the *economic* warm glow and pure altruism hypotheses. The utility rationalization for warm glow hypothesis holds only that, *ceteris paribus*, larger *voluntary* contributions lead to greater satisfaction. It is inherently mute as to whether larger *involuntary* contributions lead to the same gains in satisfaction, and is therefore consistent with the possibility that contributions lead to the same warm glow regardless of whether they are voluntary or involuntary. Thus, rejecting the hypothesis that the arguments of

[15] There are conceptual problems with neural measures of well-being, which I discuss in Section 2.1.

W exclude t and b does not favor pure altruism over warm glow giving. Likewise, the utility rationalization for the pure altruism hypothesis holds only that, *ceteris paribus*, a larger budget for the charity leads to greater satisfaction when it results from *voluntary* contributions. It is entirely consistent with the possibility that larger budgets lead to greater gains in satisfaction when they result from voluntary rather than involuntary contributions (e.g., because discretion is valued). Thus, rejecting the hypothesis that $W = U$ does not favor warm glow giving over pure altruism. At best, that evidence speaks to a normative proposition concerning the intrinsic value of free choice, not to positive questions concerning behavior.

Leaving aside the various issues discussed above, any rigorous neural test of the pure altruism and warm glow hypotheses would require a careful examination of the relationships between the associated choice patterns, computational algorithms, and neural processes. Harbaugh et. al. essentially assume that the brain must implement either choice pattern by coding for utility,[16] that the neural activities which they measure completely encompass utility, and that the measured activities are not contaminated by other functions. Their justifications for these assumptions are not apparent.

Example #3: Expected utility. A number of studies have discovered neural circuitry that, among other functions, appears to encode expected payoffs. For example, in a study of monkeys, Platt and Glimcher [1999] found that the firing rate of certain neurons in the lateral intraperietal cortex (LIP) is highly correlated with the expected value of an anticipated reward (the volume of juice). Unfortunately, such evidence sheds little if any light on the validity of expected utility theory. At best, it shows that a particular neural response converges to a measure of expected payoff in a simple, stationary environment after repeated trials. It does not show, for example, that humans have neural circuitry that encodes ex-

[16] Evidence of correlations between voluntary choices and neural responses, which the authors present, does not establish that the brain implements choices by coding for utility through those responses. We can once again frame the issue as a computer programming task. It is clearly possible to write a program that implements choice patterns associated with either pure altruism or warm glow giving without coding explicitly for utility, but that still generates variables that are correlated with choice.

pected payoff when they merely informed of objective probabilities (without repeated trials), or even more importantly, when they are provided with no objective information, so that probabilities are subjective. Even if such circuitry exists, other systems could influence decisions, causing violations of expected utility theory.[17] Notably, Camerer [2007] acknowledges that evidence of Bayesian neural mechanisms "is in sharp contrast with many cognitive psychology experiments showing that Bayesian principles are violated when intelligent humans evaluate abstract events."

A rigorous neural test of expected utility theory would require a careful examination of the relationships between choice axioms, computational algorithms, and neural processes. We can once again frame the issue as a computer programming task. It is clearly possible to write a program that implements a choice rule consistent with Savage's [1954] axioms without coding explicitly for expected utility.[18] Likewise, it is also possible to write a program that codes for expected payoffs, but that nevertheless leads to choices that are inconsistent with those axioms. Thus, evidence of this type is inherently incapable of distinguishing between expected utility theory and other hypotheses concerning choice under uncertainty.

1.7 Can an understanding of neural processes usefully guide model selection?

The number of empirical models that an economist could construct to describe any particular decision as a function of conventional explanatory variables is vast. Even if neuroeconomics does not provide new variables of interest (the topic of Sections 1.4 and 1.5) or an independent foundation for testing one model against another (the topic of Section 1.6), it could conceivably generate suggestive findings that informally guide the search for an appropriate empirical model in useful directions, leading to more rapid and effective identification of the best predictive relationship. I will discuss the two main aspects of model selection: variable

[17] Once again, it may be possible in principle to establish through microstimulation of the pertinent brain areas that specific neural activity plays a causal role in decision making. However, such a finding would not address the concerns expressed in the text.

[18] For example, following the example provided earlier, one can code for a binary relation over lotteries that satisfies Savage's axioms.

selection and the choice of functional form.

First consider variable selection. Neuroeconomic evidence could in principle motivate the inclusion of particular conventional variables in specific behavioral models. Suppose, for example, that mandated transfers to others influence brain activity in centers linked to reward-processing, as the evidence in Harbaugh et. al. (2005) suggests. While such evidence would not prove that altruism motivates behavior, it might well *suggest* such a hypothesis to an empirical economist, who might then investigate the predictive power of behavioral models that incorporate related variables (e.g., measures of potential externalities). The effects of those variables might prove interesting in their own right, and their inclusion might purge the estimated effects of other conventional variables of otherwise spurious correlations with the behavior of interest. Similarly, an examination of neural evidence concerning the processes that govern attention might suggest that consumers are potentially susceptible to tax illusion, and that they will respond differently depending on whether a product is tagged with tax-inclusive or tax-exclusive prices. Such evidence might lead an empirical economist to examine empirical models that separately include explanatory variables measuring posted prices and hidden taxes.

While acknowledging the possibilities described in the preceding paragraph, a skeptic might nevertheless question whether neuroeconomics is likely to make such contributions in practice. Empirical economists have other sources of guidance and inspiration, such as introspection and research from psychology. Indeed, neural studies such as Harbaugh et. al. [2005] are themselves motivated by hypotheses imported from other fields. I doubt that Harbaugh et. al. [2005] would have searched for neural correlates of altruism had other work in the social sciences (which they cite) not pointed toward altruism as a significant motivational factor. Likewise, economists formulated and tested conjectures concerning tax illusion based on a common-sense understanding of attention, without the benefit of neuroeconomic evidence; see in particular Chetty, Looney, and Kroft (2007), and Finkelstein (2007). Empirical economists who are not persuaded to investigate the roles of pertinent

variables in behavioral relationships on the basis of other considerations are unlikely to be convinced by the neural evidence. To uniquely motivate the inclusion of a potential explanatory variable that empirical economists have ignored, a neuroeconomist would literally have to stumble across some unexpected environmental correlate of brain activity. I do not dismiss that possibility, but neither does it convince me that the field holds great potential for conventional positive economics.

Even if research on the neurobiology of decision making had provided the impetus for investigating altruism, tax illusion, or some other phenomenon, it seems unlikely that an empirical strategy for estimating the function η would have been influenced by the details of the neurobiological evidence. Rather, that evidence would have merely *motivated* (to use Gul and Pesendorfer's term) an examination of functional forms that include the pertinent variables. It is not at all obvious that an economist who possesses a deep understanding of the motivating scientific evidence would be any better equipped to estimate η than one who simply apprehends the pertinent psychological principles intuitively.

In addition to suggesting that certain variables may play roles in particular behavioral relationships, neuroeconomic evidence may also indicate that others play no role. Such evidence could motivate exclusion restrictions. Indeed, formal neural tests of exclusion restrictions are conceivable in principle, even without precise knowledge of the computational algorithms that govern decision-making. We can once again frame the issue as a computer programming task. To implement a choice mapping that depends on a particular variable, computer code must reference that variable. For any neural process that implements the same computational algorithm, there must presumably be some neural response to the variable's value. Consequently, the absence of any response would formally justify an exclusion restriction in the behavioral relationship.

Next consider the choice of functional form. In principle, the nature of neurobiological response mechanisms may suggest particular empirical specifications. For example, there is some evidence that temporal difference reinforcement learning (TDRL) models accurately

describe the operation of neural systems governing dopamine learning (Schultz, Dayan, and Montague, 1997, and Schultz, 1998, 2000). These parsimonious, tightly parameterized learning models could guide the formulation of empirical behavioral relationships in settings that involve the accumulation of experience. Because other learning processes may also influence choices, the neural evidence cannot *prove* that one functional form is better than another for the purpose of predicting behavior. However, it could lead economists to examine particular parsimonious specifications that they might not otherwise consider, and some of these may outperform more conventional alternatives.

A mere catalog of such possibilities will never suffice to convince the skeptics, nor should it. Mainstream economists should acknowledge the conceptual possibilities discussed above, and exercise intellectual tolerance and patience while neuroeconomists explore them. Neuroeconomists should recognize in turn that the burden of proof is squarely on their shoulders. Skeptical reactions define a specific challenge: *Provide an example of a novel economic model derived originally from neuroeconomic research that improves our measurement of the causal relationship between a standard exogenous environmental condition – one with which economists have been historically concerned – and a standard economic choice.* Unless the neuroeconomics community eventually rises to that challenge, the possibilities discussed in this section will eventually be dismissed as unfounded speculation.

1.8 Can neuroeconomics improve out-of-sample predictions?

Sometimes, economists wish to predict behavior under completely novel conditions (for example, a new and untried public policy). There is no assurance that reduced form behavioral models will perform well in such contexts, especially if the novel conditions are qualitatively distinct from any that have preceded them. In contrast, a good structural model, based on a deeper understanding of behavior, may permit reasonable projections even when fundamental environmental changes occur. Many neuroeconomists believe that their field will provide such models.

By way of analogy, suppose a computer has been programmed to make selections for

choice problems that fall into a number of distinct categories, but the tasks for which we have observed its choices belong to a subset of those categories. We could potentially develop a good positive model, conceivably along the lines of standard economic theories (e.g., utility maximization), that predicts the computer's choices for problems within the categories for which we have data. However, based on that limited data, projecting choices for problems within the remaining categories is guesswork. Now suppose that someone obtains the computer code. In that case, even without additional choice data, we could accurately predict the computer's decisions in *all* circumstances. When neuroeconomists suggest that an understanding of the brain's computational algorithms will permit more reliable out-of-sample behavioral predictions, they are making an analogous claim.

Unfortunately, the issue is not quite so straightforward. The analogy is convincing only if we assume that the totality of all decision processes within the brain will be reduced to a precise computational algorithm. If, as is more likely, neuroeconomists only succeed in mapping a subset of the brain's neural circuitry to computational algorithms, out-of-sample prediction will remain problematic. To pursue the analogy a bit further, suppose we obtain the code only for certain subroutines that are activated when the computer solves problems falling within the categories for which we have data. There is no guarantee that it will activate the same subroutines for related purposes when confronting problems within the remaining categories, particularly if those problems are qualitatively different from the ones previously encountered. Without knowing how the entire program operates, including the full array of subroutines upon which it can call, as well as the conditions under which it activates each of them, one cannot simulate its operation in fundamentally new environments.

Of course, one can proceed based on the *assumption* that the brain will continue to use the same neural circuitry in the same way when confronting new classes of decision problems. But there is no way to *test* that assumption until out-of-sample observations become available, and no guarantee of greater stability at the neural level than at the behavioral level.[19] If,

[19] Just as a structural economic model can be viewed as a reduced form for a structural neural model, any structural neural model can also be viewed as a reduced form for some deeper structure, and the stability of

for example, secondary (and normally quiescent) neural systems are designed to override a primary system whenever the latter would generate behavior too far from the individual's norm, then an incomplete neural model of choice might be less stable out of sample than a behavioral model. Whether we would be better off making out-of-sample predictions from structural neural models rather than structural behavioral models is therefore a factual question that can only be settled through experience, and not through logical arguments.

Still, there are reasons to hope that consideration of evidence on neural processes might at least help us select economic models that are more reliable for the purpose of making out-of-sample projections. Imagine, for example, that an estimated within-sample behavioral relationship is equally consistent with several distinct structural economic models, each of which has a different out-of-sample behavioral implication. Suppose the available neural evidence informally persuades us (but does not prove) that one of those models is more likely to match reality. Then we might reasonably hope to obtain more accurate out-of-sample predictions from the preferred model.

Consider the following example. Currently, tens of millions of people lack health insurance coverage. One theory holds that those households have carefully assessed the costs and benefits of insurance, and concluded that it is too costly; another holds that they are inattentive to their health care needs, and hence unresponsive to costs and benefits. Both hypotheses are equally consistent with observed choices, but they have starkly different out-of-sample implications concerning the fraction who would purchase insurance if the cost of coverage were reduced well below historical levels. Can neuroeconomics help us judge between their divergent predictions? Suppose we use neural methods to measure attentiveness to health care needs, as well as value assessments for insurance coverage. The first theory informally predicts high attentiveness and high value assessments; the second has the opposite prediction. Neither finding would *prove* that the uninsured are more likely to behave one way or the other out of sample. For example, the uninsured might *start* attending to

the neural reduced form over classes of environments will depend on how that deeper structure operates.

health care issues and contemplating the benefits of insurance if they thought health care was affordable. Even so, the neural evidence would presumably influence our comfort with and degree of confidence in each theory.

All of these possibilities are of course speculative. Mainstream economists will relinquish their skepticism only when confronted with examples of superior out-of-sample prediction in contexts involving the types of environmental conditions and behaviors that economist ordinarily study.

1.9 An overall assessment

In pondering the future of neuroeconomics, I see substantial likelihood that the field will make intellectually legitimate contributions to positive economics. At the same time, a number of the potential contributions discussed in this section strike me as somewhat modest, rather special, and/or somewhat peripheral. While there is good reason to hope that some of the contributions will prove noteworthy, I have considerably more difficulty convincing myself that neuroeconomics is likely to become a central or indispensable component of standard positive economics, or that it will revolutionize the field in some fundamental way. Whether that assessment reflects the field's actual limitations or the deficient imagination of a relatively mild skeptic remains to be seen.

2 Normative economics

In standard economics, normative judgments are rooted in the choices of the affected parties. If an individual would choose option A over option B, then economists typically treat option A as better for her than option B. Henceforth, I will refer to this normative judgment as the *libertarian principle*.

The libertarian principle involves deference to each individual's judgment. The notion that such judgments merit deference reflects a particular perspective concerning well-being, one that I will embrace throughout this discussion. I recognize, of course, that some read-

ers may favor alternatives. Deference to the individual is an attractive principle because it guards against the possibility that one individual will impose personal and potentially arbitrary judgments on another.

Any contribution of neuroeconomics to normative economics would presumably take one of two forms. First, neuroeconomics might lead economists to develop a entirely new approach to measuring an individual's welfare, one that evaluates her well-being based on her neural activity rather than her choices. Second, neuroeconomic research might allow economists to improve choice-based welfare analysis without abandoning the standard normative paradigm. I will consider each of these possibilities in turn.

2.1 Can neuroeconomics offer an alternative to choice-based welfare analysis?

Prior to the revealed preference revolution, classical economists such as Francis Edgeworth, Frank Ramsey, and Irving Fisher speculated about the possibility of measuring utility directly (see Colander, 2005). Will neuroeconomics provide us with the technology to make such measurements, and ultimately replace choice-based welfare analysis with a new utilitarian paradigm? For the reasons detailed in the next two subsections, I am skeptical. Moreover, technological advances are unlikely to address the main sources of my skepticism, which are largely conceptual.

2.1.1 Problems associated with the construction of a neural welfare measure

Because the human brain relies on multiple motivational systems (Balleine et. al., 2008), we must acknowledge the very real possibility that no single type of neural response codes for overall well-being. Thus, the construction of a neural welfare measure requires the identification of all welfare-relevant neural activity. To say with confidence that an individual is better off with one alternative than another based on neural activity, it is not sufficient to demonstrate that certain activities code for certain aspects of well-being, or even that they code for well-being comprehensively in certain circumstances. We must also establish that

the identified activities do not neglect any significant aspect of well being in any pertinent circumstance. Even if we can prove that certain types of neural activity code for aspects of well-being (leaving aside for the moment the issue of *how* we might reach such a determination), the task of demonstrating that no other type of neural activity codes for any aspect of well-being is likely to prove far more challenging.

Similarly, it is important to acknowledge the possibility that there may not be a clean separation between welfare-relevant activity and other activity within the brain. The circuitry that registers pleasure or codes value may also be involved in other functions, such as information processing. To say with confidence that an individual is better off with one alternative than another based on neural activity, it is not sufficient to demonstrate that the activity in question is related to well-being. We must also establish either that it is not systematically related to anything else, or that we can somehow purify its measurement.

Even if we could identify neural activities that code comprehensively for well-being and nothing else, we would still confront the problem of aggregation. How can we identify objective principles for combining various measures of welfare-related neural activity into a single index? We might hope to discover that brain itself aggregates well-being and codes it as a single type of neural activity. But what type of evidence would allow us to distinguish that activity from the aggregated components? If, as is more likely, the neural aggregator either fails to exist or is impossible to identify, we would be forced to adopt principles of aggregation for which there is no neural foundation. The resulting welfare index then becomes a hybrid of neural and non-neural concepts, and the latter (whatever they are) as well as the former must withstand scrutiny.

In addition to requiring aggregation over various dimensions of brain activity, a neural measure of welfare would also require aggregation over time. Suppose that an individual must choose between two alternatives, A and B, with consequences at dates 0 and 1. Imagine optimistically that we discover how the brain codes an overall sense of well-being at each moment in time. Let u_t^i denote the coded level of well-being for activity i at time t. If

$u_0^A > u_0^B$, and $u_1^B > u_1^A$, is the individual better off with alternative A or B? If the value of u_0^i is unrelated to the value of u_1^i (so that we can interpret u_0^i as a measure of flow utility, rather than a forward-looking index of well-being), how would we aggregate u_0^i and u_1^i? If the value of u_0^i is found to vary with the value of u_1^i (so that it appears to be forward looking to some degree), is it then appropriate to base welfare judgments entirely on u_0^i and ignore u_1^i? How would we determine whether this effect reflects aggregation of feelings at different points in time, or immediate feelings driven by anticipated outcomes (in which case aggregation would still be necessary)? What principles would we use to determine whether u_0^i aggregates appropriately?

These various issues must, of necessity, undermine the confidence one can reasonably have in any neural welfare measure. To put the matter starkly, suppose that when the available alternatives are A and B, the individual chooses A *regardless of how or when the choice is presented* (in other words, it is impossible to induce him to choose B over A),[20] while the neural welfare measure points unambiguously to alternative B. In light of the various problems listed above, I submit that this fact pattern would lead us to suspect that the neural welfare measure, rather than a choice-based measure, is flawed. Indeed, we can construe such a pattern as evidence that the neural measure either (a) is not comprehensive, (b) has not been purged of all influences that are not welfare-relevant, or (c) involves an inappropriate judgment concerning aggregation. In other words, we can interpret comparisons between choices and a neural welfare measures as validating or invalidating the various judgments and decisions made in the processing of arriving at the neural measure. For example, we can use evidence on choice as the standard for evaluating whether certain types of neural activity code for aspects of well-being.

Could we also use non-choice data to both guide the construction of, and validate, a neural welfare measure? If so, we might then reinterpret the fact pattern described in the preceding paragraph as establishing that the choice-based measure, rather than the neural

[20] One can of course induce an individual to choose the option labeled B over the one labeled A through coercion or by offering inducements. But in that case the actual objects of choice are no longer A and B.

measure, is problematic. However, for the reasons discussed in the next subsection, I am skeptical of that possibility.

2.1.2 Problems associated with the justification for a neural welfare measure

Economists have considered using at least three types of data for normative analysis: choice, self-reported happiness (or preferences), and neural activity. We can justify the use of any particular type of data in one of two ways: either we can define welfare in terms of a measured variable (an *intrinsic justification*), or we can hypothesize that welfare is correlated with a measured variable (a *proxy justification*).

I contend that the foundations of normative analysis should be built around a welfare measure for which one can offer a coherent intrinsic justification. Proxy justifications for a *foundational* welfare measure are inherently problematic. By definition, any proxy justification references a variable for which the proxy proxies. Logically, there are only two possibilities: either the proxied variable is in principle measurable (at least in some instances), or it isn't.

Consider first the possibility that the proxied variable is (sometimes) measurable. In that case, the proxied variable, not the proxy, should be treated as the foundational welfare measure around which the welfare framework is built; the proxy should be treated merely as a proxy. For instance, if neural activity proxies for choice, and if the justification for choice-based welfare measures is intrinsic, then we should build our welfare framework around choice, and not around neural activity, using neural variables as proxies for choice when choice data are unavailable or of low quality (as discussed in the next section). Conversely, if choices proxy for neural activity, and if the justification for a neural welfare measure is intrinsic, then we should build our welfare framework around neural activity, and not around choice, using choices as proxies for neural activities when neural data are unavailable or of low quality.

Next consider the possibility that the proxied variable is *not* measurable. For example, it might be some latent measure of well-being that we have no hope of accessing directly. In

that case, one cannot determine empirically whether the correlation between the proxy and true well-being is high, low, or even positive. In other words, there is no way to validate the proxy. Any welfare framework requiring the use of a proxy that is impossible to validate is conceptually flawed.

In light of the preceding discussion, it is essential to identify the types of welfare measures for which we can in principle offer intrinsic justifications. In economics, standard welfare analysis permits one to treat choice as *embodying* welfare; one can comfortably adopt the premise that it is appropriate to defer to an individual's choices because they are choices, not because they are correlated with something else. Deference to choice is, for example, a core principle among rights theorists within the libertarian tradition (see, e.g., Mill, 1869, and Nozick, 1974). Because there is a coherent intrinsic justification for using choice as a measure of welfare, one can build a conceptually sound welfare framework around a foundation based on choice.

Intrinsic justifications for the use of self-reported happiness as a measure of welfare are necessarily more awkward. Presumably, when someone is asked to report her happiness, she introspects in an attempt to assess some internal state. But in that case, self-reported happiness reflects welfare not because it intrinsically equates with welfare, but rather because it is correlated with the pertinent internal state. Thus, self-reported happiness is more naturally justified as a proxy, rather than as an intrinsic measure of welfare. These observations point to a fundamental flaw in the argument that choices and feelings have equal ethical validity as indicia of welfare (see, e.g., Kimball and Willis, 2006): while we can measure choices, we cannot measure feelings; rather, we measure *reports* of feelings, which we take as proxies for feelings (often without explicitly acknowledging them as proxies, which is where the confusion arises). One could, of course, choose to live with the awkwardness of a welfare framework that intrinsically equates self-reported happiness with welfare, but then it would be misleading to say (as happiness researchers often do) that the framework employs measures of happiness;[21] rather, it would employ measures of self-reported happiness, and

[21] For example, Kimball and Willis (2006) write: "...some economists think happiness can't be measured

happiness itself would play no role. Accordingly, I question the suitability of self-reported happiness as a foundational welfare measure.

It is difficult to imagine an intrinsic justification for the use of any particular neural activity variable (or variables) as a measure of welfare. Without some external frame of reference, it would be impossible to say whether heightened activity in a particular portion of the brain reflects pleasure, pain, or something else entirely. Regions of the brain are not etched with such labels. We associate certain types of neural activity with pleasure only because subjects who experience that activity report pleasant sensations, or are engaged in activities which we recognize as pleasurable.[22] When using neural welfare measures, we must therefore necessarily offer proxy justifications. Accordingly, I doubt that any aspect of neural activity can adequately serve as a foundational welfare measure.

It is sometimes suggested that correlations between self-reported feelings, biometric variables, and neural measurements corroborate the use of such objects as indicia of well-being (see, e.g., Larsen and Fredrickson, 1999). However, the same objections that I raised above concerning the use of a single proxy as a fundamental welfare measure apply with equal force to any measure based on a collection of proxies: either it is possible to validate the composite proxy through comparisons with some index of true well-being, in which case our normative framework should be built around the latter index, or it is not possible to validate the composite proxy, in which case it cannot provide a foundation for a compelling normative framework. If, as argued above, the only compelling candidate for an intrinsic welfare measure is choice, then we can validate a composite proxy by comparing it with choice, but the composite proxy cannot then *substitute* for choice if and when the two conflict.

---

well. *This is just not true.* Happiness (current affect) is one of the easiest of all subjective concepts to measure." On the contrary, only *reports* of happiness are easy to measure.

[22] We may identify an activity as pleasurable either through introspection, or by asking ourselves whether the subject would choose it voluntarily.

2.2 Can neuroeconomics improve choice-based welfare analysis?

In practice, the normative choice-based methods of standard economics encounter the following two difficulties. First, many individuals appear to make inconsistent choices. Indeed, much of empirical behavioral economics involves the identification of seemingly irrelevant changes in conditions that lead to choice reversals: an individual chooses option A over option B under one condition, and option B over option A under another. Typical examples include the point in time at which a choice is made (dynamic inconsistency), the manner in which information is presented, the labeling of a particular option as the status quo, or exposure to an anchor (for a survey, see Rabin, 1998). If choices are inconsistent, how can they serve as a coherent basis for making normative judgments?

Second, economists sometimes attempt to make normative statements concerning options for which no choice data are available. This problem arises most prominently in the context of environmental economics. For example, how can we put an economic value on the environmental damage caused by an oil spill? The typical consumer does not make any choices involving significant changes in the likelihood of oil spills, nor is it practical to offer such choices experimentally. One standard approach, contingent valuation, involves hypothetical questions. But the hypothetical nature of the exercise induces a potentially large bias (see, e.g., the review in List and Shogren, 2002), and answers are sensitive to the details of elicitation protocols (List et. al., 2004). How then can we reliably evaluate welfare, using choice as a foundation, when no actual choices are available?

In this section, I argue that neuroeconomics offers potential solutions to both of these problems.

2.2.1 Normative analysis when choices conflict

Some scholars have argued that evidence of inconsistent choice patterns overturns the hypothesis that choice reveals meaningful preferences based on well-defined valuations, and undermines the legitimacy of welfare judgments based on choice (e.g., Kahneman, 1999,

Ariely, Loewenstein, and Prelec, 2003). Their objection is based on the false premise that the libertarian principle requires a *rationalization* of choice (in other words, utility or preferences), and that choice-based welfare analysis must respect that rationalization, rather than choice itself. Elsewhere, Antonio Rangel and I have argued that choice-based welfare analysis requires no rationalization of behavior (Bernheim and Rangel, 2007a,b, 2008). When choice lacks a consistent rationalization, the normative guidance it provides may be ambiguous in some circumstances, but is typically unambiguous in others. As our work demonstrates, this partially ambiguous guidance provides a sufficient foundation for rigorous welfare analysis.

Formally, we have developed a framework for welfare analysis based on a binary individual welfare relation P^*, defined (informally) as follows: xP^*y iff y is never chosen (under any condition) when both x and y are available. That relation need not be either complete or transitive, but it is *always* acyclic, which suffices for welfare analysis. Interested readers can find a more complete justification for this approach, as well as properties of the binary relation, generalizations of the standard tools of applied welfare analysis, and applications to specific behavioral models, in Bernheim and Rangel [2007a,b, 2008].

When choice conflicts are severe, our framework remains applicable, but our welfare criterion may not be particularly discerning. We have therefore proposed an agenda for refining the criterion, and have identified an important potential role for neuroeconomics.

The logic of refinements Within our welfare framework, the goal of a refinement is to make the welfare criterion more discerning while adhering to the libertarian principle by *officiating* between apparent choice conflicts. In other words, if there are some situations in which option A is chosen over option B, and other situations in which option B is chosen over option A, we can look for *objective* criteria that might allow us to disregard some of these situations, and thereby reduce the ambiguity.

What might such criteria entail? Suppose the objective information available to an individual implies that he is choosing from the set X, but he believes his opportunities are $Y \neq X$. We submit that a planner should not mimic that choice. Why would the

individual believe himself to be choosing from the wrong set? His attention may focus on some small subset of X. His memory may fail to call up facts that relate choices to consequences. He may forecast the consequences of his choices incorrectly. Or he may have learned from his past experiences more slowly than the objective information would permit. Therefore, by studying the neurobiology of attention, memory, forecasting, and learning, it may be possible to identify specific conditions under which there is a significant discrepancy between the actual choice set, X, and the perceived choice set, Y.

The following simple example motivates the use of evidence from neuroscience. An individual is offered a choice between alternatives A and B. He chooses A when the alternatives are described verbally, and B when they are described partly verbally and partly in writing. Which choice is the best guide for public policy? If we learn that the information was provided in a dark room, we would be inclined to respect the choice of A, rather than the choice of B. We would reach the same conclusion if an opthamologist certified that the individual was blind, or, more interestingly, if a brain scan revealed that the individual's visual processing circuitry was impaired. In all of these cases, non-choice evidence sheds light on the likelihood that the individual successfully processed information that was in principle available to him, thereby properly identifying the choice set X.

An application: addiction My work on addiction with Antonio Rangel (Bernheim and Rangel, 2004) provides a practical application of the agenda described in the previous section. Citing evidence from neuroscience, we argue as follows. First, the brain's forecasting circuitry includes a specific neural system that measures empirical correlations between cues and potential rewards.[23] Second, the repeated use of an addictive substance causes that

[23]Recent research indicates that the mesolimbic dopamine system (MDS) functions, at least in part, as a mechanism for forecasting hedonic responses, based on environmental cues (see Schultz, Dayan, and Montague, 1997, and Schultz, 1998, 2000). The evidence points toward a temporal difference reinforcement learning (TDRL) model of the MDS. The subject's dopamine response at the presentation of the cue codes for an expectation (or forecast), while the response at the presentation of the reward codes for a surprise (the discrepancy between expectations and observation). Learning converges when there is no longer any surprise.

system to malfunction in the presence of cues that are associated with its use.[24] Whether or not that system *also* plays a role in hedonic experience, the choices made in the presence of those cues are therefore predicated on improperly processed information, and welfare evaluations should be guided by choices made under other conditions.

As an illustration, suppose that a recovering alcoholic drinks whenever he socializes with drinkers, but at other times would happily impose upon himself a binding commitment not to drink in such situations. Because those choices pertain to precisely the same actions and circumstances, there is plainly a conflict. We resolve that conflict in favor of the precommitment, on the grounds that a decision to drink taken in the presence of a cue (social interaction with drinkers) associated with the consumption of an addictive substance (alcohol) is influenced by a neural forecast that is most likely distorted due to the substance's neurobiological properties. The cue, and even the flawed forecast itself, may also have hedonic consequences, but the individual presumably considers those consequences when deciding whether to make a precommitment that would restrict his behavior contingent on exposure to the cue.

Thus, the analysis in Bernheim and Rangel [2004] serves as proof of concept for the refinement agenda proposed in Bernheim and Rangel [2007a,b, 2008]. More generally, it suggests that research on neural processes can play an important role in the analysis of a standard normative economic question.

2.2.2 Normative analysis when choice data are unavailable

Now I turn to the second issue: how can we reliably evaluate welfare, using choice as a foundation, when no actual choices are available? In ongoing work, Colin Camerer, Antonio Rangel, and I are exploring one possible solution to this problem, involving the use of

[24] There is a large and growing consensus in neuroscience that addictive substances share an ability to activate the firing of dopamine with much greater intensity and persistence than other substances (see Nestler and Malenka, 2004, Hyman and Malenka, 2001, Nestler, 2001, Wickelgreen, 1997, and Robinson and Berridge, 2003). As a result, the dopamine response occurring with the presentation of a reward (consumption of the substance) *always* registers a surprise (Di Chiara, 1999), even with experience, which implies that temporal difference reinforcement learning cannot converge (Redish, 2004). Consequently, when an addict encounters a drug-related cue, the MDS pleasure forecast is necessarily exaggerated.

neuroeconomic methods. At this stage, our work is still preliminary, so I will confine my remarks to a brief description of our agenda.

As discussed in Section 1.5, neuroeconomic methods may enable us to predict accurately the choices that people would make from any given set of prospects by measuring their neural responses to those prospects, even when no choice is offered – indeed, even if no choice is possible. One could then supplement actual choice data with these synthetic choices for the purpose of conducting normative analysis. For example, one might accurately forecast the choice that an individual would make between an environmental outcome (such as the avoidance of the environmental damage resulting from an oil spill) and various monetary payoffs, thereby associating that outcome with an economic value. The standard choice-theoretic welfare framework would be retained; one would simply use synthetic choices rather than actual choices when the latter are unavailable.

3 Conclusions

In my opinion, the potential for the emerging field of neuroeconomics to shed light on traditional economic questions has been overstated by some, unappreciated by others, and misunderstood by many. With respect to positive economics, the case for studying the neural foundations of decision-making is hardly self-evident. Certain claims, such as the suggestion that it is possible to formulate neural tests of conventional behavioral hypotheses, appear at this point to have limited merit. Nevertheless, neuroeconomics could in principle contribute to conventional positive economics in a number of ways, which I have attempted to catalog in the first portion of this paper. Because many of those potential roles are both speculative and narrow, I question whether the impact of neuroeconomics on the analysis of conventional positive economic issues is likely to be revolutionary.

I see greater potential in the area of normative economics. I do not believe that neuroeconomics will provide us with the technology to measure utility directly, and thereby ultimately replace choice-based welfare analysis with a new utilitarian paradigm. However,

I have argued that it holds the potential to improve choice-based welfare analysis in two ways. First, by shedding light on the manner in which the brain processes information, it can provide objective criteria for officiating between apparently conflicting choices. Second, it may allow us to predict the choices that people would make from any given set of prospects based on their neural responses to those prospects. Such predictions would permit us to conduct choice-based welfare analysis even when no choice is actually available.

Many neuroeconomists have been surprised and frustrated to learn that skepticism concerning their field's potential among mainstream economists runs deep. How can they combat that skepticism? First, neuroeconomists need to do a better job of articulating specific visions of the field's potential contributions to mainstream economics. Such an articulation would ideally identify a standard economic question of broad interest (e.g., how taxes affect saving), and outline a conceivable research agenda that could lead to specific, useful insights of direct relevance to that question. Vague assertions that a deeper understanding of decision-making processes will lead to better models of choice do not suffice. Second, it is essential to avoid hyperbole. Exaggerated claims simply fuel skepticism. Sober appraisals of the field's potential, including its limitations, will promote its acceptance more effectively than aggressive speculation that involves loose reasoning or otherwise strains credibility. Third, the ultimate proof is in the pudding. To convert the skeptics, neuroeconomists need to accumulate the right type of success stories – ones that illuminate conventional economic questions that attracted wide interest among economists prior to the advent of neuroeconomic research.

References

References

[1] Andreoni, James, and John Karl Scholz, "An Econometric Analysis of Charitable Giving with Interdependent Preferences," *Economic Inquiry* 36(3), July 1998, 410-428.

[2] Ariely, Dan, George Loewenstein, and Drazen Prelec, "Coherent Arbitrariness: Stable Demand Curves without Stable Preferences," *Quarterly Journal of Economics* 118(1), 2003, 73-105.

[3] Balleine, B.W., Daw, N.. and O'Doherty, J., in P. W. Glimcher, E. Fehr, C. F. Camerer, and R. A. Poldrack (eds.), *Neuroeconomics: Decision Making and the Brain*, Elsevier: New York, 2008, forthcoming.

[4] Bernheim, B. Douglas, "Taxation and Saving," in Alan Auerbach and Martin Feldstein (eds.), *Handbook of Public Economics*, Volume 3, North-Holland, 2002, 1173- 1249.

[5] Bernheim, B. Douglas, and Antonio Rangel, "Addiction and Cue-Triggered Decision Processes," *American Economic Review* 94(5), 2004, 1558-90.

[6] Bernheim, B. Douglas, and Antonio Rangel, "Beyond Revealed Preference: Choice-Theoretic Foundations for Behavioral Welfare Economics," mimeo, Stanford University, 2007a.

[7] Bernheim, B. Douglas, and Antonio Rangel, "Toward Choice-Theoretic Foundations for Behavioral Welfare Economics," *American Economic Review Papers and Proceedings* 97(2), 2007b, 464-470.

[8] Bernheim, B. Douglas, and Antonio Rangel, "Choice-Theoretic Foundations for Behavioral Welfare Economics," in Andrew Caplin and Andrew Schotter (eds.), *The Methodologies of Modern Economics*, Oxford University Press, forthcoming, 2008.

[9] Camerer, Colin F., "Neuroeconomics: Using Neuroscience to Make Economic Predictions," *Economic Journal* 117, March 2007, C26-C42.

[10] Camerer, Colin F., George Loewenstein, and Drazen Prelec, "Neuroeconomics: Why Economics Needs Brains," *Scandinavian Journal of Economics* 106(3), 2004, 555-579.

[11] Camerer, Colin F., George Loewenstein, and Drazen Prelec, "Neuroeconomics: How Neuroscience Can Inform Economics," *Journal of Economic Literature* 43, March 2005, 9-64.

[12] Carman, Katherine G., *Three Essays on Household Behavior*, Ph.D. dissertation, Stanford University, 2003.

[13] Chetty, Raj, Adam Looney, and Kory Kroft, "Salience and Taxation: Theory and Evidence," mimeo, University of California, Berkeley, 2007.

[14] Colander, D., "Neuroeconomics, the hedonimeter, and utility: some historical links," mimeo, Middlebury College, 2005.

[15] Di Chiara, Gaetano, "Drug Addiction as Dopamine-Dependent Associative Learning Disorder," *European Journal of Pharmacology* 375, June 30, 1999, 13-30.

[16] Finkelstein, Amy, "EZ-Tax: Tax Salience and Tax Rates," mimeo, MIT, 2007.

[17] Glimcher, Paul W., Michael C. Dorris, and Hannah M. Bayer, "Physiological utility theory and the neuroeconomics of choice," *Games and Economic Behavior* 52, 2005, 213–256.

[18] Glimcher, Paul W., and Aldo Rustichini, "Neuroeconomics: The Consilience of Brain and Decision," Science 306, October 15, 2004, 447-452.

[19] Gul, Faruk, and Wolfgang Pesendorfer, "The Case for Mindless Economics," mimeo, Princeton University, 2005.

[20] Harbaugh, William T., Ulrich Mayr, and Daniel R. Burghart, "Neural Responses to Taxation and Voluntary Giving Reveal Motives for Charitable Donations," *Science* 316, June 15, 2007, 1622-1625.

[21] Harbaugh, William T., Ulrich Mayr, and Dharol Tankersley, "Understanding Charitable Giving, Other Regarding Preferences, and the Moral Sentiments," in P. W. Glimcher, E. Fehr, C. F. Camerer, and R. A. Poldrack (eds.), *Neuroeconomics: Decision Making and the Brain*, Elsevier: New York, 2008, forthcoming.

[22] Hsu, M., M. Bhatt, R. Adolphs, D. Tranel, and C. F. Camerer, "Neural systems responding to degrees of uncertainty in human decision-making," *Science* 310(5754), December 9, 2005, 1680–1683.

[23] Hyman, Steven, and Robert Malenka, "Addiction and the Brain: The Neurobiology of Compulsion and Its Persistence," *Nature Reviews Neuroscience* 2, 2001, 695-703.

[24] Kahneman, Daniel, "Objective Happiness," Chapter 1 in Daniel Kahneman, Ed Diener and Norbert Schwarz (eds.), *Well-Being: The Foundations of Hedonic Psychology*, Russell Sage Foundation, New York, 1999.

[25] Kimball, Miles, and Robert Willis, "Utility and Happiness," mimeo, University of Michigan, 2006.

[26] Knutson, Brian, Scott Rick, G. Elliott Wimmer, Drazen Prelec, and George Loewenstein, "Neural Predictors of Purchases," *Neuron* 53, January 4, 2007, 147-156.

[27] Kuhnen, Camelia M., and Brian Knutson, "The Neural Basis of Financial Risk Taking," *Neuron* 47, September 1, 2005, 763-770.

[28] Larsen, Randy J., and Barbara L. Fredrickson, "Measurement Issues in Emotion Research," Chapter 3 in Daniel Kahneman, Ed Diener and Norbert Schwarz (eds.), *Well-Being: The Foundations of Hedonic Psychology*, Russell Sage Foundation, New York, 1999.

[29] List, John A., Robert P. Berrens, Alok K. Bohara, and Joe Kerkevliet, "Examining the Role of Social Isolation on Stated Preferences," *American Economic Review* 94(3), 2004, 741-752.

[30] List, John A., and Jason F. Shogren, "Calibration of Willingness-to-Accept," *Journal of Environmental Economics and Management* 43(2), 2002, 219-233.

[31] McClure, S. M., D. I. Laibson, G. Loewenstein, and J. D. Cohen, "Separate neural systems value immediate and delayed monetary rewards," *Science* 306, October 15, 2004, 503–507.

[32] Mill, John Stewart, *On Liberty*, London, UK: Longman, Roberts & Green, 1869.

[33] Nestler, E.J., "Molecular Basis of Long-term Plasticity Underlying Addiction", *Nature Reviews Neuroscience* 2, 2001, 119-28.

[34] Nestler, E. and Robert Malenka, "The Addicted Brain," *Scientific American*, March 2004, 78-85.

[35] Nozick, Robert, *Anarchy, State, and Utopia*, Basic Books, 1974.

[36] Padoa-Schioppa, Camillo, and John A. Assad, "Neurons in the orbitofrontal cortex encode economic value," *Nature* 441, May 11, 2006, 223–226.

[37] Padoa-Schioppa, Camillo, and John A. Assad, "The representation of economic value in the orbitofrontal cortex is invariant for changes of menu," *Nature Neuroscience* 11, December 9, 2007, 95-102.

[38] Platt, M. L., and P. W. Glimcher, "Neural correlates of decision variables in parietal cortex," *Nature* 400, 1999, 233–8.

[39] Rabin, Matthew, "Psychology and Economics," *Journal of Economic Literature* 36(1), 1998, 11-46.

[40] Redish, A. D., "Addiction as a Computational Process Gone Awry," *Science* 306, December 10, 2004, 1944-1947.

[41] Robinson, Terry and Kent Berridge, "Addiction," *Annual Reviews of Psychology* 54, 2003, 25-53.

[42] Rustichini, Aldo, "Neuroeconomics: Present and Future," *Games and Economic Behavior* 52, 2005, 201-212.

[43] Savage, L., *The Foundation of Statistics*, New York: John Wiley and Sons, 1954.

[44] Schultz, W., "Predictive reward signal of dopamine neurons," *Journal of Neurophysiology* 80, 1998, 1-27.

[45] Schultz, Wolfram, "Multiple Reward Signals in the Brain," *Nature Reviews Neuroscience* 1, 2000, 199-207.

[46] Schultz, W., P. Dayan, and P.R. Montague, "A neural substrate of prediction and reward," *Science* 275, 1997, 1593-99.

[47] Wang, J. T.-Y., M. Spezio, and C. F. Camerer, C. F., "Pinocchio's pupil: using eyetracking and pupil dilation to understand truth-telling and deception in biased transmission games," mimeo, Caltech, 2006.

[48] Wickelgren, Ingred, "Getting the Brain's Attention," *Science* 278, 1997, 35-37.