NBER WORKING PAPER SERIES

ON THE POTENTIAL OF NEUROECONOMICS: A CRITICAL (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.

On the Potential of Neuroeconomics: A Critical (but Hopeful) Appraisal
B. Douglas Bernheim
NBER Working Paper No. 13954
April 2008, Revised February 2009
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 (see, e.g., Rangel, 2008). 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 area of inquiry, as well as interested observers, hope neuroeconomics will also eventually make foundational contributions to the various traditional fields from which it emerged, including economics, psychology, 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 disputes. My exclusive 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 substantive (as opposed to methodological) 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 [2008]

[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 deep 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 incompletely articulated. However, because I share many of the conceptual concerns raised by Gul and Pesendorfer, I also see a pressing need for a critical and systematic articulation of the field's relevance for traditional economics. 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 open-mindedness of mainstream economists, and because it helps neuroeconomists hone research agendas that would be more useful and relevant to those who examine traditional economic questions.

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

1 Positive economic analysis of decision making

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 the basis for their skepticism, and then attempt to identify specific ways in which neuroeconomics could in principle contribute to traditional positive economic analysis of decision making.

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 manner in which an individual's choice of a vector y is causally influenced by a set of conditions, x, which the individual regards as fixed (or at least predetermined).[2] For example, y might be a vector of the quantities of various goods the individual purchases, and x might include prices and sales tax rates, as well as the individual's age, gender, and income. It is worth emphasizing that we are concerned here with the type of relationship that describes the behavior of a single individual, rather than a collection of interacting individuals; thus, this framework is not intended to encompass economic models that describe equilibria in markets or games. It does, however, subsume models that depict the choices of a single individual within such contexts, and which typically serve as the building blocks for equilibrium analysis.

For the time being, we will take x to include only the types of variables normally considered by economists. We recognize nevertheless that y depends not only on x, but also on

[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.

a set of unobservable conditions, ω, which may include the types of conditions studied by psychologists and neuroeconomists, such as psychological or neural traits, as well as other factors. We hypothesize that the causal relationship between y and the conditions, (x, ω), is governed by some function f:

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

The function f could be either a simple reduced form (e.g., a collection of demand functions, each 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 instance, it could indicate that the individual's choices maximize some objective function given the available alternatives when the conditions x and ω prevail.[3] It is important to emphasize that standard economic models of decision making, whether reduced form or structural, are typically agnostic with respect to the nature of decision *processes*. In particular, no explicit assumptions are usually made concerning the inner workings of the brain. For example, we often assume that individuals act *as if* they maximize utility, not that their brains necessarily assign actual utility values to alternatives and identify the alternative with the highest assigned value. That is why economists commonly take the position that the reduced form for a structural model of decision-making embodies all of the model's empirical implications. (I will return to that issue in Section 1.5.1.) There are notable exceptions; for example, models of cognition and reasoning in strategic environments have empirical implications concerning processes (see, e.g., Costa-Gomez and Crawford, 2006). However, as process-oriented models of decision-making are still relatively uncommon in economics, I will set them aside until Section 1.5.2.

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

[3] In the latter case, an economist would typically interpret the free parameters of the objective function as aspects of preferences. However, in modern choice theory, one can view preferences and utility functions as mere constructs that economists invent to summarize systematic behavioral patterns. From that perspective, we are of course concerned with the accurate estimation of preference parameters, but only because they imply the behavioral relation f.

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 ε. There is no opportunity to recover the form of the function $\varepsilon(\cdot)$ 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 behavior attributable to specific neural traits; rather, she would focus on either the average response (by estimating a linear model in which the tax rate appears without interactions), a characteristic-specific response (by including interactions between the tax rate and other elements of x, such as gender), or the distribution of responses (by allowing for interactions between the tax rate and unobservables, as in a random coefficients model), without conditioning on neural characteristics. Accordingly, the identification of the causal relation $\eta(A \mid x)$, where x consists of standard economic variables such as prices, taxes, incomes, age, and gender, is arguably our primary objective when we undertake traditional positive economic analysis of decision making. (Whether it should remain so in light of developments in neuroeconomics is plainly another matter; I will return to that issue shortly.)

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

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

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 function Z (the process that determines of neural activity) and Y (the neural process that determines decisions). From the perspective of a neuroeconomist, any standard economic model of decision making, along with its associated function f, is a reduced form for the underlying neural processes. Neuroeconomic research can also potentially shed light on the distribution of previously unobserved neural components of ω, which is another building block of the relation η, the object of primary interest to a traditional positive economist who studies decision making.

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 the economic analysis of decision making is practically self-evident; consequently, they are puzzled by the persistent skepticism among many 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):

> "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

[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.

> 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.[5]

Notably, economists have not materially benefited from a long-standing ability to open certain black boxes. We could have spent the last hundred years developing highly nuanced theories of production processes through the study of physics and engineering, but did not. Opening other black boxes has yielded disappointing results. For example, past efforts to understand price formation processes in competitive markets were largely unsatisfactory. Relatively few economists are now interested in that topic; most are content to leave that

[5] 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.

box closed and focus on equilibria. 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?

My frequent coauthor, Antonio Rangel, a pioneer in neuroeconomics, has defended the field's relevance to economics by arguing that the goal of all science is to explain "why." Yet *every* scientific explanation begs that question because it proceeds from assumptions for which no explanation is offered. Newton's theory of gravity explains why apples fall from trees, but it does not explain why objects produce gravitational fields. If one's limited objective is to predict the point in time at which the apple will strike the ground, one does not require (and may not even benefit from) an answer to the latter question, despite its obvious scientific importance.

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 has 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 directly relevant. What has been offered along these lines to date is too vague and insubstantial to convert the skeptics.

1.3 Some specific sources of skepticism

Some neuroeconomists have attempted to offer economists a variety of affirmative motivations for opening the black box of the human mind.[6] Many mainstream economists find those motivations unpersuasive because they see neuroeconomic inquiry as largely orthogonal to

[6] See, e.g., Glimcher and Rustichini [2004], Camerer, Loewenstein, and Prelec [2004, 2005], Rustichini [2005], Glimcher, Dorris, and Bayer [2005], and Camerer [2007]

traditional economic analysis, a view that finds its most forceful articulation in the work of Gul and Pesendorfer [2008]. To identify motivations that economists would generally find persuasive, one must directly confront the logic of that view.

For the rest of this subsection, let us accept the premise that the primary historical objective of positive economics (as it pertains to individual decision making) is to recover causal behavioral relationships of the form $\eta(\cdot \mid x)$, where y describes conventional economic decisions and x is a vector of standard economic variables, such as prices, tax rates, gender, age, income, etc. (I will reexamine that premise shortly.) 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, though the functions Y and Z are obviously interesting, the questions they address directly are not ones that mainstream economists traditionally examine.

Second, because we have assumed (for the purposes of the discussion in this subsection) that the behavioral relation η contains only conventional economic variables, a traditional economist could in principle divine all of its properties from standard economic data. Distinguishing between two neural processes, (Y, Z, μ) and (Y', Z', μ'), advances the limited objective specified above 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 might reasonably hope to test between η and η' directly using conventional 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. With respect to the third consideration, even an imperfect understanding of neural processes may offer viable strategies for improving our ability to recover behavioral relations involving standard economic variables; I will describe specific possibilities in the ensuing subsections. Finally, the scope of mainstream economics (as it relates to individual decision making) is not strictly confined to the limited objective stated at the outset of this discussion. It is worth considering the possibility that neuroeconomics will allow us to accomplish other objectives and answer new questions that mainstream economists would recognize as falling within their field's boundaries.

The remainder of this section examines a number of routes through which neuroeco-

nomics could potentially contribute to standard positive economics. First, neuroeconomics will lead to the measurement of new variables, some of which may usefully find their way into otherwise conventional economic analyses. I discuss that possibility in Section 1.4. 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 1.5, the formulation of rigorous tests may prove challenging. With some notable exceptions, 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 1.6 and 1.7 examine the more modest possibility that an understanding of neural processes may provide economists with informal but nevertheless useful guidance with respect to model selection.

While I see the potential for neuroeconomics to enrich standard positive economics, I nevertheless doubt that it will revolutionize the practice of our discipline within its traditional scope. As I see it, the paths that hold the greatest promise for success involve the augmentation rather than the replacement of existing economic methods. For example, neural variables may well find their way into otherwise standard econometric analysis. That development would not necessarily require an extensive knowledge of neuroeconomic methods or a deep appreciation of neural processes; instead, an economist might simply rely on neuroscientists 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 that information, we might be able to learn something useful about the relationship between marketing and purchases. But once we have identified the implied restrictions on the latter relationship, a positive economist arguably gains nothing of value from knowledge of its neural underpinnings.[7]

[7]The manner in which marketing influences attention is, however, potentially of interest to a *normative*

1.4 Are there potential uses for neuroeconomic variables?

Neural variables play two distinct roles in the framework of Section 1.1. Some specify exogenous (or at least preexisting) neural conditions under which a decision is made, and appear as elements of ω. Others describe endogenous neural reactions to the presentation of a decision problem, and appear as elements of z. Exogenous neuroeconomic variables include neural traits that remain fixed throughout a time window that encompasses the contemplation and execution of a particular decision problem. Such traits could reflect genetics and/or the effects of environmental factors. One existing method of assessing neural traits involves the measurement of activity in the resting brain (see, e.g., Gianotti et al., 2008). Other methods might include the identification of genetic markers that influence neural architecture, or the measurement of subtle variations in brain structure.[8] In contrast, endogenous neuroeconomic variables typically measure aspects of neural activity during the process of decision making. The use of variables from these two classes raise different issues, and I will discuss them in separate subsections.

1.4.1 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 economic analyses of decision making. 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 $\eta^0(\cdot \mid x^0)$.

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.

[8] Traditionally, neuroscientists have relied on gross variations in brain structure, such as brain lesions resulting from accidents; see, e.g., Damasio and Damasio [1989].

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 neural trait that helps predict saving (a "patience marker"). Should mainstream economists greet that discovery with enthusiasm? Economics has not, after all, concerned itself historically with the relationship between neural attributes 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. Knowing that a tax policy has different effects on different demographic groups is potentially of interest, as is the measurement of the degree to which those effects vary within each demographic group; in both cases, the hypothesized finding sheds light on traditional questions about incidence and efficiency. In contrast, knowing that the within-group variance in the policy's effect is linked to a specific neural trait does not illuminate a traditional economic question. In principle, one could compute incidence and efficiency effects across neurally differentiated groups, but to what end? (One can come up with reasons for making such a calculation – I mention one speculative possibility below – but I have not yet encountered a reason that

I find compelling.)

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 poorly measured or 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 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. The use of such variables would fall squarely within the scope of conventional economics; mainstream economists seek and use proxies for hard-to-measure conditions as a matter of course. 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 marker," 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 marker" could allow us to detect the presence of peer effects by studying the relationship between an individual's giving and the neural charitable predispositions of his peers, controlling for his own predisposition.[9] 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, which was suggested to me by Antonio Rangel. Suppose 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.

[9] 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 neural charitable predisposition.

Then by collecting neural measurements for a sample of traders, one might be able to forecast short-term movements of the stock market. A skeptic might object that an idiosyncratic shock to a single investor's neural state would have no effect on the market, and any market-wide shock to investors' neural states presumably has some observable cause (such as a news report), which can then serve equally well as a predictor. But the relationship between observable events and neural reactions may be complex and difficult to model. For example, a given event may cause different levels of anxiety depending on its presentation (which is subjective) and prevailing expectations (which are tricky to measure). A forecaster armed with neural measurements might therefore perform better than one who closely monitors the news.

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 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 gender and race, are of course simply aspects of genetics. To improve the accuracy of the overall forecast for the second population, one could add other pertinent biological 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 neural traits 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 marker" 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. Neuroeconomic methods may have commercial applications in the private sector. Some neuroeconomists project a near-term future in which employers subject job applicants to tests that evaluate genetic and/or neural 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*, Twentieth Century-Fox Film Corporation, 2002). To describe and analyze resource allocation in light of such developments, economists would need to consider the behavioral roles of the pertinent neural variables.

In principle, governments could also make use of information on citizens' neural characteristics. Indeed, conventional economics offers potential justifications for that practice. For example, if the effects of a tax vary systematically with some neural trait within demographic groups (a possibility mentioned above), then the unrestricted solution to a standard optimal tax problem will involve trait-specific tax rates. Should the government contemplate such policies, it will behoove economists to study behavioral relationships involving neural traits.

In practice, ethical and political concerns will likely preclude any serious consideration of policies that discriminate based on neural characteristics, just as they currently 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. Such restrictions would obviously impact resource allocation. To understand

and evaluate that impact, economists would need to understand the *potential* roles of neural measurement in the private and public sectors.

1.4.2 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. Existing findings concerning Y suggest that it may be possible to predict certain choices from particular types of endogenous neural activity. For example, activity in the nucleus accumbens, the insula, and/or the mesial prefrontal cortex correlates with purchase decisions (Knutson, 2007) and risk-taking (Kuhnen and Knutson, 2005), while right orbital frontal cortex activation in response to ambiguity correlates with 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 which grocery store customers 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 demured 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 studies that establish correlations between endogenous brain activity and choices. 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 current 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 prices, taxes, income, gender, age, 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 choices. The estimation of choice mappings (e.g., demand curves) undeniably falls within the province of mainstream economics. However, our traditional reliance on naturally occurring choice data presents at least two problems, one practical and one conceptual.

First, empirical economists are often constrained by the quality and availability of choice data. One common problem is that such data are not generally available for samples in which environmental conditions are randomly assigned. Consequently, causal interpretations of estimated behavioral relationships are often controversial. In other contexts, such as when a firm introduces a new product, choice data are simply unavailable.

Second, 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 additional hypotheses. Typically, we assume that the relationship manifests some degree of stability across distinct moments in time and/or potentially diverse individuals. However,

except in special cases, those assumptions are untestable, and without them not even the distribution of choice mappings – let alone the choice mapping that an individual would follow at a moment in time – is identified.[10]

To some extent, one can address these problems by conducting appropriate choice experiments. In principle, one can design an experiment with random assignment of environmental conditions to collect data on any feasible choice. In addition, the strategy method arguably permits the elicitation of an individual's choice mapping at a moment in time, rather than a single choice. However, choice experiments raise new difficulties. The experimental replication of significant real-world decisions can be prohibitively costly. In addition, the legitimacy of the strategy method depends on implicit assumptions, and evidence concerning its validity is mixed (see, e.g., Oosterbeek, Sloof, and van-de-Kuilen, 2004).

Were it possible to gather accurate data by posing hypothetical choices, one could easily and inexpensively identify any individual's choice mapping within any set of potential alternatives, without invoking additional assumptions or raising concerns about non-random environmental conditions. Indeed, marketing specialists often attempt to estimate demand curves for new products based on answers to hypothetical questions. Unfortunately, such data are known to be inaccurate within many domains, and economists are therefore rightly suspicious of their use.

Neuroeconomic methods can potentially address all of the limitations discussed above by, in effect, allowing us to treat rich collections of decision problems hypothetically, while nevertheless imputing predictively accurate choices. Specifically, it is natural to hypothesize that an individual's choice among a set of alternatives bears a stable relation to the neural responses they generate as she considers them. If that hypothesis proves correct, appropriate neural measurements may permit us to predict with accuracy which alternative she would choose even when no choice is offered at the time of measurement.[11] Jezewski et al. [2009] investigate that possibility. Their findings suggest that it may indeed possible to recover

[10] These assertions are based on work in progress, which I am conducting with Debraj Ray.

[11] One can validate such predictions by offering choices unexpectedly.

the complete choice mapping that an individual would follow at a moment in time, at least for relatively small sets of potential alternatives.[12] From such information, one could, for example, construct the demand curve of a particular individual at a specific moment, without relying on potentially objectionable assumptions (aside from predictive accuracy, which is testable) or raising concerns about non-random assignment.

Neuroeconomic methods are of course far less convenient and more costly than batteries of hypothetical questions. Even so, it is likely that technological progress will reduce costs and improve convenience considerably. Such methods could therefore provide a rich source of imputed choice data, which in principle could augment or substitute for actual choices in otherwise standard economic analysis, leading to more accurate behavioral estimates and more powerful tests of consumer theory.

Presumably, neuroeconomic research will also identify limits on the reliability of neurally imputed choices. For example, predictions may become less accurate when an individual is presented with a large number of prospects. 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 complex. All of those possibilities bear careful investigation.

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, we 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 con-

[12]Specifically, the study shows that neural responses to the presentation of various prospects accurately predict the subsequent choices that subjects make from any subset of those prospects, even when subjects do not anticipate that they will confront a choice. The experiment is limited to small sets of relatively simple items, and the generality of the result is not yet known.

ceivably draw reliable inferences about private information without observing it directly. The technical feasibility of that agenda has already been established. Specifically, Wang et al. (2006) conducted an experimental study of a "biased transmission game" involving cheap talk, in which one party (the "sender") observes a state of the world and transmits a message to another party (the "receiver") who then makes a decision. Payoffs are structured so that the sender wishes to mislead the receiver as to the true state. Statistical 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 addition to facilitating improved tests of economic theories in which private information plays a role, such techniques may also have practical applications. Take the classic problem of mechanism design. As is well-known, private information can stand in the way of achieving first-best outcomes. Yet as Krajbich et al. [2009] have shown (both theoretically and experimentally), one can potentially overcome those barriers by designing mechanisms that determine outcomes as a function of both announcements and neural proxies for private information, even when those proxies are imperfect.

Unfortunately, the neuroeconomics community has yet to produce a useful non-experimental application involving the detection of private information. 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 may be challenging to design an application that is both useful and feasible.

1.5 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 tests of both standard and nonstandard (behavioral) theories of decision making (see, e.g., Camerer, 2007). The least controversial possibilities involve choice-oriented tests

based on patterns of choices or choice proxies. Such a test might exploit the availability of a new instrumental variable (as discussed in Section 1.4.1). Alternatively, if we discover that particular neural responses reliably predict choices from sets of alternatives even when no choice is offered at the time of measurement (as discussed in Section 1.4.2), we could perform tests using data on neural responses rather than actual choices, effectively treating the former as a proxy for the latter. In both cases, neuroeconomics would simply serve as a tool for measuring important economic variables; given those variable, economists would continue to apply conventional methods.

Some neuroeconomists have raised the more revolutionary possibility that, by peering into the brain, we may discover how to falsify or validate theories of choice using information concerning the structure of the neural processes behind choice.[13] This section examines the prospects for such process-oriented tests. I distinguish between two classes of economic theories: those that invoke assumptions about choice patterns, and those that invoke assumptions about thought processes. Process-oriented neural evidence is sometimes cited as supporting or undermining theories belonging to the first category, but the underlying arguments tend to involve imprecise reasoning and conceptual leaps of faith, which undermines the credibility of the agenda. Conceptually legitimate tests are conceivable, but the task of laying a rigorous foundation for such a test is considerably more challenging than some have suggested. The potential relevance of process-oriented neural evidence for theories belonging to the second category is more readily apparent.

1.5.1 Testing theories that invoke assumptions about choice patterns

With some notable exceptions (which I discuss in the next subsection), most economic theories of decision making invoke assumptions about choice patterns, and are agnostic with

[13] In light of recent technological advances, the prospects for uncovering the neural structure of the brain's decision-making apparatus have noticably improved. For example, using either microstimulation to induce activity within a localized neural structure (a technique applied extensively in primate research, e.g., Hanks, Ditterich, and Shadlen, 2006), or Transcranial Magnetic Stimulation (TMS) to temporarily shut down that activity, in effect simulating a lesion (e.g., as in Camus et al., 2008), neuroscientists can determine the causal role which that localized activity plays in decision making.

respect to the nature of decision processes. No explicit assumptions are made concerning the inner workings of the brain. As an illustration, consider the Weak Axiom of Revealed Preference (WARP), which lies at the core of standard decision theory. WARP holds that an individual who chooses x but not y from a constraint set $X \supset \{x, y\}$ will not choose y from any other constraint set containing x. WARP is obviously a statement about choices; it involves no assumption about the mechanical characteristics of the decision making procedure. More generally, our disciplinary agnosticism with respect to process accounts for Gul and Pesendorfer's [2008] contention that neural evidence cannot shed light on standard economic hypotheses.

Of course, a choice axiom cannot hold unless the neural processes that govern choice are capable of delivering decisions that conform to the axiom; 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, one would be able test any choice axiom. Obviously, by executing the algorithm for a variety of choice problems, one could generate artificial choice data and perform conventional choice-oriented tests. Alternatively, without actually executing the algorithm, one might logically deduce from its properties whether it is consistent with the axiom, thereby performing a process-oriented test.

In practice, neuroeconomics is very far 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 extremely simple tasks and functions. Yet even partial information concerning the general decision algorithm might suffice for our purposes. Knowing only that the algorithm belongs to a certain class, we might nevertheless be able to deduce either that it must generate choices that conform to a given axiom, or that it cannot do so. While that possibility is conceptually coherent, laying the necessary groundwork for such a test is likely to prove extremely challenging.

To illustrate the difficulties, consider the textbook model of economic decision making, which implies that people behave as if they maximize utility. (I will focus on this example to keep the discussion concrete; neural tests of other decision theories encounter similar problems.) Emerging evidence supports a neural maximization (NM) hypothesis, which holds that an individual always chooses the alternative that elicits the highest level of activity within certain brain structures during the process of deliberation (see, e.g., Padoa-Schioppa and Assad, 2006, Plassmann, O'Doherty, and Rangel, 2007, Kable and Glimcher, 2007, Tom et al., 2007, or Hare et al., 2008). Superficially, support for the NM hypothesis may appear to vindicate the textbook model by justifying a *literal* interpretation of utility maximization.[14] However, because the NM hypothesis is consistent with any conceivable choice pattern, supportive evidence sheds no light on the validity of WARP, and consequently neither bolsters nor undermines the textbook model. Moreover, even if the NM hypothesis proves false, an individual might nevertheless employ a choice algorithm that respects WARP, in which case the textbook model would be valid. For example, she might make choices by sequentially eliminating alternatives based on checklists of desirable properties, a possibility suggested by various psychologists, including Tversky [1969], Bereby-Meyer, Assor and Katz [2004], Brandstätter, Gigerenzer and Hertwig [2006], and Katsikopoulos and Martignon [2006]. As shown by Mandler, Manzini, and Mariotti [2008], such procedures satisfy WARP; indeed, they can rationalize virtually the same set of choice patterns as utility maximization. Therefore, even if the individual has no recognizable preferences, her behavior may still conform to the textbook model. Because the NM hypothesis is neither necessary nor sufficient for the validity of the textbook model, it cannot provide the basis for a formal test, at least by itself.

The possibility remains that the NM hypothesis sets the stage for legitimate tests of the textbook model and other economic theories. Consider, for example, the following

[14] For example, Kable and Glimcher [2007] write: "From an economic perspective, these findings indicate the existence of a neural valuation process that is strikingly similar to the representations of subjective values that are employed in revealed preference theories."

neural version of "independence of irrelevant alternatives" (henceforth labeled NIIA): the ordering of any two alternatives according to the level of activity they elicit in the pertinent brain structures during the process of deliberation is unaffected by the availability of other alternatives. Plainly, the NM and NIIA hypotheses together imply WARP. However, it is important to bear in mind that neither condition is *necessary* for WARP. Indeed, even if we treat neural maximization as a maintained hypothesis, WARP may be satisfied even if NIIA is not. WARP permits context-specific reversals of the neural ordering provided they are always confined to unchosen alternatives. Inconsequential reversals might occur, for example, if the brain attends to alternatives closely only if they appear to be serious candidates in light of the opportunity set. When an alternative is plainly dominated within an opportunity set, the level of neural activation it elicits during deliberation might then be a noisy measure of "subjective value," and thus the neural ordering of such alternatives might be inconsequentially context-dependent.

Still, one might hope to bolster the textbook model by demonstrating that the NM and NIIA hypotheses are both valid. The simplest approach is to measure activation in the pertinent brain structures during deliberation for a collection of choice problems and check for violations of those hypotheses. But that approach raises an obvious question: why not skip the neural measurements and just examine the experimental choices to determine whether they satisfy WARP? As long as the NM hypothesis holds, there will be consequential reversals of the neural ordering if and only if the observed choices exhibit violations of WARP. Thus, with respect to testing WARP, evidence concerning consequential neural reversals simply reiterates what we already know from the choices.

Some might argue that experimental confirmation of NIIA (specifically, the absence of both consequential and inconsequential neural reversals) would provide insight into the structure of neural responses, and thereby justify greater subjective confidence that WARP will hold outside the test sample. But if choices satisfy WARP within the test sample, then obviously the neural responses associated with those particular choices must satisfy *some*

condition that is sufficient for WARP. Should we have more faith in WARP outside the test sample simply because the sufficient neural condition for WARP that is satisfied within that sample happens to be NIIA (in combination with NM), rather than something else?

An alternative approach is to develop tools for identifying and testing hypotheses concerning features of the neural decision-making circuitry that have specific implications for the pattern of brain activation during decision making. In combination with NM, confirmation of such a hypothesis would permit one to draw general inferences concerning choices. To illustrate, suppose we deduce from the physical wiring of the brain that activation in response to any alternative x will necessarily be given by a relationship of the form $f(u(x,\theta), X \cup X^H)$, where X is the current opportunity set, X^H is the union of recent opportunity sets, θ summarizes other neural preconditions pertinent to choice (such as hunger), and f and u are scalar-valued functions with f monotonically increasing in u. The hypothesized finding along with NM would imply that the individual maximizes $u(x,\theta)$, with currently and recently available alternatives normalizing the neural scale. WARP would follow immediately, and with generality. The available evidence suggests that responses in the pertinent brain structures may indeed take the hypothesized form: those responses depend on X and X^H (Tremblay and Schultz [1999]), but not on which alternatives are currently available when $X \subset X^H$ (Padoa-Schioppa and Assad [2008]). Whether it is feasible to establish the generality of those properties (and/or others) by tying them to specific features of the neural architecture that generates signals in the pertinent brain structures (as demanded by the agenda outlined in this paragraph) remains to be seen.

Having discussed a particular example at some length, it is useful to describe what a neural test of an "as-if" choice theory would entail more abstractly and generally. Each possible neural architecture (or decision process), n, implements a particular computational algorithm, $a = A(n)$. In turn, every possible computational algorithm, a, implements a particular choice correspondence, $c = H(a)$. 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 WARP), we would need to identify testable features of the set of neural architectures, N_x, that generate choice correspondences in C_x (formally, $N_x = \{n \mid H(A(n)) \in C_x\}$). 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 $E \cap N_x = \varnothing$, and fail to reject it if $E \cap N_x$ is non-empty. 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.

For any x, characterizing the set N_x is likely to prove extremely challenging. However, matters are not hopeless. Neuroeconomics will presumably progress by formulating and testing increasingly specific theories of neural decision processes. We can express any such theory in the form $n \in T$, where T is some set of neural architectures (for example, those conforming to the NM hypothesis). Even if the set N_x is analytically intractable, we may be able to usefully characterize the set $T \cap N_x$. 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.

The preceding discussion underscores the general proposition that a valid process-oriented neural tests of any choice axiom will require a thorough understanding of the relationships between the properties of choice mappings, the characteristics of decision algorithms, and the features of neural architectures. Fundamentally, the problem is one of identification: we are asking whether it is possible to recover specific information concerning the choice mapping from particular observations concerning the neural architecture. As a general matter, the problem of identification is one that economists have studied extensively and are methodologically equipped to address. The standards for identification in neuroeconomics should be no different than in other areas of economics. It is therefore disconcerting that some neuroeconomists have proceeded (at least implicitly) as if the foundations for neural tests of economic hypotheses are either obvious or easily motivated. 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.[15] The paper is sometimes interpreted as providing a neural test of the popular β-δ model of quasihyperbolic discounting. That interpretation is inappropriate. Even assuming that the observed neural activity encodes subjective valuations and that those valuations govern decision making (hypotheses that are not addressed by this particular experiment), the evidence does not rule out the possibility that the valuations are time-consistent, or 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. 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 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 discussion in McClure et al. [2004] is not completely clear on this point,[16] a friendly reading of the paper suggests that the authors had in mind a more reasonable

[15] In a subsequent study, Kable and Glimcher [2007] reached different conclusions. However, my focus here is not on the neural evidence *per se*, but rather on the inferences concerning economic theories of choice that one can validly draw from hypotheses involving neural processes (regardless of whether those hypotheses ultimately prove to be right or wrong).

[16] 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.

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 mandatory 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 than mandatory transfers (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 voluntary self-sacrifice (in the case of the second). Unfortunately, that interpretation is problematic. Even assuming that the measured neural activity cleanly codes for "utility" (which is by no means apparent from the experiment), the "test" is fundamentally flawed because it reflects confusion concerning the nature of pure altruism and warm glow giving as *economic* hypotheses.

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 when 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). Even if we provisionally accept the premise that those hypotheses are meaningful and testable,[17] they have no positive implications concerning choice (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 over some range with the charity's total budget held fixed. 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 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*, larger voluntary contributions lead to greater satisfaction over some range when the charity's budget varies dollar-for-dollar with the contribution. 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

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

(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.

1.5.2 Testing theories that involve assumptions about thought processes

Within the subfield of behavioral economics, some theories of decision making invoke assumptions about thought processes – particularly regarding the types of factual information considered or ignored, or the manner in which it is processed – rather than choice patterns.[18] Leading examples include theories of reasoning in games, as well as models of memory, attention, and learning. While one could argue that those theories admit "as-if" interpretations, in most cases their positive implications seem difficult to justify if not far-fetched unless one takes the thought-process assumptions literally.[19] Consequently, formal tests of those assumptions have a legitimate and potentially important place in positive behavioral economics.

Recent experimental work by Costa-Gomez and Crawford [2006] exemplifies this agenda (see also Chong, Camerer, and Ho, 2004). Behavioral game theorists have proposed several competing models of strategic reasoning to account for the prevalence of non-equilibrium choices among subjects who have no previous experience playing particular games. Leading candidates include "level k thinking" (Lk) and "level k dominance" (Dk).[20] Costa-Gomez

[18] According to Sen [1973], even the standard textbook model of decision making involves thought-process assumptions: "the rationale of the revealed preference approach lies in the assumption of revelation and not in doing away with the notion of underlying preferences, despite occasional noises to the contrary." Sen reasoned that the consistency axioms upon which economists usually rely have no force unless one assumes that preferences actually drive choices. His argument ignores the fact that certain classes of mechanistic decision processes can generate choice correspondences that satisfy the same consistency axioms, as noted in Section 1.5.1.

[19] In addition, thought-process assumptions sometimes have important normative implications. For example, one might evaluate welfare differently depending on whether an individual actually ignores certain types of information, or simply acts as if she ignores that information. I turn to normative issues in Section 2.

[20] In a two-player game, a level 1 thinker (L1) best-responds to a uniform prior over her partner's decisions; a level k thinker (Lk) best-responds to a level k-1 thinker. Level k dominance (Dk) implies that a player performs k rounds of deleting dominated strategies and then best-responds to a uniform prior over her

and Crawford categorize subjects according to those theories (and a few others) by studying each subject's choices and information search patterns across a collection of sixteen two-person guessing games. Each game is characterized by six parameters: an upper and lower bound on each player's choice, and two scalars that define each player's target as a multiple of the other's choice. The games are chosen to generate a high degree of differentiation between the choice patterns predicted by each theory.

Using a computer interface known as Mouselab, Costa-Gomez and Crawford observed the order in which subjects looked up the parameters of each game while deliberating. Those search patterns are relevant because each theory has a natural procedural interpretation. For example, it is natural to assume that an L1 player will look up parameters in the following order: her opponent's bounds, her own target parameter, her own bounds. In contrast, an L2 player will follow a different search pattern: her own bounds, her opponent's target parameter, her opponent's bounds, her own target parameter (at which point she may also remind herself of her own bounds). Costa-Gomez and Crawford treat those procedural assumptions as maintained hypotheses.[21] They classify subjects according to the theory that most likely accounts for the observed data – either their choices alone, or both their choices and search patterns, allowing for noise in both. Consideration of the data on information search patterns leads them to reclassify a sizable group of subjects.[22]

Of course, one can acknowledge Mouselab's value while doubting the usefulness of brain scans and biometric measurements. The point of the preceding discussion is simply to establish that one can test certain economic theories of decision making using evidence on thought processes, either instead of or in addition to choice patterns. If, as Crawford [2008] poetically suggests, we turn to Mouselab because it allows us to peer through "windows of

opponent's remaining alternatives.

[21] It is therefore worth noting that a player with a sufficiently good memory could look up a game's parameters in any order irrespective of the model of reasoning that accounts for his behavior. Thus, in any given instance, Costa-Gomez and Crawford could discard one theory of strategic reasoning in favor of another because their procedural assumption, and not the first theory, is incorrect.

[22] Because their method does not generate measures of statistical confidence in the subjects' classifications, they do not provide a formal test of any theory (that is, a probability statement concerning its validity) for any particular subject.

the strategic soul," then it makes sense to seek larger and clearer windows. Whether more advanced technologies will permit us to gather pertinent information on thought processes, particularly regarding the processing of factual information, certainly remains to be seen, but that possibility cannot be dismissed.

Recent work by Coricelli and Nagel [2008] hints at the new "windows" that more advanced technique may open. Their study shows that, relative to subjects who are classified as low-level strategic thinkers according to their behavior, those who are classified as high-level strategic thinkers have considerably greater activation in areas of the brain that tend to engage when a subject is instructed to think about a situation from another person's perspective (specifically, the dorsal and ventral medial prefrontal cortex, abbreviated mPFC). Certainly, that evidence provides no more than a hint concerning thought processes. Based on what we currently know, we cannot rule out the possibility that a subject might engage in high-level strategic reasoning with unremarkable mPFC activation, or in low-level strategic thinking despite pronounced mPFC activation. Consequently, Coricelli and Nagel's evidence does not provide an adequate foundation for testing formally between theories of strategic reasoning – at least not without strong maintained hypothesis in which we may have very little a priori confidence.[23] Still, it is a beginning. A neural counterpart to Mouselab (for example, a scanning procedure that registers the use of particular factual information) would have obvious value for economists seeking to test theories of strategic reasoning; that said, no such counterpart currently exists, nor can we be confident that it will exist in the foreseeable future.

Theories that invoke assumptions about thought processes are not confined to pure strategic reasoning. Other examples include models of memory (such as Mullainathan, 2002), attention to information (such as Akerlof and Dickens, 1982), and learning (such as Erev and Roth, 1998). Non-choice evidence concerning the validity of various underlying thought

[23] On the other hand, as long as we have some reason to credit those maintained hypotheses, the type of neural observations studied by Coricelli and Nagel are potentially useful for refining (or enhancing confidence in) subjects' classifications, and hence for forecasting choices out of sample.

process assumptions already exists, and indeed is commonly cited as justification for those assumptions. Neural methods can in principle contribute additional evidence. For example, Mullainathan cites Kandel, Schwartz, and Jessell [1991] to provide a neurobiological basis for his assumption concerning the importance of rehearsal in memory. Similarly, the hypothesis that choice depends upon particular information could be falsified by evidence demonstrating the absence of neural responses upon presentation of that information. Regardless of whether useful tests prove difficult to formulate and implement in practice, there is nothing conceptually wrong with the agenda.

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

The number of empirical models 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 Section 1.4) or an independent foundation for testing one model against another (the topic of Section 1.5), 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 selection and the choice of functional form.

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 descriptive and predictive power of behavioral models that incorporate related variables (e.g., measures of potential externalities). 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 motivate the inclusion of an important explanatory variable that empirical economists have otherwise ignored, a neuroeconomist would literally have to come across some significant and *unexpected* environmental correlate of brain activity. I do not dismiss that possibility, but neither does it alone convince me that the field holds great potential for conventional positive economic analyses of decision making.

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.

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.

1.7 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 hope 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 and highly differentiated 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. As long as 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 ways 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.[24] If, for example, secondary (and normally quiescent) neural systems override a primary system whenever the latter would generate behavior too far from the individual's norm, then an incomplete neural model of choice (specifically, one that omits the secondary systems) might be less stable out of sample than a behavioral model. By way of analogy, structural models of reasoning in games have notoriously performed poorly out of sample, apparently because reasoning is highly context-specific. 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

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

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.

As an example, consider once again the findings of Coricelli and Nagel [2008], discussed in Section 1.5.2. Suppose a subject's observed choices are consistent with both L1 and L2 thinking (because the within-sample games were not chosen to differentiate between those theories), but that we can rule out other models of strategic reasoning. Now imagine we are asked to predict the subject's choices out of sample in games for which the L1 and L2 models have different implications. Choices alone provide no basis for preferring one model over the other. But suppose a brain scan taken during within-sample deliberation reveals a high level of mPFC activity. That finding would not *prove* the individual engaged in high level strategic thinking within sample, let alone that he would do so out of sample in clearly differentiated settings. Even so, we would presumably take the previously documented correlation between mPFC activity and high-level strategic reasoning into account when forecasting his choices, and place greater weight on predictions derived from the L2 model.

All 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.

1.8 An overall assessment

In pondering the future of neuroeconomics, I see substantial likelihood that the field will make intellectually legitimate contributions to the positive economic analysis of decision making. While there is reason to hope that some of the contributions will prove noteworthy, at this point in time neither the field's actual accomplishments nor my speculative musings persuade me that it 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; I hope I am proven wrong.

2 Normative economics

Any contribution of neuroeconomics to normative economics would presumably take one of two forms. First, neuroeconomics might play a role in the development of an entirely new approach to measuring an individual's welfare, one that evaluates her well-being based at least in part 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? Even allowing for technological advances, I see fundamental conceptual problems with that agenda.

The central premise for "neural utilitarianism" – that welfare is reducible to neural impulses – has disconcerting implications. Picture a world in which a malevolent authority surreptitiously attaches everyone to a machine that replicates the full spectrum of neural activity associated with a lifetime of highly pleasurable and satisfying experiences while exploiting them for a dark purpose. I wager that this prospect is almost universally abhorrent; as far as I know, no one roots for the machines in the film *The Matrix* (Warner Bros. and Village Roadshow Pictures, 1999). Though such nightmare scenarios are the stuff of science

fiction,[25] they make the point that any compelling notion of welfare must encompass more than neural impulses. We often consider ourselves better off when we have actual autonomy, liberty, and a firm grasp on reality even if, as a consequence, we must relinquish appealing illusions and experience less pleasurable neurobiological sensations. Consideration of those intangibles leads to a welfare paradigm based on informed choice rather than brain activity.

If despite my squeamishness we are nevertheless determined to construct a neural welfare index, we must first confront and overcome at least five conceptual and practical challenges. The first challenge is to identify regions of the brain that generate welfare-relevant neural responses. Unfortunately, those regions are not etched with functional labels. To proceed, we require criteria for determining whether activity in a particular portion of the brain reflects pleasure, pain, or something else entirely. I will return to that central issue shortly.

The second challenge is to assure ourselves that we have not neglected any brain regions in which neural activity codes for important aspects of well-being. Even if we can prove that certain types of neural activity are welfare-relevant (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 reflects any other aspect of well-being is likely to prove far less tractable. There is a risk that a neural welfare index will focus disproportionately on the welfare-relevant signals that are most easily identifiable and measurable. Economists might unitentionally overemphasize certain aspects of well-being (possibly physical sensations) while underemphasizing others (possibly abstract emotions).

The third challenge is to filter out spurious signals within brain regions that manifest welfare-relevant neural activity. The neural circuitry that registers welfare-relevant signals may also be involved in other cognitive functions, in which case neural measures of well-being may contain spurious components. One can of course treat the resulting noise as measurement error, but the signal-to-noise ratio may be low and the noise may be non-random (for example, it may be systematically related to features of the environment, such

[25] At least, that is what the machines would like us to believe.

as complexity).

The fourth challenge is to devise an objective criterion for aggregating welfare-relevant neural signals in different brain regions into a single index of well-being. 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 seems 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 fifth challenge is to devise an objective criterion for aggregating welfare-relevant neural signals over time. Suppose 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 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 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?

The first challenge is the most fundamental but in some ways least problematic. To identify regions of the brain that generate welfare-relevant neural responses, we must of necessity ultimately rely on correlations with directly interpretable indicators of welfare – specifically, either choices or expressions of well-being (such as self-reported happiness, life satisfaction, and related concepts, all of which I will henceforth subsume under the heading of "happiness" for the sake of brevity). In other words, we classify a particular type of brain

activity as pleasurable only because people consistently choose alternatives that generate it, or because they report pleasure while experiencing it. Accordingly, while I can imagine that neural methods may eventually build upon choice-based and/or happiness-based welfare analysis, I do not see how they could spawn an entirely new and conceptually independent approach to welfare. For the rest of this section, I will focus on ways in which neural methods might build upon happiness-based welfare analysis; I discuss potential contributions to choice-based welfare analysis in Section 2.2.

In my view, self-reported happiness does not by itself provide an adequate foundation for rigorous welfare analysis. Though a full discussion of my objections to happiness-based welfare analysis is beyond the scope of this paper, I offer the following example (and refer the interested reader to Bernheim [2009] for a more complete discussion). When we ask people to express their happiness using a unitless scale (e.g., one to seven), we compel them to invent their own normalizations; they decide for themselves what each potential response signifies. The chosen normalization may well depend on aspects of the environment. For example, people may use the mid-point of the scale to denote a "typical" state of happiness within some context-specific domain. Once we admit the natural possibility that people use context-specific normalizations, identifying changes in underlying happiness from data on self-reported happiness becomes highly problematic.

Contributors to the happiness literature have informally suggested that we can validate and/or enhance the reliability of happiness analysis by integrating neural data. For example, they argue 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). But that argument is circular; it demonstrates only that the variables in question have something in common, not that they individually or collectively reflect true well-being.

For my part, I doubt that neural methods can build upon happiness analysis in a way that leads to a viable and rigorous welfare framework. The case for that approach rests on

the implicit premise that we can judge whether a given alternative to a standard happiness index brings us closer to a measure of true well-being. But true well-being is not directly measurable (assuming we reject the notion that choices perfectly reveal preferences). How then can we ever hope to satisfy the premise? The fact that the hypothesized alternative would involve *neural* data (either in addition to or instead of self-reported happiness) is of no particular consequence. As I have emphasized, neural activity in particular brain regions is considered welfare-relevant only because it correlates with either choice or self-reported happiness. On the one hand, if we focus on brain regions that are considered welfare-relevant because neural activity has been found to correlate with choice, and if we then use information concerning that activity to "improve" self-reported measures of happiness, we are essentially taking the position that correlates of choice trump self-reports, in which case choice is presumably the best indicator of well-being. On the other hand, if we focus exclusively on brain regions that are considered welfare-relevant because neural activity has been found to correlate with self-reported happiness, then it is difficult to see how we could use information concerning that activity to determine when self-reports are unreliable, let alone to improve them, or to judge whether particular alternatives constitute improvements.

It is useful to restate my objection using the formal language of identification. I will define an *environment*, E, to consist of the external processes that govern experiences, as well as all welfare-relevant preconditions. An environment may or may not present an individual with decisions (present and future), and the preconditions may include past choices. An environment maps to a vector of neural signals, $s = S(E)$, which pertain to sensations associated with physical states (such as hunger and fatigue), abstract emotions (such as shame or relief), or feelings induced by expectations of future sensations. I allow for the possibility that internal well-being, u, depends both on neural signals and on the environment: $u = V(s, E)$. I include E as an argument of U because the nature of the environment (e.g., whether the individual has autonomy) may color hedonic experience. Whether or not V depends directly on E, we can write well-being as a function of the environment and nothing

else:

$$u = U(E) \equiv V(S(E), E). \tag{2}$$

The mapping U is not observed. Our central objective is to identify it up to a monotonic transformation, so we can rank environments according to the well-being they generate.

Let h denote self-reported happiness. Reports presumably depend on well-being, but they may also depend directly on specific sensations (e.g., if a particular feeling is salient when the question is posed) and/or aspects of the environment (e.g., if the respondent normalizes the happiness scale based on recent experience). Thus, I write $h = R(u, s, E)$. Notice that we can write h as a function of E alone:

$$h = H(E) \equiv R(U(E), S(E), E). \tag{3}$$

By assessing self-reported happiness in a wide range of environments, we can observe the mapping H. By taking neural measurements, we also observe components of the mapping S. The question before us is whether we can recover the mapping U from that information.

Having formulated the problem in this way, it should be obvious that identification is hopeless without further restrictions. In particular, for any mappings U, H, and S, there exists functions R and V (neither of which is directly observable) such that equations (2) and (3) hold. For example, R and V might depend only on E, in which case one can simply take $V(E) = U(E)$ and $R(E) = H(E)$. Even if we insist that well-being is responsive to neural signals and that self-reported happiness is responsive to well-being, we are no better off.[26]

The preceding discussion poses a straightforward challenge to those who advocate supplementing happiness-based welfare analysis with neural data: explicitly state the mathematical assumptions that permit us to fully or partially identify true well-being from available data,

[26] To see why, consider the possibility that V and R might be separable: $V(s, E) = V_1(s) + V_2(E)$, and $R(u, s, E) = R_1(u) + R_2(s) + R_3(E)$. For any U, H, S, V_1, R_1, and R_2, one can simply take $V_2(E) = U(E) - V_1(S(E))$ and $R_3(E) = H(E) - R_1(U(E)) - R_2(U(S(E)))$.

and justify them. Identifying restrictions certainly exist; for example, assuming that greater well-being leads to higher levels of reported happiness (monotonicity of R in u), and that those reports do not depend directly on either sensations or the environment (invariance of R with respect to s and E), rankings according to u and h plainly coincide, so H alone identifies U up to a monotonic transformation (R). However, those identifying restrictions are untenable, as are all others I have considered; see Bernheim [2009] for details.

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),[27] while the neural welfare measure points unambiguously to alternative B. Personally, I would be unwilling to overrule the individual's choice and declare him better off with B. On the contrary, I would be inclined to assume that the neural measure overlooked some consideration that was important to the individual.

2.2 Can neuroeconomics improve choice-based welfare analysis?

Naturally, one can also raise objections to the choice-based normative methods of standard economics, including the following two difficulties. First, many people 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

[27] 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.

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 may help to address both 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 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 choice-based welfare analysis requires a *rationalization* of choice (in other words, utility or preferences), and that the associated normative judgments must respect that rationalization, rather than choice itself. Elsewhere, Antonio Rangel and I have argued that choice-based welfare analysis requires no rationalization for behavior (Bernheim and Rangel, 2007, 2008, 2009). When choice lacks a consistent rationalization, the normative guidance it provides may be ambiguous in some circumstances, but 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 (within the welfare-relevant choice domain) 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 prop-

erties of the binary relation, generalizations of the standard tools of applied welfare analysis, and applications to specific behavioral models, in Bernheim and Rangel [2009].

When choice conflicts are severe, our framework remains applicable, but our welfare criterion may not be particularly discerning. If we could find an objective basis for disqualifying certain choices (that is, for removing them from the welfare-relevant choice domain), the severity of choice conflicts would diminish, and our criterion would become more discerning. But how can we disqualify choices while remaining true to the tenets of choice-based welfare analysis? The answer is that those tenets only require respect for *informed* choices. As I explain below, standard welfare analysis actually *requires* us to disqualify or appropriately reinterpret choices upon determining from an examination of the facts (or apparent facts) to which the decision-maker actually had access that he was either misinformed or less well-informed than we had imagined. A more stringent application of the doctrine of respect for informed choice would lead us to make that determination based on the facts (or apparent facts) to which the decision maker *effectively* had access. The resulting refinement agenda defines a potential role for neuroeconomics.

The logic of the refinement agenda Suppose a planner (P) must choose between two actions, A and B, the consequences of which affect only one individual (I) and depend upon a parameter θ, which P always observes (call the true value θ^T). P looks to I's choices for guidance. Upon learning that I has sometimes chosen A and sometimes B when the value of the parameter has been θ^T (which P assumes I can also observe), P fears that those choices have ambiguous implications. In an effort to understand I's apparently conflicting choices, P examines them more closely, and makes one of the following two discoveries.

First, suppose P learns that, in some instances, contrary to his assumption, I did not actually have the opportunity to observe θ. Moreover, whenever I did have that opportunity, I chose A. In that case, according to standard choice-based welfare principles, P should choose A, not B. P must either reinterpret the choice situations in which I selected B in light of I's less-than-perfect information concerning θ, or disqualify those choice situations

for the purpose of welfare analysis if he cannot determine the extent of I's knowledge. But now observe that precisely the same conclusions would follow if P discovered that, whenever I chose B, I simply neglected to observe θ^T despite having the opportunity to do so. It makes no difference whether I lacked the opportunity to observe θ, neglected to observe it due to inattention, or observed and then forgot it whenever he chose B; in each case, I's decision was effectively not informed by the observation that $\theta = \theta^T$, and hence should not guide a planner who observes θ^T.

Next, suppose P learns that, whenever I chose B, I was provided with incorrect information concerning θ (specifically, that $\theta = \theta^F \neq \theta^T$). In that case, according to standard choice-based welfare principles, P should again choose A, not B. P must either reinterpret the choice situations in which I selected B in light of I's misinformation, or disqualify those choice situations for the purpose of welfare analysis if he cannot nail down the misinformation (specifically, the value of θ^F). But now observe that precisely the same conclusions would follow if P discovered that, whenever I chose B, I misinterpreted the available factual data as implying that $\theta = \theta^F$ rather than $\theta = \theta^T$. It makes no difference whether I was provided with incorrect data or misinterpreted correct data whenever he chose B; in either case, I's decision was not correctly informed by the observation that $\theta = \theta^T$, and hence should not guide a planner who observes θ^T.

The preceding discussion implies that evidence concerning the manner in which people observe and process factual information can in principle shed light on the extent to which specific choices are *effectively* informed, and hence appropriate as guides for normative analysis. Despite having access to particular information, a decision maker may be insufficiently attentive, fail to recall important facts that relate choices to consequences, forecast the consequences of his choices incorrectly, or learn from his past experiences more slowly that the available information would permit. Therefore, by studying the neurobiology of attention, memory, forecasting, and learning, it may be possible to identify specific conditions under which people are either effectively misinformed, or effectively less well-informed than they

would be if they correctly processed all the available factual information.

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 identified his alternatives and their consequences.

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. The system learns by adjusting the estimated correlations in response to surprises.[28] Second, unlike other consumption goods, addictive substances *directly* stimulate the neurobiological activity that codes for surprises. Consequently, the repeated use of an addictive substance causes that system to measure empirical correlations incorrectly, and hence to forecast exaggerated rewards in the presence of cues that are associated with its use.[29] Whether or not that system *also* plays a role in hedonic experience, the choices

[28] 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.

[29] 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

made in the presence of those cues are therefore predicated on improperly processed factual information. Therefore, welfare evaluations should be guided by choices made under other conditions.

As an illustration, suppose 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 [2007, 2008, 2009]. 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 neuroeconomic methods. At this stage, our work is still preliminary, so I will confine my remarks to a brief description of our agenda.

---

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.

As discussed in Section 1.4.2, 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.[30]

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 economic analyses of decision making, 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 be overstated. 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. Based on that catalog, I question whether the impact of neuroeconomics on the analysis of conventional positive economic issues is likely to be revolutionary.

I see somewhat 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

[30] In principle, neural measurements could play the same predictive role in happiness-based welfare analysis. However, the value of such predictions would be modest at best. Why use brain scans to construct noisy predictions of a subject's answers to questions about happiness and/or satisfaction when we can simply pose those questions directly?

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 factual information, it can provide objective criteria for disqualifying certain choices for the purpose of welfare analysis. Second, it may allow us to predict the choices that people would make from any given set of prospects based on their passive neural responses. Such predictions would permit us to conduct choice-based welfare analysis even when no choice is actually possible.

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, they must 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. Most economists are not convinced by vague assertions that a deeper understanding of decision-making processes will lead to better models of choice. Second, it is essential to avoid hyperbole. Exaggerated claims fuel skepticism. 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] Akerlof, George, and William Dickens, "The Economic Consequences of Cognitive Dissonance," *American Economic Review* 72(3), 1982, 307-319.

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

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

[4] Bereby-Meyer, Y., A. Assor, and I. Katz, "Children's choice strategies: the effects of age and task demands," *Cognitive Development* 19, 2004, 127-146.

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

[6] Bernheim, B. Douglas, "Behavioral Welfare Economics," *Journal of the European Economics Association*, 2009, forthcoming.

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

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

[9] Bernheim, B. Douglas and Antonio Rangel, "Choice-Theoretic Foundations for Behavioral Welfare Economics," in Andrew Caplin and Andrew Schotter (eds.), *The Founda-*

tions of Positive and Normative Economics: A Handbook, Oxford: Oxford University Press, 2008, 155-192.

[10] Bernheim, B. Douglas, and Antonio Rangel, "Beyond Revealed Preference: Choice-Theoretic Foundations for Behavioral Welfare Economics," *Quarterly Journal of Economics*, 2009, forthcoming.

[11] 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.

[12] Brandstätter, B., G. Gigerenzer, and R. Hertwig, "The priority heuristic: making choices without trade-offs," *Psychological Review*, 413, 2006, 409-432.

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

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

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

[16] Camus, Mickael, Neil Halelamien, Hilke Plassmann, Shinsuke Shimojo, John O'Doherty, Colin Camerer, and Antonio Rangel, "rTMS over the right dorsolateral prefrontal cortex decreases goal values during decision-making," mimeo, Cal Tech, 2008.

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

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

[19] Chong, Juin-Kuan, Camerer, Colin F., and Ho, Teck-Hua, "A Cognitive Hierarchy Model of Games," *Quarterly Journal of Economics* 119(3), August 2004, pp. 861-98.

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

[21] Coricelli, Giorgio, and Rosemarie Nagel, "Neural correlates of depth of reasoning in medial prefrontal cortex," mimeo, Institut des Sciences Cognitives – CNRS, 2008.

[22] Costa-Gomez, Miguel A., and Vincent P. Crawford, "Cognition and Behavior in Two-Person Guessing Games: An Experimental Study," *American Economic Review* 96(5), December 2006, 1737-1768.

[23] Crawford, Vincent, "Look-ups as the Windows of the Strategic Soul," in Andrew Caplin and Andrew Schotter (eds.), *The Foundations of Positive and Normative Economics: A Handbook*, Oxford: Oxford University Press, 2008, pp. 249-280.

[24] Damasio, Hanna, and Antonio R. Damasio. *Lesion Analysis in Neuropsychology.* Oxford University Press, 1989.

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

[26] Erev, Ido, and Alvin E. Roth, "Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria," *American Economic Review* 88 (4), 1998, 848-881.

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

[28] Gianotti, Lorena R.R., Daria Knoch, Pascal L. Faber, Dietrich Lehmann, Roberto D. Pascual-Marqui, Christa Diezi1, Cornelia Schoch, Christoph Eisenegger, and Ernst Fehr, "Tonic Activity Level in the Right Prefrontal Cortex Predicts Individuals' Risk Taking," mimeo, University of Zurich.

[29] 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.

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

[31] Gul, Faruk, and Wolfgang Pesendorfer, "The Case for Mindless Economics," in Andrew Caplin and Andrew Schotter (eds.), *The Foundations of Positive and Normative Economics: A Handbook*, Oxford: Oxford University Press, 2008, pp. 3-42.

[32] Hanks , T.D. , J. Ditterich , and M.N. Shadlen, "Microstimulation of macaque area LIP affects decision-making in a motion discrimination task," *Nat. Neurosci.* 9, 2006, 682 – 689 .

[33] 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.

[34] 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.

[35] Hare, T., J. O'Doherty, C. Camerer, W. Schultz, and A. Rangel, "Dissociating the Role of the Orbitofrontal Cortex and the Striatum in the Computation of Goal Values and Prediction Errors," *Journal of Neuroscience* 28, 2008, 5623-5630.

[36] 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.

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

[38] Jezewski, Sean, B Douglas Bernheim, Colin Camerer, and Antonio Rangel, "Neural responses to stimuli predict unanticipated future decisions," mimeo, Cal Tech, 2009.

[39] Kable, J., and P. Glimcher, "The neural correlates of subjective value during intertemporal choice," *Nature Neuroscience* 10, 2007, 1625-1633.

[40] 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.

[41] Kandel, Eric, James Schwartz, and Thomas Jessell. *Principles of Neural Science.* New York, NY: Elsevier Science Publishing, 1991.

[42] Katsikopoulos, K. V., and L. Martignon, "Naive heuristics for paired compar-isons: some results on their relative accuracy," *Journal of Mathematical Psychology*, 50, 2006, 488-494.

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

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

[45] Krajbich, Ian, Colin Camerer, John Ledyard, and Antonio Rangel, "Neural Mechanism Design: Using Neuro-technology to Solve the Free-Rider Problem," mimeo, Cal Tech, 2009.

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

[47] 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.

[48] 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.

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

[50] Mandler, Michael, Paola Manzini, and Marco Mariotti, "A million answers to twenty questions: choosing by a checklist," mimeo, Unversity of London, 2008.

[51] 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.

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

[53] Mullainathan, Sendhil, "A Memory-Based Model of Bounded Rationality," *Quarterly Journal of Economics* 117(3), 2002, 735-774.

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

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

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

[57] Oosterbeek, Hessel, Randolph Sloof, and Gijs van-de-Kuilen, "Cultural Differences in Ultimatum Game Experiments: Evidence from a Meta-analysis," *Experimental Economics* 7(2), June 2004, 171-88

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

[59] 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(1), January 2008, 95-102.

[60] Plassmann, H., J. O'Doherty, and A. Rangel A,"Orbitofrontal cortex encodes willingness to pay in everyday economic transactions, *Journal of Neuroscience* 27, 2007, 9984-9988.

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

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

[63] Rangel, Antonio, "The Computation and Comparison of Value in Goal-Directed Choice," in P. W. Glimcher, E. Fehr, C. F. Camerer, and R. A. Poldrack (eds.), *Neuroeconomics: Decision Making and the Brain*, Elsevier: New York, 2008, 423-438.

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

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

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

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

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

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

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

[71] Sen, Amartya K., "Behavior and the Concept of Preference," *Economica* 40, 1973, 241-59.

[72] Tom, S., C. Fox, C. Trepel, and R. Poldrack, "The Neural Basis of Loss Aversion in Decision-Making Under Risk," *Science* 315, 2007, 515-518.

[73] Tremblay, L., and W. Schultz, "Relative reward preference in primate orbitofrontal cortex," *Nature* 398, 1999, 704–708.

[74] Tversky, A., "Intransitivity of preferences," *Psychological Review*, 76, 1969, 31–48.

[75] 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.

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