Neuroeconomics: Using neuroscience to make economic predictions

Colin F. Camerer

Div HSS
Caltech
camerer@hss.caltech.edu

ATTN Printer: Short Form Title: Neuroeconomics… to make economic predictions

11/1/2006 1:15 PM. This paper was prepared for the Hahn Lecture, Royal Economic Society, Nottingham UK, April 20, 2006 and is forthcoming in *The Economic Journal*. Thanks to all my collaborators whose joint work is reported (Ralph Adolphs, Meghana Bhatt, Ming Hsu, Michael Spezio, Dan Tranel, Joseph Wang), to RA's Min Kang and Alex Brown, to sceptics for forcing us to think harder and write more clearly about the enterprise, and to many neuroscientists (especially John Allman, Paul Glimcher, John O'Doherty and Read Montague) for tutoring and advice over the last few years.

1. Introduction

Neuroeconomics seeks to ground microeconomic theory in details about how the brain works (see Zak, 2004; Camerer et al., 2005 [\?]; Chorvat and McCabe, 2005; Sanfey et al., forthcoming). Neuroeconomics is a subfield of behavioural economics (behavioural economics uses empirical evidence of limits on computation, willpower and greed to inspire new theories; see Mullainathan and Thaler, 2000; Camerer, 2005). It is also a subfield of experimental economics because neuroeconomics requires mastery of difficult experimental tools which are new to economists (discussed in further detail in section II below). And to many neuroscientists, the greatest promise of neuroeconomics is to supply theories and experimental designs for neuroscience. These neuroscientists feel that the kinds of models and tasks economists use routinely can contribute to "systems neuroscience" understanding of higher-order cognition, which are challenging for neuroscientists who are used to focusing on very fine details of neurobiology and specific brain areas.

To modern economists, the neuroeconomic approach seems to be a sharp turn in economic thought. Around the turn of the century, neoclassical economists made a clear methodological choice, to treat the mind as a black box and ignore its details for the purpose of economic theory (Bruni and Sugden, forthcoming). In an 1897 letter Pareto wrote

> It is an empirical fact that the natural sciences have progressed only when they have taken secondary principles as their point of departure, instead of trying to discover the essence of things. ... Pure political economy has therefore a great interest in relying as little as possible on the domain of psychology (quoted in Busino, 1964, p. xxiv).

Pareto's view that psychology should be deliberately ignored was partly reflective of a pessimism of his time, about the ability to ever understand the brain well enough to use neural detail as a basis for individual economizing. (This pessimism was also manifested in the behaviourist psychology of Watson and Skinner, who turned attention away from the "mentalism" of their time to stimulus-response relations and conditioning.)

As William Jevons wrote a little earlier, in "Theory of Political Economy"

> I hesitate to say that men will ever have the means of measuring directly the feelings of the human heart. It is from the quantitative effects of the feelings that we must estimate their comparative amounts (Jevons, 1871).

This turn-of-the-century pessimism about understanding the brain led directly to the rise of "as if" rational choice models in neoclassical economics. Models of this sort posit individual behaviour which is consistent with logical principles, but do not put any evidentiary weight on direct tests of whether those principles are followed. For example, if a consumer's choices are transitive and complete, then she acts as if she attaches

numerical utilities to bundles of goods and choose the bundle with the highest utility, but direct measurement of utility is thought to be irrelevant as a test of the theory.

The ignorance of psychology that Pareto explicitly advocated was cemented by Milton Friedman's (1953) development of "positive economics". Friedman, and the many economists influenced by his view, advocated two separate principles for judging theories which use assumptions A to make a formal prediction P:

1. Assumptions A should be judged by the accuracy of the predictions P they mathematically imply.
2. Since false assumptions can yield accurate predictions, even if assumptions appear false their empirical weakness should be tolerated if they lead to accurate predictions P.

I wholeheartedly endorse the first principle (1), but not the corollary principle (2).

Here's why: First, if assumptions A are false but lead to an accurate prediction, they presumably do so because of a hidden "repair" condition R (that is, (not-A and R) → P is a more complete theory at both ends than A→ P). Then the proper focus of progressive research should be specifying the repair assumption R and exploring its implications, in conjunction with more accurate assumptions.

Second, the importance of making good predictions (1) is precisely the reason to explore alternative assumptions grounded in psychological and neuroscientific facts. We do this in behavioural economics because we hope that models based on more accurate assumptions will make some interesting new predictions, and better predictions overall.

As-if models based on dubious assumptions clearly work well in many respects, and always will (just as expected value is still a useful tool for some kinds of analysis, even though it is a severe restriction of expected utility). But tests of the predictions that follow from as-if rational choice have also established many empirical anomalies. Behavioural economics describes these regularities and suggests formal models to explain them (e.g. Camerer, 2005).

Debates between rational-choice and behavioural models usually revolve around psychological constructs, such as loss-aversion (Kahneman and Tversky, 1979), the role of learning and limited strategic thinking, a preference for immediate rewards, and precise preferences over social allocations, which have not been observed directly. But technology now allows us to open the black box of the mind and observe brain activity directly. These direct observations can only enhance the development of theories which are based on more accurate assumptions *and* make better predictions as a result.

An analogy to organizational economics illustrates the potential of neuroeconomics (see also Sanfey et al., 2006). Until the 1970's, the "theory of the firm" was basically a reduced-form model of how capital and labour are combined to create a production function. The idea that a firm just combines labour and capital is obviously a

gross simplification— it neglects the details of principal-agent relations, gift exchange and efficiency wages, social networks and favour exchange in firms, substitution of authority for pricing, corporate culture, and so forth. But the gross simplification is useful, for the purpose of building up an industry supply curve.

Later, contract theory opened up the black-box of the firm and modelled the details of the nexus of contracts between shareholders, workers and managers. The *new* theory of the *firm* replaces the (perennially useful) fiction of a *profit-maximizing firm* which has a single goal, with a more detailed account of how *components* of the *firm— individuals, hierarchies,* and *networks*-- interact and communicate to determine *firm* behaviour.

Neuroeconomics proposes to do the same by treating an individual economic agent like a firm. The last sentence in the previous paragraph can be exactly rewritten to replace firms and the components of firms with individuals and neural components of individuals. Rewriting that sentence gives this one: The *neuroeconomic* theory of the *individual* replaces the (perennially useful) fiction of a *utility-maximizing individual* which has a single goal, with a more detailed account of how *components* of the *individual— brain regions, cognitive control,* and *neural circuits*-- interact and communicate to determine *individual* behaviour.

The rapid emergence of various dual-self or dual-process approaches testifies to how well economic theory can be adapted to study the brain as an organization of interacting components. Fudenberg and Levine (forthcoming) emphasize the struggle between a long-run player and a short-run player, adapted from game-theoretic models (see also Thaler and Shefrin's prescient, 1981, "planner-doer" model[1]). Benhabib and Bisin (2005) emphasize the constraint that controlled "executive" processes put on automatic processes. Bernheim and Rangel (forthcoming) emphasize "hot" impulsive states (akin to automatic process, but perhaps driven by visceral factors like drug craving or hunger) and "cold" states. Loewenstein and O'Donoghue (2004) emphasize deliberate processes and affective ones. Brocas and Carillo (2005) emphasize how a cortical control process constrains an emotional process which may be asymmetrically informed. So far, there is little direct neural evidence testing these various models and comparing them. Doing so is an obvious immediate direction for research (and will contribute to basic neuroscience as well).

It is important to note that the focus of neuroeconomic research so far is largely on microeconomics foundations of consumer choice, valuing risky gambles, and strategic thinking. It remains to be seen whether neural measurement will be useful for understanding macroeconomic phenomena like consumer confidence or stock market bubbles. However, many of these macro phenomena might spring from the interaction of many brains that are tightly linked through social networks and common responses to emotional and news shocks which can be reciprocal or contagious. If so, macro models could explore how the result of brain activity has a multiplier effect in the economy.

2. Neuroscientific facts and tools

2.1. Facts

Some basic facts about the human brain are useful economists to know, to understand the evidence presented below and to provide constraint on theorizing.

The brain is weakly modular, in the sense that not every brain area contributes to every behaviour. (That is, the early phrenologists were on the right track, but had too crude a concept of how localized complex behaviours or traits like "virtue" and "sloth" were.) While the brain is modular, it is also "plastic"— responsive to environment as brain 'software' is gradually 'installed'. Plasticity is most obvious in childhood development but seems to continue well into adolescence. Plasticity is the reason why neuroscientists usually bristle at the term "hard-wired", which economists often use casually.

While neuroscientists often focus on specific brain areas which are cyto-architecturally distinct (i.e., they have distinct tissue, neurons, and neurotransmitters), for tasks economists are interested in the proper focus is "circuits" of multiple brain areas. The importance of circuitry also implies that the right kinds of models are computational ones in which well-understood components collaborate to create behaviour.

Attention and consciousness are scarce, and the brain is evolved to off-load decisions by automating activity through learning. Automaticity means that people are capable of creating tremendous expertise which relies on subconscious intuition and pattern recognition. It also means that overcoming automated behaviour takes scarce conscious effort and is often a source of mistakes in "Stroop tasks".[2]

The human brain is basically the primate brain with extra neocortex; and the primate brain is a simpler mammalian brain with some neocortex. This evolutionary history is the main reason why experiments with animals are so informative about human behaviour. (To think otherwise is economic creationism.) For example, rats become biologically addicted to all substances that humans become addicted to (nicotine, opiates, alcohol, etc.). Our shared evolutionary past, and inherited brain regions, do not imply that humans always behave like monkeys (though we sometimes do). Our shared past just implies that when humans struggle to control animal impulses (such as drug addiction), the struggle is between the neocortex and older temporal-lobe areas. Knowing which areas are involved in the struggle is useful for crafting theory and for prescribing treatments.

2.2. Tools

Much of the potential of neuroeconomics comes from relatively recent improvements in technology for measuring brain activity (particularly fMRI), and in matching older technologies (such as eyetracking and EEG) with new tasks.

fMRI uses magnetic resonance imaging, popular for decades for medical diagnosis, at rapid frequencies to measure oxygenated blood flow in the brain

(which is correlated with neural input). The spatial resolution of fMRI is about 3 cubic millimetre voxels and its temporal resolution is 2 seconds. Stronger magnetic fields are unlikely to provide much more improvement (and may post health risks, which modern 3-tesla magnets do not); but improvement may come from innovation in experimental design and statistics.

Positron Emission Tomography (PET) is an earlier scanning technology which injects radioactive solution (usually glucose with a radioactive marker). PET temporal resolution is worse than fMRI (minutes rather than seconds) but glucose is a more direct correlate of neural activity than blood flow.

fMRI and PET are good for roughly identifying areas that are active in a task. Once candidate circuits are established, it is useful to ask whether behaviour is changed when parts of the circuit are broken or disrupted.

Studies of patients with brain lesions are useful for testing hypotheses from fMRI. If a patient with damage to area X cannot perform a task T normally, then area X is part of a normal circuit for doing T. (Lesion data are reported below in a study of the Ellsberg paradox in ambiguous choice.) Transcranial magnetic stimulation (TMS) can "knock out" or activate brain areas, and hence is useful for knowing what targeted areas do. The animal model is also useful because invasive surgeries and genetic engineering can be done with animals, as a substitute for exogenous lesions and correlational studies in humans.

A much more detailed level of data comes from recording activity of a single neuron at a time, mostly from primates (and rarely, from human neurosurgical patients in whom electrodes have been planted to detect locations of epileptic seizures to locate surgical targets).

Older tools continue to be useful. The electroencephalogram (EEG) records very rapid (millisecond) electrical activity from outer brain areas, and can sometimes be used to interpolate activity in areas deeper in the brain. Psychophysiological recording (or skin conductance, heart rate and pupil dilation, for example) are cheap and easy too. Tracking where people are looking on a screen (eyetracking) is also very easy and useful for many questions economists ask. Directly observing the information people use to make decisions provides a second dependent variable that can be used, in conjunction with observed choices, to identify decision rules better than choices alone can.

A great strength of neuroscience is that investigators who have mastered these tools compete fiercely (for grants, students, and space in Science and Nature); their fierce competition creates a bonus for methodological innovation and weeds out weak results. The tools are also complements because each tool can compensate for the weaknesses of others (e.g., having an fMRI finding makes data from patients with lesions in the areas identified by fMRI especially valuable). Recognizing this complementarity, neuroscientists are most comfortable with ideas that are consistent with many types of data recorded in different ways at different levels of temporal and spatial resolution.

Happily for economists, many of our simplest questions can be illuminated by the simple measures (e.g., eye tracking and psychophysiological recording). Ambitious graduate students interested in this field are well advised to pick one tool that can help answer the questions they are interested in, and master it.

Neuroeconomics is likely to provides three types of evidence about economic behaviour. Examples of each type of evidence are given in the next three sections of this paper.[3] The three kinds of evidence are:

1. Evidence which show mechanisms that implement rational choice (utility-maximization and Bayesian integration of information), typically in tasks that are highly-sculpted to make decisions that are useful for survival across species (vision, food, sex and safety).
2. Evidence which support the kinds of variables and parameters introduced in behavioural economics.
3. Evidence which suggest the influence of "new" variables that are implicit, underweighted, or missing in rational-choice theory.

3. Evidence for rational choice principles

In many simple choice domains, evolution has had a long time to sculpt cross-species mechanisms that are crucial for survival and reproduction (involving food, sex, and safety). In these domains, evolution has either created neural circuits which approximate Bayesian-rational choice, or learning mechanisms that generate Bayesian-rational choice with sufficient experience in a stationary environment, putting to use highly-developed capacities for sensory evaluation (vision, taste, smell), memory, and social imitation.

For example, Platt and Glimcher (1999) find remarkable neurons in monkey lateral intraparietal cortex (LIP) which fire at a rate that is almost perfectly correlated with the expected value of an upcoming juice reward, triggered by a monkey eye movement (saccade) (see also Bayer and Glimcher, 2005). Deaner et al. (2005) find that monkeys can reliably trade off juice rewards with exposure to visual images (including images of females from behind, and faces of high and low status conspecific monkeys). Monkeys can also learn to approximate mixed-strategies in games (Glimcher et al., 2005), probably using generalized reinforcement algorithms (Lee et al., 2004). Neuroscientists are also finding prefrontal neurons that appear to express values of choices (Padoa-Schioppa and Assad, 2006 [REF]) and potential locations of "neural currency" that creates tradeoffs (Shizgal, 1997). Following a long tradition in "animal economics" (Kagel et al.,, 1995),

Chen, et al. (2006) show that capuchin monkeys respond to price changes, obeying the GARP axiom, when exchanging tokens for different food rewards.

Another literature shows that Bayesian models are accurate approximations of how different kinds of sensory information are integrated (Stocker and Simoncelli, 2006). These data are in sharp contrast with many cognitive psychology experiments showing

that Bayesian principles are violated when intelligent humans evaluate abstract events (e.g., Kahneman, 2003). It is difficult to reconcile these two literatures directly, because it is difficult to create tasks in which monkeys have to judge the kind of abstract questions people are asked—like whether basketball players have a "hot hand" or whether representative conjunctions of events (F & B) are more likely than their component events (F and B judged separately). Common paradigms that can be used across species represent a huge challenge that would be very useful for either reconciling the results across species or establishing why they differ.

4. Evidence for behavioural economics principles

This section discusses four areas in which neuroscience has established some tentative neural foundation for ideas from behavioural economics which were derived earlier from experiments and field data. The four areas are: β-δ time discounting; aversion to missing information about probability (ambiguity); nonlinear weighting of probability; and limited strategic thinking in games.

Time discounting: Extensive experiments with animals, and later with humans, established that the discount factor put on future rewards is closer to a hyperbola, $1/(1+kt)$, than an exponentially-declining discount factor δ^t. Laibson (1997) borrowed a two-piece discounting function introduced to explain parental bequests, to model "quasi-hyperbolic" discounting. In the β-δ model, agents put a weight of one on current rewards, and weight future rewards at discrete time $t>0$ by $\beta\delta^t$. (When $\beta=1$ the two-parameter function reduces to an exponential.) O'Donoghue and Rabin (1999) dubbed the β term a "present bias" and explore its implications. Various field and experimental data suggest values of β around .6-.8.[4] To search for β and δ processes in the brain, McClure et al (2004) presented subjects with choices between a current reward and a reward with a one-month delay (which activates both β and δ systems), and other choices with a one-month or two-month delay (in which the β component divides out). They find activity in areas often associated with an emotional limbic system (medial frontal cortex, cingulate, and ventral striatum) when β comes into play, and find distinct activity in lateral orbitofrontal cortex and dorsolateral cortex linked to the δ system. Their study is hardly the last word—in fact, it's the first word— but is consistent with discounting being a splice of two processes.

Ambiguity-aversion: In subjective expected utility theory, the willingness to take bets on events is taken to reveal subjective probabilities of those events. The Ellsberg paradox showed that for a small majority of subjects, when two events are equally likely but poorly understood (or "ambiguous"), revealed decision weights seem to combine judgment of likelihood and an additional factor which leads to an aversion to betting under ambiguity. Theories of nonadditive probability and set-valued probabilities loosely ascribe this ambiguity-aversion to pessimism or fear of betting in the face of unknown information. Ambiguity-aversion has been implicated in "home bias" in financial investment (a preference for investing in stocks in one's own country, or firm, or firms nearby), in "robust control" in macroeconomics, and in other economic domains (Camerer et al., forthcoming). Scottish law provides a useful practical example. In Scottish law there are three verdicts—guilty, not guilty, and "unproven". An unproven verdict results when there is too little evidence to determine guilt or innocence (often in

sexual assault cases, since Scottish law requires a corroborating witness besides a testifying victim). Unproven verdicts are usually the jury's way of expressing an aversion to rendering either verdict, often shaming a victim they believe is guilty but cannot legally find guilty because of evidentiary rules which create reasonable doubt .

Since decision theorists forming axioms are not generally thinking about brain activity adhering to those axioms, it is difficult to find descriptions which are suggestive of neural activity. But Raiffa (1963) wrote:

> But if certain uncertainties in the problem were in cloudy or fuzzy [ambiguous] form, then very often there was a shifting of gears and no effort at all was made to think deliberately and reflectively about the problem. Systematic decomposition of the problem was shunned and an over-all 'seat of the pants' judgment was made which graphically reflected the temperament of the decision maker.

Unfortunately, the "seat of the pants" is not a brain area, but Raiffa's description does a rapid emotional response in the face of ambiguity. Hsu et al. (2005) investigated ambiguity and risk using fMRI (see also Huettel et al., 2006). They found additional activation in valuing bets on ambiguous gambles relative to risky ones (such as bets on low-knowledge events, like the temperature in Tajikistan compared to high-knowledge New York). They found additional activity in the dorsolateral prefrontal area, orbitofrontal cortex (above the eye sockets, OFC) and the amygdala (a "vigilance" area, which is rapidly activated in 5-20 msecs by fearful images, even before they are consciously processed). Subjects with higher right OFC activity in response to ambiguity also had higher ambiguity-aversion parameters as estimated by a stochastic choice logit model fit to gamble valuations.

Nonlinear probability weighting: In expected utility (EU) theory, the utilities of gamble outcomes are weighted by their probability p. But many experimental studies suggest that people actually weight probabilities nonlinearly with a function $\pi(p)$, overweighting low probabilities and underweighting probabilities close to one (the "certainty effect"); see Figure 1 (from Prelec, 1998). Overweighting of low p could be important in pricing insurance and in explaining demand for lottery tickets and the high failure rate of new businesses.

Measuring neural activation in response to variation in probability is made possible by the fact that a fair amount is known about how the caudate (a temporal lobe area including the striatum) responds to anticipated reward. Hsu et al.(in preparation) set out to see whether activation in the striatum responded nonlinearly to probability of winning. They first presented simple binary gambles (p,X) which have a p chance of paying $X (otherwise they pay zero) for a few seconds, then had subjects choose between the presented gamble and a second gamble (roughly matched for expected value). The choice data enable estimation of parameters of a probability weighting function $\pi(p)$. They look at activity in the left and right caudate areas—an area in the temporal lobe associated with rewards of many types (juice, cocaine, attractive faces, money, faces of people who have cooperated with you). Controlling for the payoff amount X, there is a modest nonlinearity of activity across levels of probability p which is reasonably similar

to the nonlinear functions shown in Figure 1. This similarity of indirect estimates and direct estimates of caudate activity is not conclusive proof that the brain is weighing probabilities nonlinearly, but it is consistent with that hypothesis. A likely explanation is that probability estimation is a combination of a linear weighting and an inverse-S step function which sorts probabilities crudely into "no, maybe, yes".[5] Combining the two gives a regressive function that overweighs low p and underweighs high p, and is consistent with the brain activation.

Limited strategic thinking: In game theory, players are in equilibrium when they guess correctly what other players will do— that is, when their beliefs about other players' strategies match the actual strategies others choose. Camerer et al. (2004) describe an alternative "cognitive hierarchy" (CH) theory in which players use various steps of strategic thinking. Some step-0 players randomize, other step-1 players anticipate randomization and best-respond to it, step-2 players best-respond to a mixture of step-0 and step-1 players, and so on. Since the highest-step players anticipate correctly the distribution of what other players will do, their beliefs are in equilibrium, but the beliefs of lower-step thinkers are not in equilibrium because they do not guess correctly what higher-step players will do. This model (and earlier versions introduced by others) fits empirical data from dozens of game experiments with many different structural forms (mixed-equilibria, coordination, dominance-solvable games, and so forth). To look for evidence of limited strategic thinking in the brain, Bhatt and Camerer (2005) did fMRI of players when they made choices, and when they expressed beliefs about what other players would do. They found that when players' choices and beliefs were in equilibrium, there was almost perfect overlap in brain activity during choosing and belief expression—that is, creating equilibrium beliefs requires players to imagine how others are choosing, which uses overlapping neural circuitry with making your own choice (Figure 2). When players were out of equilibrium, there was much more activity when making a choice than when expressing a belief (as would be expected from 0- and 1-step thinkers, who are thinking harder about their own choice than they are about choices of other players). Thus, being in equilibrium is not merely a mathematical restriction on equality of choices and beliefs, it is also a "state of mind" identifiable by brain imaging.

5. Evidence for new psychological variables

My view is that the largest payoff from neuroeconomics will not come from finding rational-choice processes in the brain for complex economic decisions, or from supporting ideas in behavioural economics derived from experimental and field data (as shown by examples in the last two sections). The largest innovation may come from pointing to biological variables which have a large influence on behaviour and are underweighted or ignored in standard theory. This section lists a few speculative examples. They suggest that the concept of a preference is not a primitive (as Pareto suggested); preferences are both the output of a neural choice process, and an input which can be used in economic theory to study responses to changes in prices and wealth. This view implies that if we understand what variables affect preferences, we can shift preferences and shift behaviour (without changing prices or constraints). Whether this can be done reliably or on a large scale is not yet known. The goal at this point is just to show that understanding biology and the brain can make fresh predictions about observed

choices. At this point, there are few such predictions and they focus on small effects at the individual level. But given the youth of the field, having any such examples is suggestive and they point in interesting directions.

1. In the ambiguity study described in the last section (Hsu et al., 2005), there is a modest correlation of right OFC activity with a parameter characterizing the degree of ambiguity-aversion, which is derived from estimation using choices. (The parameter γ is derived implicitly from the weight $(E(p)^{\gamma})$ given to an event with expected or diffuse-prior probability p. The value $\gamma=1$ is ambiguity-neutrality. A value $\gamma>1$ corresponds to ambiguity-aversion; an ambiguity-averse person acts like the decision weight on an ambiguous event is lower than its expected probability.) One can extrapolate statistically from the correlation between OFC activation and γ in normal subjects to infer the behavioural value of γ that would be revealed by choices of a person with no OFC activity at all—due to a lesion in that area, say (see Figure 3). The extrapolated estimate is $\gamma=.85$ (roughly ambiguity-neutral, given sampling error). In fact, Hsu et al. also tested Ellsberg-type problems on patients with OFC damage subsuming the areas observed in fMRI. Those patients' choices exhibited a value of $\gamma=.82$. I would love to say this value was truly predicted before the fact, but it was not (both studies were conducted in parallel). In any case, there is a close link between the behavioural parameter "predicted" by extrapolating from the fMRI evidence to patients with no activity, and the extrapolated parameter is close to the figure revealed by choices. While this correspondence could be construed as consistent with axiomatic theories of ambiguity-aversion, no theory would have predicted it without the fMRI evidence to tell us what lesion patients would be roughly ambiguity-neutral.

2. Wang et al. (2006) studied experimentally a classic "biased-transmission game" that has been widely used in economics and political science. In this game, a sender observes a state S, an integer from 1 to 5 (uniformly distributed). The sender then chooses an integer message M from 1 to 5. A receiver knows the setup of the game, and learns the message, but *does not know the true state directly*. The receiver then chooses an action A from 1 to 5. (The game is like security analysts who know more about the value of a stock than you do, make a recommendation, and want you to act as if the stock is more valuable than it is, because of career concerns or other collateral interests.) In the interesting conditions, the senders earn the most if the receiver chooses S+b, where b is a known bias parameter (either 1 or 2). We try to predict the true state from the sender's message M, and from their pupil dilation (expansion of pupils) when they send their message. Pupils dilate under arousal, stress, and deception (that's why poker players wear sunglasses if they are allowed to). Statistical tests show that measuring the pupil dilation improves substantially in predicting what the true state is. Thus, a biological variable helps infer private information which is conveyed by messages, in a way that is not explicitly predicted by conventional game theory.

3. Sanfey et al. (2003) used fMRI to see what areas were differentially active in the brains of responders in an ultimatum game, when the responders received a fair offer ($4-5 out of $10) compared to an unfair offer ($1-2). They found activation in the insula (a discomfort or disgust area, perhaps measuring the emotional reaction to getting a low offer), dorsolateral prefrontal cortex (DLPFC, a planning and evaluation area), and anterior cingulate (a conflict-resolution area). They also found that whether people rejected low offers or not could be predicted with some accuracy from whether the insula was more active than DLPFC or vice versa. Building on this study, Wout et al. (2005) and Knoch et al. (2006) used repetitive TMS to disrupt the DLPFC when people received offers. Based on the fMRI evidence, they hypothesized that if the DLPFC is disrupted, the socialized response to unfairness which leads to rejection may be turned off, so that people will exhibit more innate selfishness and accept lower offers more often. Their prediction was correct. The effects are small and come from only two studies with modest sample sizes, but they show the power of a two-step process: First establish parts of neural circuitry that implement a behaviour; then stimulate or disrupt some of those parts and if see if you can *cause* a behavioural change.

4. Oxytocin is a powerful hormone in social bonding (e.g., it surges when mothers breast-feed; and synthetic oxytocin—pitocin—is administered in American hospitals to stimulate childbirth). Direct measurement from blood samples (Zak et al., 2005) suggests oxytocin is important in trust. Inspired by this evidence, Kosfeld et al. (2005) had subjects play a trust game in which one player could choose whether to invest money or keep it. If she invested, the money doubled in amount and the responder player (the trustee) could decide how much to repay and how much to keep. Half the subjects were given a synthetic oxytocin dose (three puffs in each nostril, then wait an hour) and half were given a placebo so the subjects could not tell whether they got the real pitocin or nothing. Kosfeld et al hypothesized that oxytocin would increase trust, and it did. Game theory makes predictions about structural variables that might increase trust—most reliably, whether the game is repeated or played once (which does have a strong impact; e.g., Ho et al. (forthcoming). But nothing in game theory would have predicted the effect of synthetic oxytocin.

6. Conclusion

The goal of neuroeconomics is to ground economic theory in details of how the brain works in decision making, strategic thinking, and exchange. One way to achieve this is to observe processes and constructs which are typically considered unobservable, to decide between many theories of behavioural anomalies like risk aversion, altruistic punishment, and reciprocity.

I have presented examples in which neuroeconomic evidence points to either of three conclusions. Sometimes rational-choice processes are clearly evident in brain activity (LIP neurons that fire at rates almost exactly linear in expected reward). In other cases, the variables or differences predicted by behavioural economics models are

evident— in β-δ discounting, ambiguity-aversion, and nonlinear probability weighting. In still other case, perhaps the most innovative, variables that are not a traditional focus of economic theory have perceptible effects, and sometimes strong effects: Patients with OFC damage are unusually ambiguity-neutral (which is consistent with fMRI evidence identifying the OFC as a locus of ambiguity-aversion processing); pupil dilation helps predict a player's private information when they might be lying; stimulating DLPFC increases acceptance of low ultimatum offers (because earlier fMRI work showed DLPFC activity is correlated with acceptance); and administering oxytocin makes people more trusting.

For me, thinking about how the brain implements economic decisions, compared to thinking about choices resulting from preference and belief, is like switching from watching TV in black and white to watching in colour— there are so many more variables to think about. For economic theorists, a natural way to think about these phenomena is that many biological state variables influence preferences; given those state-dependent preferences, prices and budget constraints have familiar influences. I agree with this view, except that we will never fully understand the nature of the state-dependence without facts from psychology and neuroscience. Furthermore, it is not clear whether subjects are aware of exogenous influences that alter these internal states, and how the state-dependence works when a lot of money is on the line (arousal itself can be a big state variable) and when agents are highly experienced.

There is much obvious future research. One path is to study the multiple-process approaches seriously and look for those processes directly in the brain, or as they are manifested in behavioural experiments.[6] Another is to search for evidence of distinctions that are well-established in behavioural economics (such as gain-loss differences, framing effects, emotional foundations of inequality-aversion or social image, and so forth). A more unifying approach is to take the revealed-preference model seriously and see how far its language can be stretched to accommodate neural evidence, while making new predictions rather than just giving economic names to neural processes.

7. Afterword and prologue: The "mindless" critique, and a reply from the past

Some economists feel that the central theory in economics—revelation of inherently unobservable preferences and beliefs by observed choices— is immune to empirical evidence from neuroeconomics. Their argument is that economics is only about explaining choices, and neural evidence is not choices. For example, Gul and Pesendorfer (2005) suggest one categorization of economics (which could be called "economics™", because they so assuredly legislate what economics is and is not). They write[1]

> …the requirement that economic™ theories simultaneously account for economic™ data and brain imaging data places an unreasonable burden on economic™ theories" (Gul and Pesendorfer, 2005)

Some of the examples in sections IV and V were judiciously chosen to address precisely this critique. Theories of β-δ time discounting and nonlinear $\pi(p)$ probability weighting can account for both behavioural data from many choice experiments (and many field data too), *and* are consistent with tentative evidence of neural activity. Since such theories are possible, it is really an "unreasonable burden" to ask whether other theories can do the same? Of course, theories that spring from the fertile mind of a theorist who is simply inspired by psychology, but is not beholden to a large body of facts, could prove to be useful theories too. But theories that can explain neural facts *and* choices should have some advantage over theories which explain *only* choices, if they are comparably tractable.

More fundamentally, the argument against neuroeconomics (or the case for "mindless" economics, as their paper's title calls it) rests mostly on an interesting hope, and rests a little bit on the history of economic thought. The hope is that all anomalies produced by behavioural economics and neuroeconomics can be explained (if not predicted) by the enriched language of economics— preferences, beliefs, and imperfect information and constraint. I share that hope, but only if some imperfections and constraints are allowed to be located in the brain— in which case, brain evidence is useful for understanding those imperfections and constraints and suggesting the best models of them.

A useful focus for debate is therefore how gracefully (and predictively) conventional economics language can explain the effects on observed choices (and inferred unobservable states) of brain lesions, pupil dilation, TMS stimulation, and oxytocin. Any conventional accounts which absorb these effects semantically, and then make predictions about them, will be welcomed as interesting neuroeconomics.

The history of economic thought part of the "mindless" case is more clearly settled. Gul and Pesendorfer write that "Populating economic™ models with 'flesh-and-blood human beings' was never the objective of economists™." But Colander (2005) reminds us how interested classical economists were in measuring concepts like utility directly, before Pareto and the neoclassicals gave up.

---

[1] In the passages quoted from their paper, of course, the TM superscripts do not appear.

Edgeworth dreamed of a "hedonimeter" that could measure utility directly; Ramsey fantasized about a "psychogalvanometer"; and Irving Fisher wrote extensively, and with a time lag due to frustration, about how utility could be measured directly. Edgeworth wrote:

> …imagine an ideally perfect instrument, a psychophysical machine, continually registering the height of pleasure experienced by an individual…From moment to moment the hedonimeter varies; the delicate index now flickering with the flutter of the passions, now steadied by intellectual activity, low sunk whole hours in the neighbourhood of zero, or momentarily springing up towards infinity…"

The interest of these early economists in measuring utility directly was to establish a biological cardinal utility scale. In any case, given their ambitions, it is hard to believe at least some of these important figures would not be interested in using the modern tools that we do have. If Edgeworth were alive today, would he be making boxes, or recording the brain?

Footnotes

1. Benabou and Pyciak (2002) show how the Gul-Pesendorfer (2001) model of preferences under temptation is mathematically equivalent to a rent-seeking competition between two brain areas, linking the preferential approach to the multiple-selves approach.

2. In the classic Stroop task, people are asked to name the color of ink a word is printed in. Under time pressure, people invariably state the word rather than the color (e.g., if the word "black" is printed in green ink, they say "black", not "green") at first, though they can learn over time. The Stroop task is now used as a generic term for any automated response which must be overridden by cognitive control. The game "Simon says" is an example. Another example is when Americans visit England. Americans are used to looking to the left for cars approach them was they cross the street, but in England cars approach from the right. Many Americans are killed every year because of a Stroop mistake. The fact that avoiding a Stroop mistake takes conscious effort also predicts that Americans whose conscious attention is absorbed elsewhere when they are crossing the street in England— talking on a cell phone, for example— are more likely to be killed than those who are not distracted.

3. Note that the length of the three sections is not intended to reflect either the accumulated regularity in each of the three areas, or likely future results. The last section is longer because it presents a more novel perspective, and most directly meets the critique that neuroeconomics does not provide new insight.

4. Angeletos et al (2001), Della Vigna and Paserman (2005), Tanaka et al. (2006, http://www.hss.caltech.edu/~camerer/Growth-nth.pdf) and Brown et al. (2006) all report estimates from savings data, unemployment data, abstract experiments in Vietnam, and dynamic savings rewards with temptation (respectively) with β around .6-.8.

5. Attention and adaptation probably also play crucial roles. While some risks are overweighed, others might be dismissed entirely because they are not imagined or attended to. There is no experimental paradigm to turn on and off attention to low probability risks; having one would be useful, as would field measurements of actual attention to risks.

6. For example, the Bernheim-Rangel, Fudenberg-Levine, and β-δ time preference models all predict that subjects who are tempted by immediate rewards will make different decisions if current choices are not consumed until a time sufficiently far in the future (so that the "hot self", "short-run player", or "present-biased" current player's myopic preferences are disabled). Brown et al. (2006) find the first direct evidence of such an effect in dynamic savings experiments, when thirsty subjects decide how much of a thirst-slaking beverage to consume. When subjects have to "order in advance", by making choices at period t which are not consumed until period t+10, they consume less and earn more overall rewards. Calibrating β-δ parameters to actual decisions yields

sensible estimates of δ=.90 and β=.76-.85 (the latter depends on whether agents are sophisticated about their present bias, or naïve).

References

Angeletos, G.-M., Laibson, D., Repetto, A., Tobacman, J. and Weinberg, S. (2001). 'The hyperbolic consumption model: Calibration, simulation, and empirical evaluation', *Journal of Economic Perspectives*, vol. 15 (3) (Summer), pp. 47-68.

Bayer, H. M. and Glimcher, P. W. (2005). 'Midbrain dopamine neurons encode a quantitative reward prediction error signal', *Neuron*, vol. 47 (1) (07 July 2005), pp. 129-141.

Benabou, R. and Pyciak, M. (2002). 'Dynamic inconsistency and self-control: A planner-doer interpretation', *Economics Letters*, vol. 77 (3) pp. 419-424.

Benhabib, J. and Bisin, A. (2005). 'Modeling internal commitment mechanisms and self-control: A neuroeconomics approach to consumption-saving decisions', *Games and Economic Behavior*, vol. 52 (2) (August), pp. 460-492.

Bernheim, B. D. and Rangel, A. (forthcoming). 'Behavioral public economics: Welfare and policy analysis with fallible decision-makers', in (P. Diamond and H. Vartiainen, ed.), *Economic institutions and behavioral economics*, Princeton University Press.

Bhatt, M. and Camerer, C. F. (2005). 'Self-referential thinking and equilibrium as states of mind in games: Fmri evidence', *Games and Economic Behavior*, vol. 52 (2) (August), pp. 424-459.

Brocas, I. and Carrillo, J. (2005). 'The brain as a hierarchical organization', University of Southern California.

Brown, A. L., Camerer, C. F. and Chua, Z. E. (2006). 'Learning and visceral temptation in dynamic savings experiments ', Caltech.

Bruni, L. and Sugden, R. (forthcoming). 'The road not taken. Two debates on economics and psychology', *Economic Journal*, .

Busino, G. (1964). 'Note bibliographique sur le cours', in (V. Pareto, ed.), *Epstolario*, pp. 1165-1172, Rome: Accademia Nazionale dei Lincei.

Camerer, C., F. (2005). 'Behavioral economics', London.

Camerer, C. F., Ho, T. H. and Chong, J. K. (2004). 'A cognitive hierarchy model of games', *Quarterly Journal of Economics*, vol. 119 (3) (August), pp. 861-898.

Chen, M. K., Lakshminarayanan, V. and Santos, L. (forthcoming). 'How basic are behavioral biases? Evidence from capuchin-monkey trading behavior', *Journal of Political Economy*.

Chong, J.-K., Camerer, C. F. and Ho, T. H. (forthcoming). 'A learning-based model of repeated games with incomplete information', *Games and Economic Behavior*.

Chorvat, T. R. and Mccabe, K. (2005). 'Neuroeconomics and rationality', *Chicago-Kent Law Review*, vol. 80 (3) (August), pp. 1235-1255.

Colander, D. (2005). 'Neuroeconomics, the hedonimeter, and utility: Some historical links', Middlebury College.
Deaner, R. O., Khera, A. V. and Platt, M. L. (2005). 'Monkeys pay per view: Adaptive valuation of social images by rhesus macaques', *Current Biology*, vol. 15 (29 March 2005) pp. 543-548.
Della Vigna, S. and Paserman, M. D. (2005). 'Job search and impatience', *Journal of Labor Economics*, vol. 23 (3) (July), pp. 527-588.
Friedman, M. (1953). *The methodology of positive economics*.
Fudenberg, D. and Levine, D. (forthcoming). 'A dual self model of impulse control', *American Economic Review*.
Glimcher, P. W., Dorris, M. C. and Bayer, H. M. (2005). 'Physiological utility theory and the neuroeconomics of choice', *Games and Economic Behavior*, vol. 52 (2) (August), pp. 213-256.
Gul, F. and Pesendorfer, W. (2001). 'Temptation and self-control', *Econometrica*, vol. 69 (6) (November), pp. 1403-1435.
Gul, F. and Pesendorfer, W. (2005). 'The case for mindless economics', Princeton University, November.
Hsu, M., Bhatt, M., Adolphs, R., Tranel, D. and Camerer, C. F. (2005). 'Neural systems responding to degrees of uncertainty in human decision-making', *Science*, vol. 310 (5754) (December 9, 2005), pp. 1680-1683.
Hsu, M., Chen, Z. and Camerer, C. F. (in preperation) Nonlinear probability weighting in the brain', Caltech.
Huettel, S. A., Stowe, C. J., Gordon, E. M., Warner, B. T. and Platt, M. L. (2006). 'Neural signatures of economic preferences for risk and ambiguity', *Neuron*, vol. 49 (5) (2 March 2006), pp. 765-775.
Jevons, W. (1871). *Theory of political economy*.
Kagel, J., Battalio, R. C. and Green, L. (1995). *Economic choice theory: An experimental analysis of animal behavior*. Cambridge: Cambridge University Press.
Kahneman, D. (2003). 'A psychological perspective on economics', *American Economic Review*, vol. 93 (2) (May), pp. 162-168.
Kahneman, D. and Tversky, A. (1979). 'Prospect theory - analysis of decision under risk', *Econometrica*, vol. 47 (2) (March), pp. 263-291.
Kosfeld, M., Heinrichs, M., Zak, P. J., Fischbacher, U. and Fehr, E. (2005). 'Oxytocin increases trust in humans', *Nature*, vol. 435 (7042) (02 Jun 2005), pp. 673-676.
Laibson, D. (1997). 'Golden eggs and hyperbolic discounting', *The Quarterly Journal of Economics*, vol. 112 (2) (May), pp. 443-477.
Lee, D., Conroy, M. L., Mcgreevy, B. P. and Barraclough, D. J. (2004). 'Reinforcement learning and decision making in monkeys during a competitive game', *Cognitive Brain Research*, vol. 22 (1) (December), pp. 45-58.
Loewenstein, G. and O'donoghue, T. (2004). 'Animal spirits: Affective and deliberative processes in economic behavior', Carnegie Mellon University,

Mcclure, S. M., Laibson, D. I., Loewenstein, G. and Cohen, J. D. (2004). 'Separate neural systems value immediate and delayed monetary rewards', *Science*, vol. 306 (5695) (October 15, 2004), pp. 503-507.

Mullainathan, S. and Thaler, R. (2000). *Behavioral economics: Entry in international encyclopedia of the social and behavioral sciences*. Massachusetts Institute of Technology.

O'Donoghue, T. and Rabin, M. (1999). 'Doing it now or later', *American Economic Review*, vol. 89 (1) (March), pp. 103-124.

Platt, M. L. and Glimcher, P. W. (1999). 'Neural correlates of decision variables in parietal cortex', *Nature*, vol. 400 (6741) pp. 233-238.

Prelec, D. (1998). 'The probability weighting function', *Econometrica*, vol. 66 (3) (May), pp. 497-527.

Raiffa, H. (1961). 'Risk, ambiguity, and the savage axioms: Comment', *The Quarterly Journal of Economics*, vol. 75 (4) (Winter), pp. 690-694.

Sanfey, A. G., Loewenstein, G., Cohen, J. D. and Mcclure, S. M. (Forthcoming). 'Neuroeconomics: Cross-currents in research on decision', *Trends in Cognitive Sciences*.

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E. and Cohen, J. D. (2003). 'The neural basis of economic decision-making in the ultimatum game', *Science*, vol. 300 (5626) (June 13, 2003), pp. 1755-1758.

Shefrin, H. M. and Thaler, R. H. (1988). 'The behavioral life-cycle hypothesis', *Economic Inquiry*, vol. 26 (4) (October), pp. 609-643.

Stocker, A. A. and Simoncelli, E. P. (2006). 'Noise characteristics and prior expectations in human visual speed perception ', *Nature Neuroscience* vol. 9 (4) (April), pp. 578-585.

Tanaka, T., Camerer, C. F. and Nguyen, Q. (2006). 'Poverty, politics, and preferences: Experimental and survey data from vietnam', California Institute of Technology.

Thaler, R. H. and Shefrin, H. M. (1981). 'An economic theory of self-control', *The Journal of Political Economy*, vol. 89 (2) (April), pp. 392-406.

Wang, J. T.-Y., Spezio, M. and Camerer, C. F. (2006). 'Pinocchio's pupil: Using eyetracking and pupil dilation to understand truth-telling and deception in biased transmission games', Caltech.

Wout, M. V. T., Kahn, R. S., Sanfey, A. G. and Aleman, A. (2005). 'Repetitive transcranial magnetic stimulation over the right dorsolateral prefrontal cortex affects strategic decision-making', *Neurorepor*, vol. 16 (16) (7 November), pp. 1849-1852.

Zak, P. J. (2004). 'Neuroeconomics', *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences*, vol. 359 (1451) (29 November 2004), pp. 1737-1748.

Zak, P. J., Kurzban, R. and Matzner, W. T. (2005). 'Oxytocin is associated with human trustworthiness', *Hormones and Behavior* vol. 48 (5) (December), pp. 522 – 527.

Fig. 1: Prelec (1998) probability weighting function and estimated shapes from various experimental studies

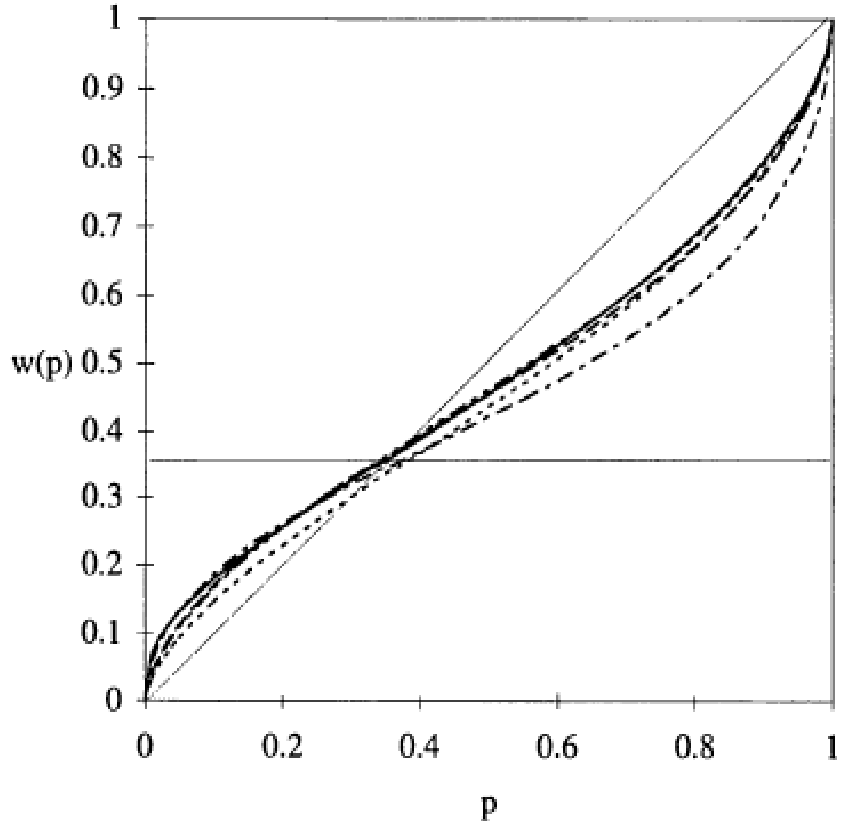
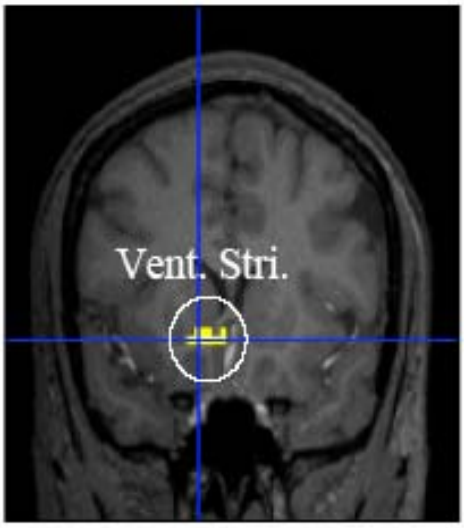
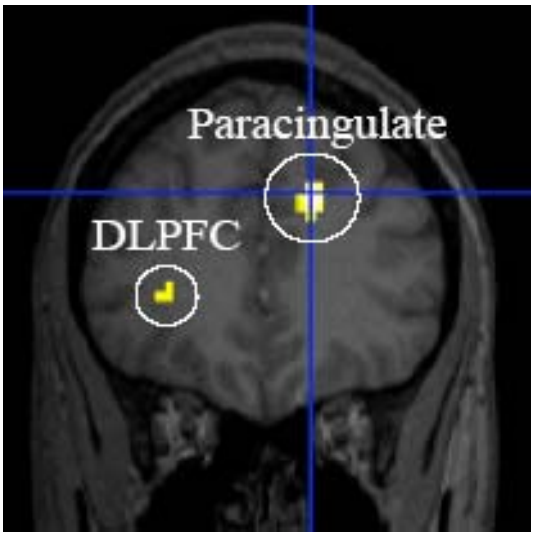


Fig. 2: Differences in brain activity during choosing a strategy and expressing a belief about another player's strategy (Bhatt and Camerer, 2005). Equilibrium trials (A) show only a difference in ventral striatum (a reward anticipation area). Out-of-equilibrium trials (B) show stronger activity in choosing than in belief expression (highlighting paracingulate and dorsolateral prefrontal (DLPFC) areas), which suggests subjects are not reasoning strategically about other players.

(A) In equilibrium



(B) Out of equilibrium



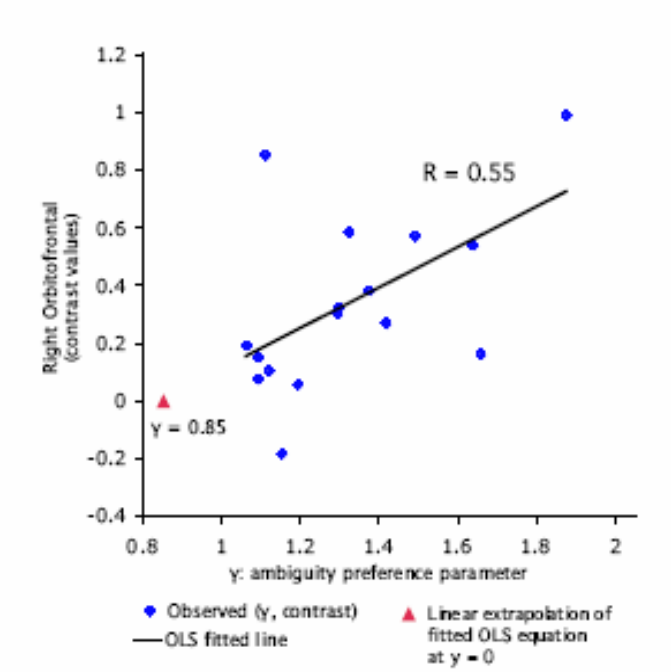


Fig. 3: Correlation between individual-specific ambiguity-aversion parameters γ estimated from choices (x-axis, higher γ is more ambiguity-aversion) and differential activity in right orbitofrontal cortex in ambiguous vs. risky gamble evaluation (y-axis). Positive correlation (r=.55) indicates more ambiguity-averse people have more differential activity in ROFC. Extrapolating to a person with no OFC activity (y=0) gives an inferred ambiguity-aversion γ of .85. The actual behavioural parameter derived from choices of patients with OFC lesions was γ=.82.