Handbook of Experimental Economics
Editors: John Kagel and Alvin Roth

Neuroeconomics

Colin Camerer (California Institute of Technology),
Jonathan Cohen (Princeton University),
Ernst Fehr (University of Zurich),
Paul Glimcher (New York University),
David Laibson (Harvard University)

Camerer, Division HSS, Caltech, camerer@hss.caltech.edu;
Cohen, Princeton Neuroscience Institute, Princeton University, jdc@princeton.edu;
Fehr, University of Zurich, Department of Economics, ernst.fehr@econ.uzh.ch;
Glimcher, Center for Neural Science, New York University, glimcher@cns.nyu.edu;
Laibson, Department of Economics, Harvard University, dlaibson@harvard.edu.

We acknowledge financial support from the Moore Foundation (Camerer), the National Science Foundation (Camerer), and the National Institute of Aging (Cohen, Laibson).

Introduction

"One may wonder whether Adam Smith, were he working today, would not be a neuroeconomi[st]"

Aldo Rustichini (2005).

Neuroeconomics is the study of the biological microfoundations of economic cognition and behavior. Biological microfoundations are neurochemical mechanisms and pathways, like brain regions, neurons, genes, and neurotransmitters.[1] Economic cognition includes memory, preferences, emotions, mental representations, expectations, anticipation, learning, information processing, inference, simulation, valuation, and the subjective experience of reward. In general, neuroeconomic research seeks to identify and test biologically microfounded models that link cognitive building blocks to economic behavior. If successful, neuroeconomic research will improve economists' ability to forecast behavior (e.g., Bernheim et al 2011, Fehr and Rangel 2011).

Neuroeconomics is a big tent. Neuroeconomic research *requires* some curiosity about *neurobiology*, but neuroeconomic research does not necessarily require a departure from classical economic assumptions (e.g., rationality and dynamic consistency). A classical economist would be a neuroeconomist if she wanted to study the biological mechanisms that influence rational decision-making. For example, neuroeconomic research provides insights about the sources of preference heterogeneity. To be a neuroeconomist you need to take an interest in the operation of the brain, but you don't need to prejudge its rationality.

Neuroeconomics includes both theoretical modeling and empirical measurement. At the moment, the majority of neuroeconomic research is focused on measurement. However, this may change as a rapidly growing body of empirical knowledge provides discipline and catalyzes theoretical integration.

Neuroeconomists use many different empirical methods, though neuroimaging is by far the dominant methodology now – especially functional magnetic resonance imaging (fMRI).[2] Neuroimaging technologies enable researchers to measure brain activity during problem solving, game-playing, choice, consumption, information revelation, and almost any conceivable type economic activity. Neuroeconomic research also uses a diverse body of complementary data sources, including neuropharmacological exposures, cognitive load manipulations, response time measurements, transcranial magnetic stimulation (a technology that temporarily alters normal cognitive functioning in a localized region of the brain), genotyping, analysis of patients with neural anomalies (e.g. brain lesions), and the study of animal models (e.g. rats or monkeys).

There are four principal motivations for pursuing neuroeconomic research.

First, some researchers are willing to study neuroscience for its own sake. Few economists share this sentiment.

Second, neuroeconomic research will likely provide a new way of (imperfectly) measuring human well-being. For example, neural activity has been shown to correlate with reports of subjective well-being (EEG cite), receipts of reward (Schulz et al 1998, Knutson et al 2006), and revealed preferences (Glimcher 2003; de Quervain et al 2005). Camerer (2006) writes that:

---

[1] Neurotransmitters are molecules that carry neurochemical signals from one neuron to another.

[2] Other neuroimaging methods include magnetic resonance imaging (MRI), positron emission tomography (PET), and electroencephalograms (EEG)

> "Colander (2005) reminds us how interested classical economists were in measuring concepts like utility directly, before Pareto and the neoclassicals gave up. 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 neighborhood of zero, or momentarily springing up towards infinity…" Doesn't this sound like the language of a wannabe neuroeconomist? (except that it's more flowery). Now we *do* have tools much like those Edgeworth dreamed of. If Edgeworth were alive today, would he be making boxes, or recording the brain?"

A precise hedonimeter is not available -- and probably never will be -- but neuroimaging techniques for *imperfectly* measuring hedonic states *are* available and are likely to dramatically improve with the resolution of imaging technologies. However, it remains to be seen if such hedonic measurements will be accepted by economists. It is plausible that economists will prefer to *exclusively* use revealed preferences, leaving little or no role for correlated neural activity as a complementary signal of well-being. Nevertheless, it seems likely that neural activity and self-reports will eventually be accepted as measurements that complement standard methodologies for inferring well-being. After all, revealed preference can itself be interpreted as a noisy measure of well-being (Luce 1959, McFadden 1981), so neural measures are likely to be useful supplementary covariates.[3]

Third, neuroeconomics will serve as a catalyst for model development. Neuroscientific data and neuroscientific models have inspired economists to develop many new economic models: e.g., Bernheim and Rangel 2005, Fudenberg and Levine 2006, Benabou and Tirole 2006, Brocas and Carrillo 2007 UPDATE THIS LIST OF THEORIES.

Fourth, neuroeconomics will provide a new, powerful way to test economic models which ambitiously specify both how choices depend on observables, and what computational mechanism leads to those choices. Of course, few economic models make specific neural (or even cognitive) predictions. However, when economic models do make neural predictions, these predictions provide an additional domain for testing these theories. Theories that successfully explain both choice data *and* neural data have many advantages over theories that only make choice predictions. A theory that explains both types of data will inevitably predict some surprising new effects of treatment variables on choice (besides the usual suspects of prices, information and income). For example, Weber et al (2010) were motivated by neural fRMI evidence about the circuitry of time preference computations to predict that disruption of a specific brain region (right DLPFC) would cause people to act more impatiently. As hypothesized, disruption in that area did actually change choices between immediate and delayed actual monetary amounts. This type of predicted treatment effect could have not have come from a model without neural detail.

In this list of four motivations, motivation one – neuroscience for its own sake -- is relevant primarily for neuroscientists. Motivation two – an imperfect hedonimeter -- relies on future acceptance of neural measurements of well-being. Motivations three and four, using neural

[3] Discrete choice models (e.g., Logit) have alternatively been interpreted as models with decision noise, like game-theoretic trembles, or models in which true utility has a stochastic component. In fact, these perspectives are both sensible and mutually compatible.

evidence for model selection and testing, are much more likely to prove useful and gain acceptance.

This analysis does not claim that economics can't get by without neuroscience. Economics certainly does not *have to build neural foundations*. There is no economic model that could *only* be derived with the benefit of a neuroscientific antecedent. There is no choice-based theory that can *only* be studied with neuroscientific data. However, neuroscience is useful because it can accelerate the pace of economic research. As a profession, economists are extremely adept at conjecturing detailed competing theories.

For example, there are many different theories of negative reciprocity. Is the *preference for punishing defectors* reputation-driven? Is punishment motivated by a reputation concern coupled with the implicit belief that we are always being watched, even in an "anonymous" laboratory experiment? Is punishment a knee-jerk response with evolutionary origins? Or do we get real instantaneous pleasure from punishing defectors? Distinguishing these theories with field data, or experimental choices, is challenging, though not impossible. Using a combination of choice data and neural data helps us make these conceptual distinctions, revealing that pleasure is at least part of the equation (Quervain et al XXX).

Even with a blindfold, a pedestrian could walk across a college campus. But she would move travel more efficiently without it. Likewise, economists should remove our own methodological blindfold. At the moment, the cost of wearing a neuroscientific blindfold is not great, since neuroscience is in its infancy. However, as neuroscience continues to rapidly advance it will become overwhelming clear that neuroscientific insights improve our economic vision.[4]

This chapter reviews the nascent, but rapidly growing literature in neuroeconomics, paying particular attention to experimental methods. The paper is divided into six sections, which are modular, so they don't need to be read sequentially. Each chapter was drafted by a different expert.

Section 1 discusses basic neurobiology, which is needed to understand the scientific questions and methodology (including measurement) used in neuroeconomics. Section 2 discusses neuroscience methods, with emphasis on neuroimaging and the challenges of designing experiments for subjects inside scanners. The rest of the chapter discusses four active topics of neuroeconomic research: Risk (Section 3); Intertemporal choice and self-regulation (Section 4); Social preferences (Section 5); Strategic behavior (Section 6). These do not span all parts of neuroeconomics, but they describe some areas of special interest in which progress is being made.

[4] Becker and Murphy (1988) conjecture: "People get addicted not only to alcohol, cocaine, and cigarettes but also to work, eating, music, television, their standard of living, other people, religion, and many other activities." Within their model, 'addiction' is simply adjacent complementarity in consumption (consuming more based on past consumption). However, to a neuroeconomist addiction to drugs is a biological process marked by increasing tolerance, withdrawal upon cessation, and sensitivity of use to environmental cue 'triggers' associated with past use (Laibson, 2001). So the economic and neuroeconomic approaches can be distinguished empirically. Becker and Murphy's claim about the breadth of their theory could then be tested on a neuroeconomic basis (along with using choices, price, and future price expectations).

1 Neurobiological Foundations

Neuroeconomics reflects a reductionist approach to social science that rests on two premises. First, that explanatory systems for describing human choice behavior can be developed at neuroscientific, psychological and economic levels of analysis. Second, that there will be consistent and understandable mappings among these levels of explanation. If both of these assumptions are correct, then studies of choice and decision at any of these levels can be used to inform and constrain explanatory models generated at other levels.

While the second of these premises remains controversial, it may be valuable to look to the history of the natural and physical sciences in assessing the likelihood that this will be validated by future empirical work. At the end of the 1800s a group of interdisciplinary scholars argued that quantum theory could provide a similar mapping between chemistry and physics which would allow for accelerated model development in both fields. The result was an enormously fertile period in the history of both of those disciplines and a permanent mapping between chemistry and physics. In the 1980s a similar trend could be observed in the relationship between biology and much of psychology. Only two decades later, just who is a neuroscientist and who is a psychologist can be very difficult to determine at a typical University. We believe that neuroeconomics may find itself today at the same crossroads. What this means for economics is that as these mappings are identified, a flood of algorithmic constraints from neuroscience will become available to economists. In a similar way, normative models and empirical behavioral models from economics will play a larger role in constraining neurobiological models.

An important barrier to the importation of these constraints into economics, however, is a lack of knowledge about the brain and unfamiliarity with neuroscientific vocabulary. The pages that follow therefore provide a basic primer on the vertebrate brain. For the neophyte interested in learning more about the brain we recommend an introductory undergraduate text like Rosenzweig's "Biological Psychology". For advanced material the reader is referred to standard graduate texts: "Principles of Neural Science" or "Fundamental Neuroscience". [finish cites]

The Cellular Structure of the Brain

Like all organs the vertebrate brain is composed of *cells*, self-sustaining units that are typically about a thousandth of an inch in diameter. The brain is composed of two types of cells, called *glia* and *neurons*. Glia are support cells that play structural and metabolic roles in the maintenance of the brain. It is neurons, or nerve cells, that perform computations and serve as the foundation for mental function. **Figure 1** shows a cartoon of a fairly typical neuron. The large bulbous center of the cell, or *cell body*, contains all of the machinery necessary to keep the cell alive. Extending from the cell body are long thin processes called *dendrites*. These extensions serve as the inputs to a nerve cell, the structural mechanism by which signals from other nerve cells are mathematically integrated and analyzed during neural computation. Also extending from the cell body is a single long thin process called the *axon*. The axon serves as an output wire for the nerve cell. Axons may be quite long, in rare cases almost a meter, and nerve cells use these axons to broadcast the outputs of their dendritic computation to other nerve cells, even if those recipient cells are quite distant. They accomplish this connection to other nerve cells at the end of the axon, the tips of the axons making physical contact with the dendrites of other neurons. The cellular specialization at this contact is called the *nerve terminal*. The nerve ending-to-dendrite junction allows a receiving neuron to add, subtract, multiply, divide or even mathematically integrate the many continuous real-valued signals that its dendrites receive from the nerve terminals that impinge upon it.

To better understand this process, however, we next have to understand what it means for a nerve cell to send a 'signal' to another nerve cell. Formally, signals in nerve cells are called *action potentials* (or more colloquially *spikes*) and they reflect a rather simple electrochemical process that is now well understood. Like all cells, nerve cells are surrounded by membranes that restrict the flow of chemicals both into and out of the cell (**Figure 2**). These membranes particularly restrict the flow of the positively charged atom sodium (the active ingredient in table salt). The critical feature that this regulation of flow, and the separation of electrical charge that it imposes, creates is a stable equilibrium between two physico-chemical forces. The high concentration of sodium outside the cell sets up a diffusive force which acts to equalize the concentration of sodium inside and outside the cell by driving sodium inside the cell. In opposition, an electrical force (involving the positively charged ion potassium, which is overrepresented inside the cell at equilibrium) seeks to distribute electrical charge equally by driving sodium outside the cell. Because of the construction of the membrane, these two forces reach a stable equilibrium state at which the inside carries a negative charge with regard to the outside (a measure of the electrical force) which is opposed by an equal and opposite diffusive force. This equilibrium state is called the *resting potential*, and perturbations of this equilibrium induced by transient changes in the strength of the diffusive force serve as the conceptual centerpiece for all neural computation.

These perturbations turn out to be quite easy to induce. This is accomplished by opening and closing mechanical channels that span the membrane. Consider an openable *ion channel* (**Figure 3**), a hollow tube spanning the membrane with a hole that can be opened and which when opened permits a single sodium atom to cross the membrane. When a few hundred of these channels open on a dendrite the result is that the dendrite is driven to a new equilibrium state by the movement of sodium, by diffusion, into the cell. This new equilibrium, one associated with a stronger diffusive forced created by the open channels, is characterized by a commensurate change in the electrical force, in this case a shift to a higher voltage inside the cell. What opens these tiny ion channels? The answer is that chemicals, called neurotransmitters, transiently open channels of this type located on the dendrites. Sodium channels are not, however, the only type of channel located on the dendrites. Other classes of channels can cause the local voltage to transiently shift to a lower voltage equilibrium. By mixing and matching both channel types and neurotransmitters we can therefore construct a kind of instantaneous mechanical adding machine. One neurotransmitter opens voltage increasing channels. The more neurotransmitter, the more open channels, the higher the voltage in that dendrite. Another opens voltage decreasing channels. The physical membrane reacts by effectively averaging these electrical fields and the instantaneous electrical field across the entire dendrite is thus an equilibrium state in which the voltage is a (surprisingly linear) readout of the sum of the neuron's inputs.

The next step in neural computation within a single neuron involves a nonlinear threshold. The ion channels along the axon, it turns out, are different from those in the dendrites. These ion channels open to allow sodium to enter the cell freely whenever the voltage near them exceeds a fixed threshold. Consider now what this means. Whenever the dendritic 'computation' (the summed voltage in a region of the cell) exceeds a fixed threshold, these *voltage-gated sodium channels* all open, thus driving the entire cell to a new equilibrium that has a much higher voltage. What this means in practice is that once the voltage of the cell is high enough to trigger the opening of voltage-sensitive channels in the axon near to the dendrites, those channels open. This in turn drives the voltage even higher up. That in turn activates adjacent channels in the axon that although far away from the dendrite are subsequently opened by this more proximal shift in the equilibrium voltage. What happens, thus, is a wave of equilibrium shifts, realized as a change in the electrical state of the cell, which propagates down the axon to the axon terminal. This wave of activation is the action potential and importantly it is always of the same voltage – the one

specified by the equilibrium state induced by these voltage sensitive channels. It is this mechanism that allows a cell to signal to the nerve endings, which may be a meter away, that the voltage of the cell body has crossed a specified threshold.

It is critical to recognize, however, that we have transformed a continuous and largely linear variable, membrane voltage, into a discrete single event. How then can nerve cells communicate the kinds of continuous real numbers that we need for meaningful computation? The answer is that the action potential itself is automatically reset after about a thousandth of a second. A second action potential is then generated if the voltage in the dendrites remains above threshold. Because of the mechanics of the channels, the higher the voltage the sooner this second action potential occurs. The result is that the rate of action potential generation, the frequency with which action potentials are generated, is a roughly linear function of dendritic voltage. In practice this means that the number of action potentials generated per second by a cell is the continuous variable onto which any neural calculation must be mapped. This variable ranges from about 0 to 100 action potentials per second (or Hertz, the units of frequency) for a typical neuron. Note that this is a positively valued range, which imposes some interesting computational constraints. Negative values can be encoded by assigning two neurons to the encoding, one for positive values and one for negative values. Alternatively, negative values can be encoded by defining 50 action potentials per second (or some other frequency) as '0'. Both encoding techniques have been observed in the mammalian brain for different subsystems. The range is also, in practice, finite because of limited precision at several points in the system. This can be overcome by dedicating more than one neuron to the encoding of a single real number, a technique also widely observed in the vertebrate nervous system[5].

What happens to these action potentials next, after they reach the nerve terminal? The answer is that each action potential triggers the release of a tiny quantity of neurotransmitter from each terminal (**Figure 4**). This neurotransmitter then diffuses across a truly tiny space, called a *synapse*, that separates each nerve terminal from the dendrite with which it communicates. Lying at the far side of the synapse, on the surface of the dendrite, are the same ion channels that we encountered when discussing dendritic function above. These were the ion channels that were opened or closed by neurotransmitter molecules. These neurotransmitter molecules thus serve to open ion channels in those dendrites causing the membrane of the post-synaptic cell to change voltage. This completes the passage of the signal through a single neuron and initiates a new computation at the next neuron. Neuronal computation is thus incremental and serial, with chains or networks of neurons performing parallel mini-computations is continuous time.

At a micro-scale, networks of neurons can be viewed as largely linear devices that can perform essentially any specifiable computation either singly or in groups. And a large segment of the theorists and empiricists in neuroscience devote their time to the study of neural computation at this level. Neuronal recording studies conducted by neuroeconomists in monkeys take advantage of this fact by measuring, one neuron at a time, the rate at which action potentials are generated as

[5] Let me draw attention to how obviously cardinal and linear is this discussion of firing rates as encoding schemes. To a neurobiologist, who is essentially an algorithmic engineer, this is the most natural way to imagine firing rates. Perhaps somewhat surprisingly, there is also a huge amount of data to support the conclusion that firing rates actually are linear with important environmental variables. Perhaps even more surprisingly, the activity level of a given neuron during rest actually does correspond, in most cases, to the default state of the variable being encoded. One simple example of this is the representation of the speed of a moving object in the visual system. Within a fixed range of speeds for each neuron, firing rates in cortical area MT are highly linear encoders of this highly abstract property with almost all variance accounted for by the Poisson structure of fixed neuronal noise. REF: Maunsell and VanEssen J Neurophys 1983.

a function of either the options that a monkey faces or the choices that he makes. This allows them to test the hypothesis, for example, that to within a linear transform the neurons of a particular brain region encode in their spike rate the expected utility of an option. Of course this observation implies that the kind of stable mapping rules that link chemistry and physics seem to reach from economic theory all the way down to single neuron function, a point that this chapter seeks to make clear.

A final point that needs to be made before we leave the study of neurons is that all of these processes - the generation of action potentials, the release of neurotransmitter, and the maintainence of dendritic electro-chemical equilibrium - are metabolically costly. All of these processes consume energy in the form of oxygen and sugars. In fact, this is one of the most costly metabolic processes in the human body. Over 20% of the oxygen and sugar we employ as humans is used in the brain, even though the brain represents only about 3% of the mass of the human body. So it is important to remember that more neural activity means more metabolic cost. This has two important implications. First, minimizing this activity is a central feature of the cost functions that lie behind neural computation. Second, this metabolic demand is what is measured in most human brain scanning experiments. To the degree that this metabolic cost is a linear function of neuronal activity, measurements of metabolic state reflect the underlying neural activity.

From Neurons to Networks

Studies of single neurons do show evidence of a clear mapping between economic theory and brain function, but it is also critical to understand the size of the human brain when one is considering the function of single neurons. The human brain is composed of about a hundred billion neurons. The average neuron receives, on its dendrites, inputs from hundreds of other neurons and in turn makes synaptic contacts at its nerve endings with hundreds of other neurons. If we were to imagine that 10^6 neurons encoded (for example) expected utility (to within a linear transform), and that those neurons were randomly distributed in the brain, then it would in practice be impossible to find those neurons if one was looking for them one at a time. The existence of a second hidden cost function, however, solves this problem for neuroscientists. It turns out that axons are particularly costly to maintain and the result is that evolution has shaped the human brain in a way that minimizes total axonal length. To achieve axonal minimization, two principles seem to be widely adhered to in the neural architecture. Neurons engaged in related computations tend to be grouped closely together and communication between distant groups of neurons tends to employ highly efficient coding schemes that use a minimum number of axons.

These *ex ante* constraints, and a wealth of empirical evidence, now support the conclusion that the brain is a set of modular processing stages. Discrete regions of the brain typically perform specific computations and pass their computational outputs in a highly compact form to other brain areas for additional processing. We need to maintain, however, a clear mapping between an analysis at the level of neurons and an analysis at the level of brain areas. Single neuron studies of decision making in monkeys are an example of this kind of mapping. Those studies often measure the rate of action potential generation in neurons that serve as outputs from brain areas and as such provide information at both of these levels of analysis.

Both the human and monkey brain can be divided into three main divisions based on converging evidence from developmental, genetic, physiological and anatomical sources. These three divisions are, front to back, the *telencephalon*, or forebrain, the *diencephalon* and the *brainstem*

(**Figure 5**). For the purposes of neuroeconomic study the telencephalon, which all vertebrates possess in some form, will be our focus.

The telencephalon itself can be divided into two main divisions that will be familiar to many neuroeconomists, the *cerebral cortex* and the *basal ganglia*. Of those two, the more evolutionarily ancient structure is the basal ganglia.

The basal ganglia are composed of a number of sub-regions in humans that lie beneath the cerebral cortex. There are five of these regions that are most important. The *caudate* and *putamen* together are known as the striatum. The striatum, and in particular the lower, or ventral, striatum is of particular interest because activity here appears to encode option value during choice tasks (Levy et al. 2010). These areas receive extensive inputs from the frontal cortex and send almost all of their outputs to two other nuclei of the basal ganglia, the *globus pallidus* and the *substantia nigra pars reticulata*. Speaking generally, the caudate and putamen are the main input areas of the basal ganglia and the globus pallidus and substantia nigra pars reticulata are the main output areas. These output areas project, through a dedicated relay, back to the frontal cortex. The core circuit of the basal ganglia is thus a loop that takes information from the frontal cortex and passes it back to the frontal cortex after further processing. The one remaining critical region of the basal ganglia is composed of the dopaminergic neurons of the *ventral tegmental area* and the *substantia nigra pars compacta*. These dopaminergic neurons receive projections from the output nuclei of the basal ganglia as well as from many other areas and project both to the frontal cortex and the input nuclei of the basal ganglia. The dopamine neurons have been of particular interest because there is now overwhelming evidence that these neurons encode a reward prediction error signal appropriate for error-correction based learning (e.g., Calin et al., QJE 2010).

The cerebral cortex of the telencephalon is much larger than the basal ganglia in most primate species and is surprisingly homogenous in structure. Essentially all cortex is a 6-layered sheet (**Figure 6**) with each of the layers showing very specific functional specializations. Layer 5, for example, always contains a specific class of cells that send axons out of the sheet to make connections with other distant regions in the cortex. This 6-layered structure also means that the cortex is, at least structurally, a sheet like device. This is obvious on gross inspection. The crinkled surface of the brain reveals that the cerebral cortex is a folded sheet that has been crumpled up to fit inside the skull. Beneath this folded sheet are dense runs of axons for interconnections between different places in the cortex. The sheet itself, composed largely of cell bodies, is referred to as *grey matter*. The dense runs of axons beneath it are referred to as *white matter*. For hundreds of years this sheet has been divided into 4-5 main subdivisions. These are not functional subdivisions but rather names of convenience. These main divisions are the frontal, parietal, occipital, and temporal lobes. Until recently the insula was considered an independent fifth lobe although it is now often referred to as part of the frontal lobe.

Despite this casual parcellation into lobes, until the twentieth century it was widely believed that cortex was homogenous not only with regard to its anatomy but also with regard to its function. That conclusion was successfully challenged when it was demonstrated that sub-areas in the cortex served quite specific functional roles. Ultimately, this led the famous German Neurologist Corbinian Brodmann to conclude that there are tiny differences between the anatomical structure of different regions of the cortex, differences so small that they had been overlooked in the preceding two centuries. Based on these tiny differences Brodmann divided the cortex into a large number of numerically labeled sub-areas and cortical sub-areas.

The principal Brodmann-area subdivisions, at a functional level, parcellate the cortex into a series of areas with known interconnectivities and discrete functions. Both of these properties are

important. The connectivities are surprisingly sparse in the sense that each cortical area connects with only a few other areas, and these connections are identical across normal individuals. The functions are often surprisingly discrete and now very well defined for some areas.

One final area that deserves mention anatomically is the *amygdala*. The amygdala is a portion of the telencephalon that is not classically considered part of the cerebral cortex or the basal ganglia. The amygdala is of particular interest because a wealth of studies now suggest that the psychological state of fear can be mapped to activation of the amygdala. Generalizing from these observations has led to the suggestion that psychologically defined emotional states may well map to neurally localizable activity. The good news is that this seems to be the case for fear. The bad news is that there is no compelling evidence, as yet, for such specific localization of other psychologically defined emotions.

Summary of Neurobiology

For an economist interested in neuroscience there are two central messages about the foundations of neuroscience. The first is that there seem to be clear and consistent mappings between events at the neural level and events at the behavioral level. The second, which follows from the first, is that the details of neurobiological function provide valuable constraints for economic theories. What this points out in turn is the critical need for basic neurobiological literacy amongst neuroeconomists.

Figure 1: A Neuron

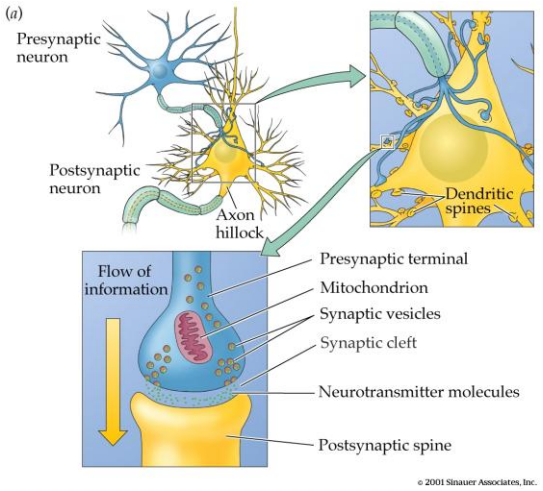


Figure 2: A Membrane

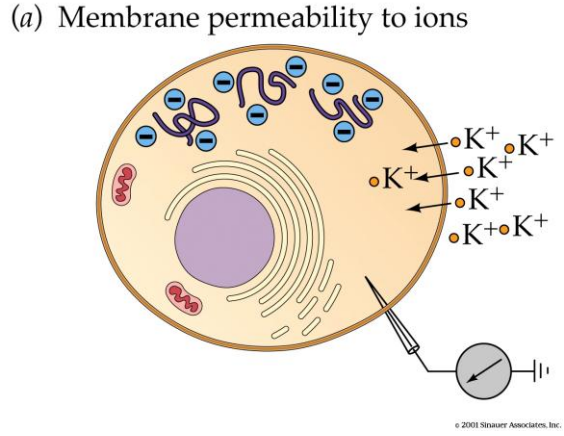


Figure 3: An Ion Channel

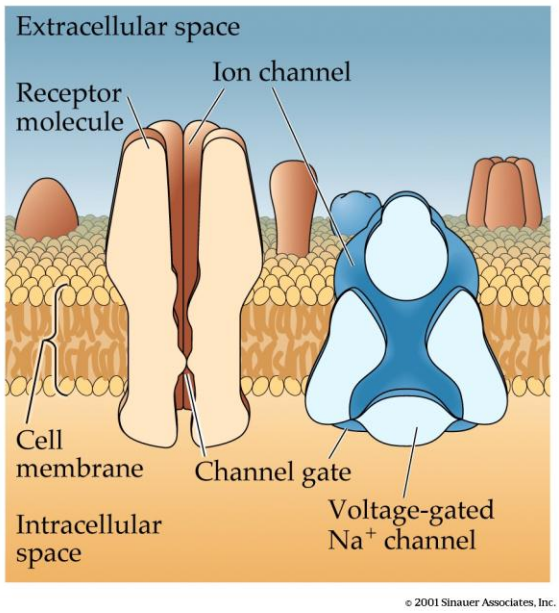


Figure 4: A synapse

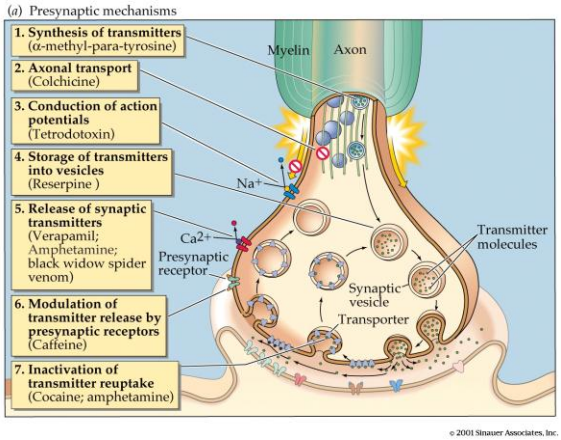
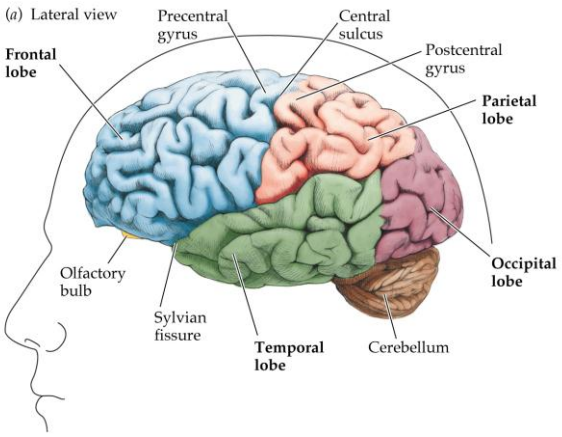
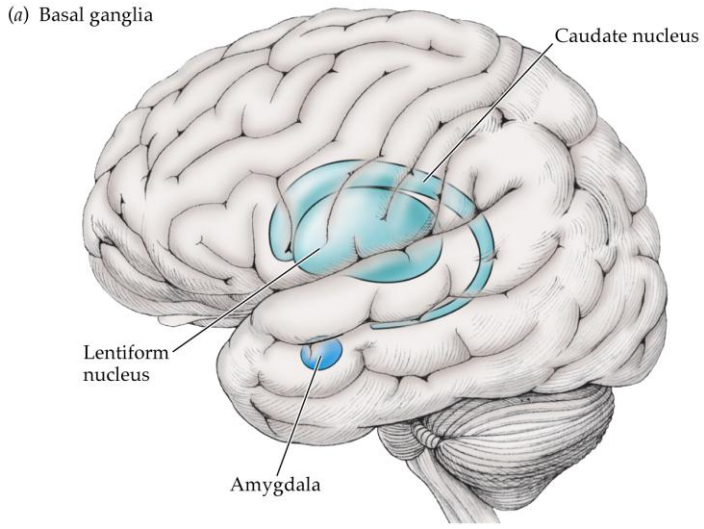
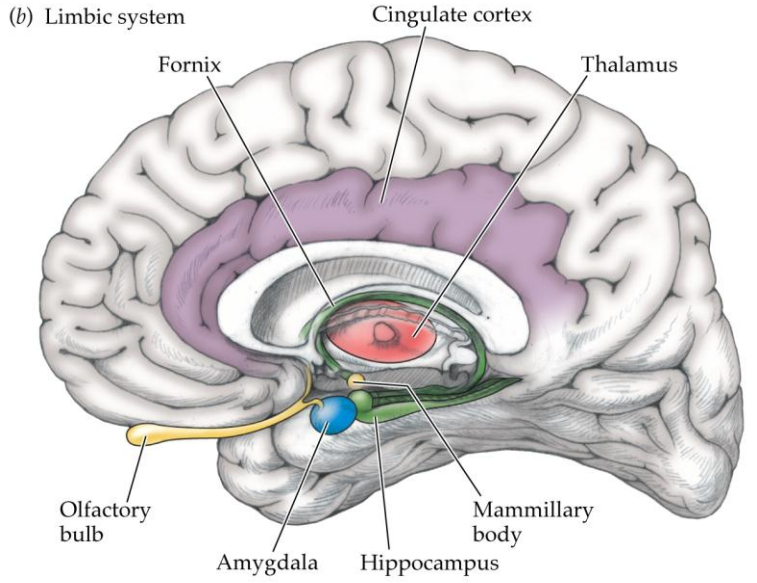


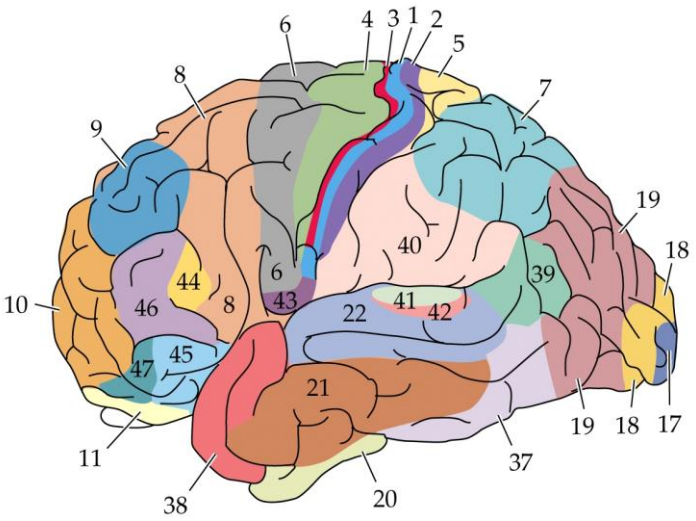
Figure 5: Brains

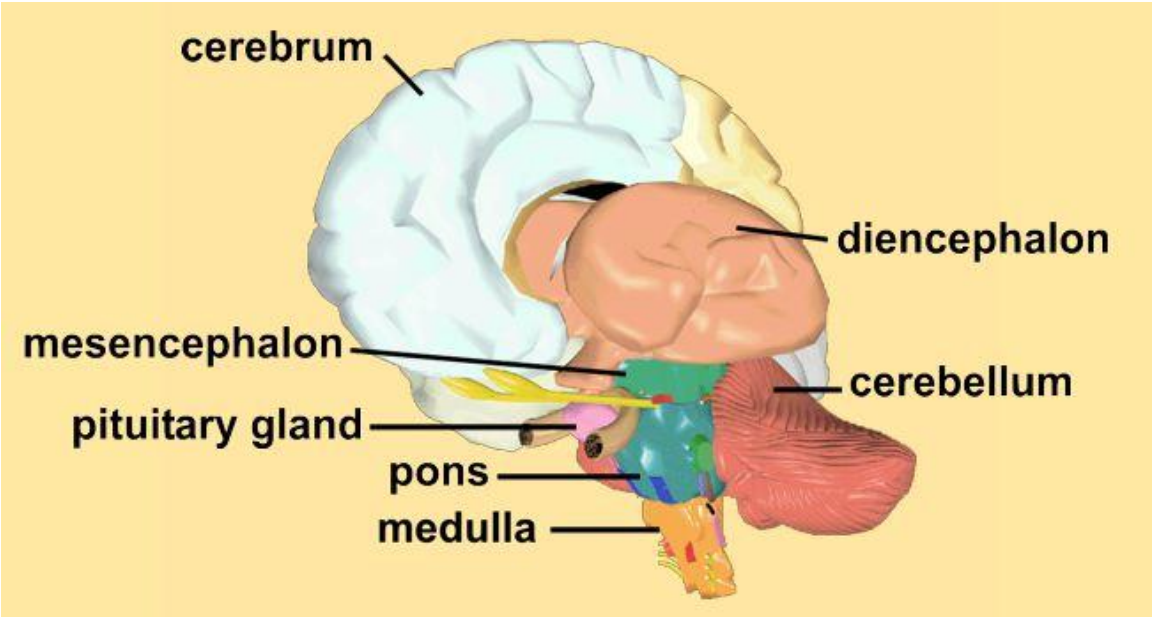






(*c*) Cytoarchitectonic map of regions of the cortex





2: Functional MRI (fMRI): A Window into the Working Brain

An understanding of the human brain remains one of the greatest challenges of science. One of the primary impediments to meeting this challenge has been the ability to measure brain activity associated with mental function. Methods for non-invasively measuring brain electrical activity in humans, or electroencelphalography (EEG), have been available for over 80 years (Berger, 1929). While these have produced useful information about the timing of some neural processes, the inhomogeneity of electrical conductivity across the brain, limits their spatial resolution. Alternative methods that provide better spatial resolution are available, such as magnetoencephalography, or MEG (Hämäläinen et al., 1993). However, like EEG, these are restricted to measuring cortical activity (where there are sufficient numbers of geometrically aligned cells to produce a coherent signal), and thus miss the operation of deeper structures thought to be involved in reward processing (e.g., basal ganglia and brainstem neuromodulatory nuclei).

To date, the most successful efforts to measure brain activity take a less direct approach than recording neural activity from the scalp. These neuroimaging methods exploit an observation first made by Roy and Sherrington in 1890 (Roy & Sherrington, 1890): that neural activity is associated with increased blood flow to the active brain region. Although the precise mechanisms that mediate the relationship between neural activity and blood flow remain incompletely understood, this relationship has been used successfully to measure regional brain activity. The first of these methods to be developed involved the injection of radiotracers into the blood stream, and the measurement of their distribution within the brain while the subject is engaged in mental activity (Phelps et al., 1975). A major advantage in these methods, including positron emission tomography (PET) and single positron emission computed tomography (SPECT), is that they can be used to radioactively label agents that selectively bind specific neurotransmitter receptors. This has been especially useful in evaluating the function of neurotransmitter systems in psychiatric disorders. However, safety limitations on exposure to radioactivity restrict the spatial resolution of the brain activation (about 5mm) and the temporal precision of the measurement (one longitudinal observation can be taken per minute). Another approach to measuring activity-related changes in blood flow uses optical recordings, which exploit signatures in the spectrum of light scattered by blood-born hemoglobin. Non-invasive optical recordings use near-infrared spectroscopy (NIRS; Villringer et al., 1993) since light in this part of the spectrum penetrates the scalp. Although the high temporal resolution, relatively low cost, and portability of this method make it useful for some specialized applications (e.g., studying infant brains), it is still limited by low sensitivity and spatial resolution. By far, the most common approach currently used to measure human brain activity is functional MRI (fMRI).

Functional MRI and the BOLD signal

The ability of MRI to detect changes in blood flow (which is referred to as the BOLD signal, for Blood-oxygen-level dependence) was first reported by three separate laboratories less than twenty years ago, in 1992 (Bandettini et al., 1992; Kwong et al., 1992; Ogawa et al., 1992). This method relies on two fortuitous phenomena of physics and physiology: 1) oxygenated and deoxygenated hemoglobin molecules have distinguishable effects on the signals detected using magnetic resonance imaging (MRI); and 2) increases in blood flow to areas of increased neural activity appear to exceed the demands of aerobic metabolism, paradoxically increasing the density of oxygenated hemoglobin. Exploiting these effects, MRI can be used to detect a blood oxygen

level dependent (BOLD) signal that is sensitive to relative changes in local blood flow. This, in turn, can be used to index neural activity. MRI can also be used to measure neural activity in other ways (e.g., using arterial spin labeling, or ASL, to directly measure perfusion; Williams et al., 1992) and to map anatomy (e.g., diffusor tension imaging, or DTI, to image fiber pathways; Le Bihan, 1995; Buxton, 2001). However, fMRI using the BOLD signal is by far the most common technique used to learn about brain function.

Because the BOLD signal reflects changes in blood flow rather than neural activity directly, it is limited in several ways. Most importantly, it responds slowly to neural activity, first appearing about 2 secs after a triggering event, peaking at about 4-6 seconds, and abating after about 10 seconds. While highly sensitive to even very brief neural events (lasting as little as 500 msec), the BOLD signal reflects these events in a delayed and diffused manner. Analyses try to compensate for this nonlinear effect (by incorporating models of the typical hemodynamic response function, or HRF). However, these rely on assumptions that are not always accurate or generalizable, and compromise precision. Because it reflects hemodynamic changes rather than direct neural activity, the BOLD signal is also limited in spatial resolution (with a current lower limit of about 1 mm).

These limitations notwithstanding, the method has proven remarkably successful in identifying neural activity associated with a wide array of mental processes. These range from visual perception and the control of overt motor actions, to subtler intervening ones such as recollection, decision making, inference and emotional evaluation. The ability of fMRI to localize such activity has been validated by comparing results with those from complementary methods, including other imaging methods, as well as simultaneous recordings of the BOLD signal and direct electrical recordings in non-human primates (Disbrow et al., 2000; Logothetis et al., 2001) and in human patients with implanted electrodes (Mukamel et al, 2005). Because it is non-invasive, and owing to the wide availability of MRI scanners, fMRI has become a mainstay of research on human brain function.

Design considerations

Scanning parameters. Several factors govern the effectiveness of an fMRI study, ranging from pulse sequence design (how the MR scanner is "tuned") and the alignment of scans within the brain, to the design of the behavioral paradigm used to engage mental functions of interest. Choice of pulse sequence has a strong impact on the nature and quality of the data acquired, but is beyond the scope of this article (the interested reader is directed to Haacke, Brown, Thompson & Venkatsean, 1999). However, it is worth noting that a typical study involves longitudinal samples from about 10,000 brain loci (about 3cc each) taken every two seconds for about 45 minutes. It is also worth noting that both pulse sequence design and scan placement can affect signal drop out (known as "susceptibility artifact"). This occurs in brain areas that are near air passages (such as the sinuses), including ones of particular relevance to decision making and valuation such as the orbitofrontal cortex (lower surfaces of the frontal lobes) and amygdala (along the inner surface of the temporal lobes). Scans can be tuned to compensate for these effects, but this can sacrifice coverage or sensitivity in other brain areas (akin to the problem of backlighting in photography). Newer hardware designs which address this problem are beginning to emerge (akin to high dynamic range [HDR] imaging in photography), and should be commonplace in the near future.

Experimental design and the subtractive method. Equal in importance to scanning considerations is the behavioral design of the experiment. The most common approach to identifying brain areas associated with a particular cognitive function uses subtractive logic (Donders, 1868/1969):

Contrast an "experimental condition" in which the participant is performing a task of interest (for example, a decision between two options) with a "control" condition in which the participant is required to process all of the same stimuli and responses, but does not engage in the process of interest (for example, observe the choice options, but simply press a button after they are seen, without choosing between them). The data are then analyzed by subtracting signals observed in the control condition from those in the experimental condition. This is usually done using simple t-tests or, for factorial designs, multiple regression or analysis of variance (ANOVA). The potential flaws of this design are obvious (e.g., the subtraction is most informative if the sensory and motor processes are carried out in precisely the same manner in the control and experimental conditions). However, as a matter of practice, this approach has been surprisingly successful (as evidenced by converging evidence using a variety of other methods).

Parametric designs. A variant on the subtractive method, that is more sensitive, is the use of a parametric design that relies on additive factors logic (Sternberg, 1969). In this case, a series of conditions are designed to engage the process of interest in a graded fashion (for example, an increasingly difficult decision). The data are then analyzed to identify areas showing a graded increase in the BOLD signal that corresponds to the experimental manipulation (e.g., Braver et al., 1997). This is usually done using regression, to identify areas in which the BOLD signal is predicted by regressors that describe the experimental manipulation(s). Like subtraction, these parametric designs are also sensitive to critical assumptions (e.g., about the functional form of neural responses and the BOLD signal's response to the experimental manipulation). Once again, despite potential pitfalls, this approach has proven to work surprisingly well (in the sense of producing results that are later corroborated by other methods).

Neural adaptation. A variant on the parametric approach takes advantage of the well-documented phenomenon of repetition suppression, a form of adaptation or habituation at the neural level (Grill-Spector & Malach, 2001; Krekelberg et al., 2006). The neural response to a preferred stimulus decreases when the stimulus is repeated sufficiently rapidly (over seconds or even minutes). This provides a method for distinguishing neural responses to different types of stimuli that do not lend themselves naturally to manipulations of strength in a standard parametric design (e.g., different categories of visual objects).

Trial sequencing. Two additional and critical design considerations are the pace of the experimental task, and how experimental conditions are organized across trials. Considering only the BOLD signal, it is ideal to separate every trial event (e.g., stimulus presentation, decision, and motor response) by at least 6 and preferably as much as 12 seconds. This allows direct discrimination of the BOLD response to each event. However, this not only compromises the rate of data collection, but also can interact with cognitive variables (such as participants' strategies and/or motivation in performing the task). Methods have been developed to analyze more rapid event-related designs (Buracas & Boynton, 2002; Burock et al., 1998; Friston et al., 1999; Liu, 2004), with events occurring as quickly as every 3-4 seconds. However, such analyses must make assumptions about the form of the hemodynamic response function (HRF) in order to "deconvolve" the BOLD signal response to a given event from overlapping effects of previous ones. Empirical studies suggest that the form of the HRF appears to be moderately consistent both across brain areas and individuals — at least within regions in which it can be directly estimated (e.g., primary sensory and motor cortex) — and so most approaches use a pre-specified, canonical approximation of the HRF. However, the extent of variation in the HRF is not yet fully understood, especially for regions in which it is difficult to measure (e.g., those supporting more abstract cognitive functions such as decision making), and thus caution is warranted. This is compounded by the fact that the HRF is best characterized in response to brief, punctuated neural events. However, many cognitive processes can be protracted (e.g., complex forms of decision

making), and therefore are more difficult to model using standard rapid event-related techniques. Although some progress has been made in this area (e.g., Donaldson et al., 2001; Greene et al, 2001; Visscher et al., 2003), it remains a challenge for BOLD-based imaging methods.

Blocked designs. The discussion above assumes that each trial is analyzed separately, responding to controlled or behaviorally-generated events (called "event-related" designs). However, sometimes it is advantageous to block trials by experimental condition, so the appropriate analysis looks for sustained activity throughout an entire block of similar events. These block designs can provide greater power to detect an effect, if the mental processes involved transpire over a longer time frame (e.g., active maintenance of a mental set; Braver et al., 2003). However, blocked designs are compromised by the fact that the predominant source of noise in the BOLD signal is low frequency (minutes), and therefore may be inextricably confounded with block effects.

Naturalistic designs. Finally, it is worth mentioning that a relatively new direction is to use more naturalistic experimental designs, in which participants engage either in self-directed tasks (e.g., reflect on the day's events) or common activities (such as movie watching). The approach to interpreting such data relies heavily on correlational analysis, either between brain regions within an individual (to identify regions of brain activity that co-vary, presumably reflecting task-relevant circuits), or across individuals (to identify regions that vary similarly in response to similar stimulus conditions). The most noteworthy example of this is work demonstrating that, over large areas of the brain, there are remarkably high correlations in brain activity across individuals watching the same movie (Hasson et al, 2004). These approaches may be moving closer to observations of brain function at a level comparable to the complex dynamics involved in naturally-occurring decision making processes.

Image Analysis

fMRI data often require extensive pre-processing in order to minimize the impact of nuisance variables (such as machine noise, head movement, etc.). Most of these methods are now standard. However, there are several important considerations that warrant discussion here, including alignment of imaging data across individuals for group averaging, corrections for multiple comparisons, exploratory analyses versus hypothesis testing, and univariate vs. multivariate methods.

Group averaging. Averaging imaging data across individuals can considerably improve power to detect subtle effects. To perform group averaging, the brains of each individual must be appropriately jointly aligned. This is complicated by the fact that human brain anatomy varies considerably across individuals. There are several methods for group alignment, that vary in sophistication by how they morph brain maps onto one another (Fischl et al, 1999; Klein et al., 2009; Talairarch & Tournoux, 1988; Woods et al., 1998; van Essen et al., 2001). However, all these methods face a common limitation: they attempt to align brains according to anatomic features, such as the shapes of the cortical folds (gyri and sulci). Unfortunately, the relationship between function and anatomic structure is not identical across individuals. For example, while the vertical meridian separating the left and right visual fields typically lies within the same fold of primary visual cortex (the calcarine fissure), its precise location (i.e, whether it lies along one bank of the fold or the other) is known to vary considerably across individuals. Thus, aligning anatomic landmarks may not succeed in precisely aligning parts of the brain that perform the same function. This can introduce noise into group-averaging, and limit spatial resolution. Methods are currently under development that align images based on functional (rather than anatomic) landmarks (e.g., Sabuncu et al., 2010). Success in this effort should considerably

improve the sensitivity and spatial resolution of fMRI, while also providing new information about features of functional organization that are universal across brains.

Exploratory analysis and multiple comparisons. Whether analyzing images from a single brain or multiple brains, most methods apply variants of the general linear model (t-tests, ANOVA, or linear regression). These tests are typically applied independently to each voxel (volumetric pixel) within the image. This step is an exploratory analysis designed to determine which voxels (or clusters of adjacent voxels of a specified size) show a significant effect of the experimental manipulation. Voxels that meet a specified level of statistical significance are then shown (usually by colors indicating their level of significance) in an activation map. One problem with this approach is that image sets are usually made up of a large number of voxels (at least 10,000 and sometimes over 100,000). Thus, the threshold used for statistical significance must be corrected to take account of this massive number of comparisons, and avoid a preponderance of Type 1 errors ("false positives"). The simplest ways of doing this is to divide the threshold by the number of comparisons (Bonferroni correction). However, this risks being overly conservative (resulting in type II error, in which a genuine effect does not 'survive' this correction and appears to be insignificant). This has driven the development of more sophisticated methods, such as cluster size thresholding which takes advantage of the fact that voxels showing truly significant effects are likely to be contiguous with one another (Forman et al., 1995; Poline et al., 1997). However, these methods can be complex, and subject to misuse (Smith & Nichols, 2009; Vul et al., 2009). Therefore, it is important to attend carefully to the assumptions they make (e.g., about the independence of voxels).

Hypothesis testing and regions of interest. An alternative to the use of whole brain, exploratory analyses is to specify, a priori, regions of interest (ROIs) in which effects are expected to occur, and then restrict hypothesis testing to those areas. This limits the number of comparisons, and thus lowers the expected false discovery rate. However, when a significant effect is observed in an analysis restricted to a given ROI, it is not possible to assert that the effect is specific to that brain region since others have not been tested. In practice, the best studies use a combination of the methods described above, initially using exploratory methods to identify regions of activity, and then confirming positive findings in subsequent experiments using an ROI-based, hypothesis-testing approach. The most solid findings come from a sequence of experiments and methods (ideally coming from different research groups) proceeding in this way.

Correlational analyses. The methods described above are being combined increasingly with correlational analyses, that examine the relationship of activity within specified voxels both to other voxels within the brain (to identify task-relevant circuits, and sometimes referred to as "functional connectivity analyses"), as well as to other physiological variables of interest (such as galvanic skin response, pupil diameter, eye movements, etc.), behavior (such as reaction time, accuracy, decision outcomes, etc.), and psychometric and demographic factors (such as personality, age, gender, etc.). Such analyses have the potential to provide valuable information about the relevance of observed neural activity to mental function and behavior. However, such analyses also carry risks that have recently been the subject of some attention (Kriegeskorte et al., 2009; Vul et al., 2009). In particular, such analyses must attend to the same problem of multiple comparisons (in this case, the number of correlations) as other analysis methods.

Multivariate pattern analysis. Finally, perhaps the most important development in the analysis of neuroimaging data has been the move from univariate methods to multivariate pattern analysis (MVPA). Univariate methods, such as those described above, analyze images voxel by voxel seeking to identify peaks of activity (i.e., voxels or voxel clusters that exceed a statistical threshold). However, this almost certainly does not correspond to how the brain functions.

Rather, computational activity is distributed over many regions, some of which may be more subtly engaged — but no less important — than others. This has recently been addressed by the application of machine learning classifier algorithms. These are "trained" on one set of imaging data to identify distributed patterns of activity that reliably predict specific mental states or behaviors (e.g., the perception of a particular type of object, or a particular outcome of a decision). The patterns of activity identified in the training data are then tested on a separate set of data, to determine the generality of their ability to predict mental states or function. Such methods have proven to be successful in a variety of domains, including the ability to identify the orientation of a line (Kamitani & Tong, 2005) or class of object being visually observed (Haxby et al., 2001), the class of an object being recollected (Polyn et al., 2005), the syntactic class of a linguistic stimulus (Mitchell et al., 2008), and the value of public goods in a designed mechanism of exchange (Krajbich et al., 2009). These methods represent the leading edge of neuroimaging research, promising to greatly enhance the sensitivity with which fMRI (and other methods) can be used to track neural activity underlying ongoing mental processes in human participants.

3. Risky choice

This section describes neural activity during risky choice. There are three topics: statistical moments and evaluation of risky choice; prospect theory; and causal experiments and their implications for economics.

Statistical moments

One popular model of risky choice is that statistical moments of reward distributions are weighted and integrated to form a choice value. This approach is popular in finance, where risk and return of asset values are integrated to determine value, and in behavioral ecology studies, where animals are assumed to respond to mean and variance in foraging for food. A moments-based approach also follows from a Taylor expansion of expected utility, so it should approximate choice for local small-scale decisions.

Several studies indicate that the mean and variance of rewards of different types are encoded in brain activity (e.g., Figure 1, from (Platt and Huettel, 2008)). Average mean reward seems to activate striatal regions (Preuschoff et al., 2006). The striatum is activated by many different types of rewards including money, attractive faces, anticipation of curiosity-provoking trivia, and prediction error.

The variance of rewards, often thought of as risk, seems to activate the insula, a region involved in interoceptive integration of emotional and cognitive information (Craig, 2002;Mohr et al., 2010). One study suggested, from the time course of BOLD signals, that expected reward evaluation occurs rapidly, in striatum, and variance response occurs more slowly (a couple of seconds later) in insula cortex (Preuschoff et al., 2006).

Prospect theory

Prospect theory is a psychophysically based theory of how risks are evaluated and combined, which makes small modifications from expect utility theory. The central modification is that outcomes are encoded relative to a reference point. In addition, the decision disutility from anticipated losses is assumed to weigh disproportionately more strongly than gains, captured by a

loss aversion parameter lambda. Objective probabilities are assumed to be weighted nonlinearly, so that low probabilities are overweighted and higher probabilities are underweighted.

Establishing the neural circuitry underlying prospect-theoretic valuation, and comparing it with candidate circuitry for expected utility, could prove to be a fruitful way to compare theories and develop new predictions (based on the location of apparent circuitry (see (Fox and Poldrack, 2008)).

An implication of reference dependence is that descriptions of choices which are equivalent in their consequences but differ in the implicit reference point, could lead to different choices. For example, in one study with medical students and doctors in leading hospitals, the subjects preferred radiation as a treatment for cancer because the immediate mortality rate from surgery was 10%, (and zero for radiation) and the five-year mortality rate for surgery was 66%, rather than 78% for radiation treatment. However when exactly the same statistics were described as survival rates, that is, the percentage of people surviving rather than dying, surgery looked much more attractive because the short-run survival rates were 90% and 100%, and the long-run survival rates 34% rather than 22%.

One imaging study looked at brain activity during response to loss and gain framed choices for monetary gambles (De Martino et al., 2006). They looked for an interaction effect between choosing sure things rather than gambling, which operates differently for gains and losses (a neural source of reflection effects). They found activity in the amygdala in response to the typical default choice, which is a sure thing for gains and a gamble for losses. Dorsal medial cingulate cortex was also differentially activated in the unusual choices (gambling over sure gains and accepting a sure loss). A further study showed that subjects with SS alleles of 5HTT neurotransmitters showed larger framing effects (Roiser et al., 2011). Furthermore, the pattern of activity in amygdala in response to framing is evident in the SS allele subjects but is completely absent in subjects with different genetic makeup. A further study (De Martino et al., 2008) showed that autistic adults are less susceptible to this framing effect. This is ironic, because framing effects are often thought of as violations of rational choice principles; if the autistic subjects obey those principles more than normal subjects, then perhaps they should not be considered benchmarks of rationality.

A key component of prospect theory is loss aversion, the disproportionate disutility from losing relative to equal sized gains. Until recently, most evidence of loss aversion in decisions is inferred from human choices between monetary gambles with possible gains and losses. However, there is also evidence of loss-aversion in monkeys trading tokens for stochastic food rewards (Chen et al., 2006) and associated evidence of endowment effects in monkeys (Lakshminarayanan et al., 2008).

An early fMRI study (Tom et al., 2007) showed comparable neural activity in several value related brain regions during evaluation of gambles with increased gains and decreased losses. In this study the neural loss aversion, the difference in brain response to potential loss dollar for dollar, relative to potential gain, was strongly correlated (r=.85) with the degree of loss aversion inferred behaviorally from choices among gambles. While that study indicated a common basis for reduced loss and increased gain, other studies indicate different locations of brain activity for loss and gain. For example, Yacubian et al. (Yacubian et al., 2006), in a study with usually large sample sizes, found gain activity in VStr, and loss activity in amygdala and temporal lobe regions lateral to the stratum. A later study showed that two patients with selective bilateral amygdala lesions exhibited very little loss aversion (De Martino et al., 2010).

Prospect theory also posits that attitudes toward risk depend not only on valuation of outcome utility, but also on weighting of likely outcomes in the process of decision. A simple way to account for these effects is by weighting an objective likelihood of an outcome, p(X), by a transformed function $\pi(p(X))$. Several parametric weighting functions have been suggested and estimated (e.g., (Abdellaoui et al., 2010)) but we focus on the simple one-parameter function $\pi(p)=1/\exp([\ln(1/p)^{\gamma}]$ (Prelec, 1998). This function is equivalent to linear weighting of objective probability when $\gamma=1$, has increasingly nonlinear inflection for $\gamma<1$, and always rotate around a pivotal probability $p^*=1/e=.37$ (at which point $\pi(p^*)=p^*$). Some field studies of game shows and a huge sample of horse racing bets (Snowberg and Wolfers, 2010) indicate overweighting of low probabilities too.

The neural literature indicates some biological correlates of nonlinear weighting, but in much more different regions and methods than studies discussed previously in this section. An early study using a titration procedure to match gamble value (Paulus and Frank, 2006) linked inflection of $\pi(p)$ to activity in anterior cingulate (ACC). A simpler later study by Hsu et al. (Hsu et al., 2009) discovered neural activity in VStr in response to valuation of different outcome probabilities in which the neural response function matched reasonably closely the inflection derived simply from analysis of choices (Figure 2). Hsu et al. also found a modest neurometric link between variation across subjects in behavioral nonlinearity of weights and neural activity associated with nonlinearity. However, Tobler et al. (Tobler, 2008) found signals associated with nonlinearity only in left DLPFC.

Takahashi et al. (Takahashi et al., 2010) correlated D1 dopamine receptor density with more linear probability weighting (which is also associated, in the estimated Prelec (1998) function, with higher weights on all probabilities and hence more attractive valuation of gambles). Finally, Wu et al. (Wu et al., 2009) used a motor task in which "risky choice" is equivalent to reaching very rapidly (<700 msec) to a narrow physical target in order to get a large reward (a slow reach earns nothing). They estimate that low probabilities are actually underweighted in the implicit motor valuation of reward.

Causal manipulations

Conventional economic analyses typically draw predictive power by assuming stability of preferences, using previous choice data to infer preferences (e.g., by estimating demand elasticities), then—holding preferences fixed-- predicting a comparative static change in choices based on changes in information, prices, or income. However, as the neural circuitry underlying choice becomes better understood, it will be possible to causally influence neural computations reliably in various ways, and thereby change choices.[6] Indeed, several studies have already shown such causal influences in choice among risky financial gambles.

Risk-aversion seems to be causally increased by: Experiencing stress (induced by immersion of hands in cold water) (Porcelli and Delgado, 2009); stimulation ("up-regulation") of rDLPFC using tDCS [transcranial direct stimulation] (Fecteau et al., 2007); seeing negative-affect images before choice (Kuhnen and Knutson, forthcoming); and eating food (Symmonds et al., 2010). Risk-seeking seems to be causally increased by: Disrupting right DLPFC (Knoch et al., 2006); stimulation using tDCS in older adults (Boggio et al., 2010); lowering serotonin in macaques by depleting tryptophan (Long et al., 2009). Loss-aversion can be down-regulated by a

[6] These types of causal influences have been for a long time using pharmacology, and techniques like TMS to affect vision and motor movements.

perspective-taking instruction to "think like a trader" and combine losses and gains mentally (Sokol-Hessner et al., 2009). fMRI indicates this down regulation works, to some extent, by reducing amygdala activation in response to loss (Sokol-Hessner et al., 2010).

There are two lessons from these biologically causal experiments: First, exogeneous changes to the neural circuitry which makes computations leading to risk-avoiding behavior can directly change choices. These effects are not due to changes in prices, information, or constraints (in any typical sense). These effects therefore suggest a possible expansion of the rational choice view in economics to include computational circuitry. Eventually we will understand the conditions under which that computational circuitry produces choices which approximate constrained-optimal rational choice as in consumer theory.

But more generally, the ability to cause change is useful as a tool to test the depth of understanding of how the circuitry works in general. And ideally, some of these results will invite new economic hypotheses about how exogeneous changes in the economic environment will influence neural computation, and hence predict changes. For example, if causally disrupting a brain region involved in inhibition and self-control reduces self-control, and external events also place a burden on activity in that region (mimicking disruption), then one can predict that the disruptive events will affect economic choice. What these new hypotheses are, and how well their effects can be seen in highly-aggregated data, remains to be seen.

Logical rationality and biological adaptation

Finally, we note an ironic result emerging from neuroeconomics about the conflict between notions of rationality. Patients with brain damage do not exhibit the Ellsberg paradox (Hsu et al., 2005) or loss-aversion (De Martino et al., 2010). Autistic patients exhibit reduced gain-loss framing effects (De Martino et al., 2008). If the sure-thing principle, description-invariance (no framing) and dynamic consistency are considered principles of normative rationality, then why does damaging or temporarily disrupting the brain push behavior toward closer adherence with those normative principles?

One possibility is that these few studies are flukes and both damaged brains and highly sophisticated brains can adhere to rationality for different reasons. Another possibility is that "normal" behavior is adapted to solve evolutionary challenges of survival and reproduction, while conserving phylogenetically-old regions inherited from other species and adding "kludges" (Ely, 2011). The result of that type of incremental adaption will sometimes violate principles of normative rationality. Of course, this possibility suggests a course that economists and biologists have been pursuing relatively recently, of deriving choice architecture from evolutionary principles instead of logical ones.

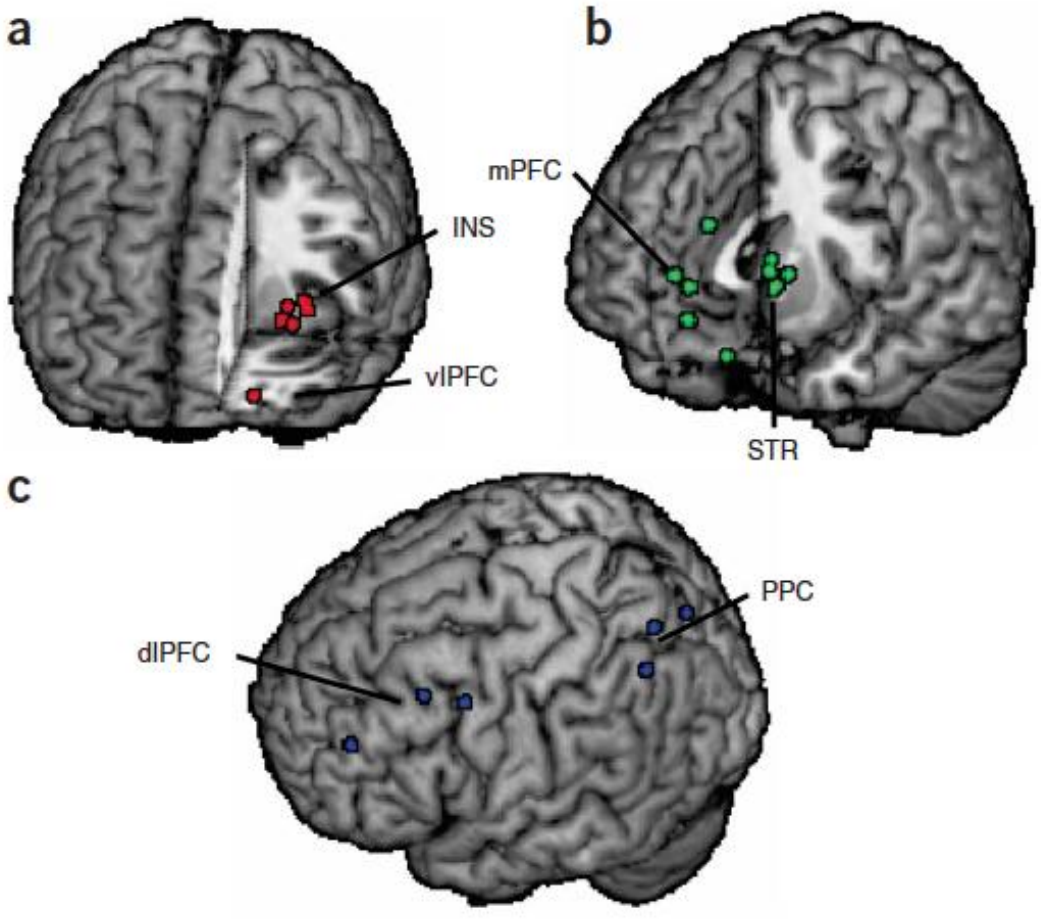


Figure 1 Brain regions implicated in decision making under uncertainty. Shown are locations of activation from selected functional magnetic resonance imaging studies of decision making under uncertainty. (**a**) Aversive stimuli, whether decision options that involve increased risk or punishments themselves, have frequently been shown to activate insular cortex (INS)[33,52,53,58] and ventrolateral prefrontal cortex (vlPFC)[61]. (**b**) Unexpected rewards modulate activation of the striatum (STR)[43,46,53,59,76], particularly its ventral aspect, as well as the medial prefrontal cortex (mPFC)[43,53,61,76]. (**c**) Executive control processes required for evaluation of uncertain choice options are supported by dorsolateral prefrontal cortex (dlPFC)[52,58] and posterior parietal cortex (PPC)[33,34]. Each circle indicates an activation focus from a single study. All locations are shown in the left hemisphere for ease of visualization.

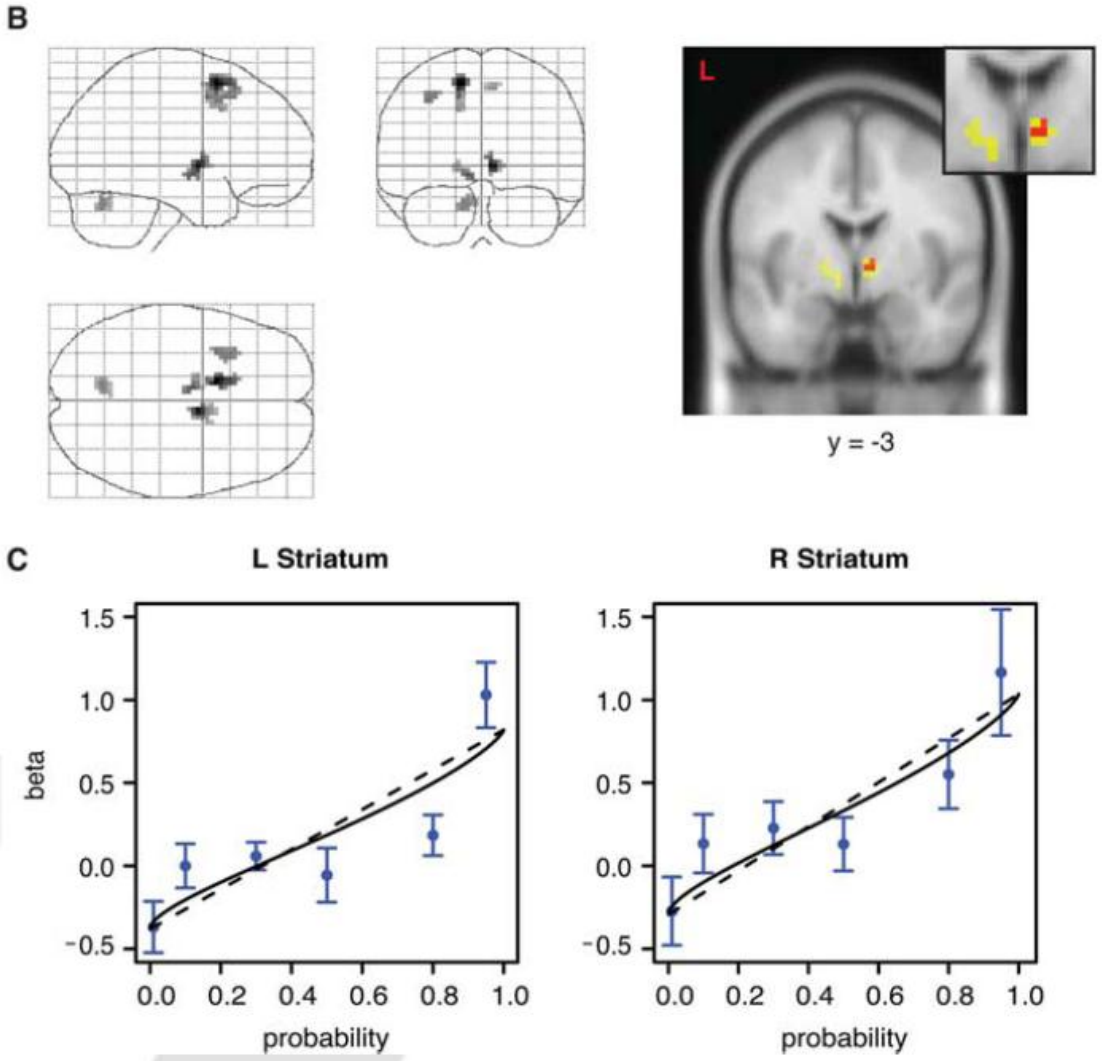


Figure 3. ***A***, The analysis decomposes expected reward responses into two terms, the linear component in p (left, dashed line) and the nonlinear $\pi(p) - p$ component (right). ***B***, Glass brain and coronal section of activations to both linear p and nonlinear $\pi(p) - p = 0.77$. Red, Regions where both linear and nonlinear terms are activated at $p < 0.001$, excluding regions where linear and nonlinear terms are significantly different at $p < 0.9$; Yellow, Regions where both linear term is activated at $p < 0.001$ and nonlinear term at $p < 0.005$, excluding regions where linear and nonlinear terms are significantly different at $p < 0.5$. For additional coronal sections, see supplemental Figure S2 (available at www.jneurosci.org as supplemental material). ***C***, Activation in left and right striatum regions (BOLD regression β coefficient) shown in ***B*** for both probability and nonlinear deviation (from yellow regions), showing that responses to both variables is insignificantly different. ***D***, Normalized GLM β coefficients of BOLD signal activation for extracted voxels (blue dots) in the left and right striatum coronal section shown in ***B*** (yellow) and Prelec function with group behavioral parameter ($\alpha = 0.77$) inferred from choices (solid black line).

Fig 2 Hsu et al.

4 Intertemporal choice and self-regulation

Intertemporal preferences continue to be one of the most active research topics in the field of neuroeconomics. In the past decade, researchers have identified many robust empirical regularities, including numerous phenomena that have been shown to be mediated by biological mechanisms. We begin this section by summarizing the body of empirical results in the intertemporal choice literature, including neuroimaging results.

There is no consensus on the theoretical interpretation of these empirical regularities (e.g., Rustichini 2008). The competing theories can be divided into three classes: multiple-self models with selves that have overlapping periods of control (e.g., Thaler and Shefrin 1981), multiple-self models with selves that have non-overlapping periods of control (e.g., Laibson 1997), and unitary-self models (e.g., Gul and Pesendorfer 2001). We turn to these theories at the end of the section and relate them to the available evidence.

Empirical regularities

The intertemporal choice literature is vast. In this subsection, we attempt to summarize some of the key empirical findings, giving extra weight to the findings with a biological interpretation.

Discount rates inferred from subject choices between smaller-sooner vs larger-later rewards, are anomalously high. Imputed discount functions have a higher discount rate at short horizons than at long horizons (e.g., Ainslie 1991, Thaler 1981, Kirby 1996). Although these findings are often cited as the basic foundational results in the intertemporal choice literature, they have turned out to be far more complicated and problematic than originally believed (for reviews see Frederick, Loewenstein and O'Donoghue 2004; Chabris, Laibson, Schuldt 2008; Halevy 2011). In fact, it is now clear that they these empirical patterns were originally misinterpreted as support for hyperbolic discounting. We summarize five of the most important methodological problems/paradoxes in the interpretation of these results. First, most of the discounting results in the literature are partially driven by reliability bias: subjects probably view later rewards as riskier than sooner rewards even when the researchers try very hard to equate reliability across different dates of reward delivery (see Andreoni and Sprenger 2010 for one way of modelling such an uncertainty effect). If such reliability bias is present and overlooked by the researcher, inferred discount rates will be biased up. Second, there is a preference for 'as soon as possible' reward receipt, even when the soonest reward is *not* available in the present (Kable and Glimcher 2006). This may also be a consequence of reliability bias: sooner rewards may be viewed as more reliable even when the "soonest" possible reward is in the future. Third, measured discount rates display sub-additivity bias (Read 1998): the product of the discount factor measured from t to $t+1$ and the discount factor measured from $t+1$ to $t+2$ is far lower than the discount factor measured from t to $t+2$. Relatedly, (Benhabib and Bisin 2006) have argued that time preferences measured in the laboratory include a 'fixed cost of delay' that is insensitive to the length of the delay. Moreover, Zauberman et al (2009) argue that subjects transform time delays with a log scale, thereby creating the observed gap between high short-run and low long-run discount rates. Fourth, the discount rates measured in laboratory experiments are too high to be consistent with commonly observed, voluntary behaviors, like accumulating retirement savings and home equity (however, Harrison et al 2009 points out that small stakes curvature in the utility function reduces this tension). Fifth, in principle, smaller-sooner vs. larger-later money choices shouldn't measure discount rates at all, but should instead measure the rate of intertemporal transformation (i.e., the

relevant interest rate for borrowing or lending; see White et al 2011). For all of these reasons, there is a growing recognition that discount rates imputed from sooner-smaller vs. larger-later reward experiments are difficult to interpret. Indeed, confounding factors like reliability bias, related uncertainty effects, and sub-additivity effects, may swamp the underlying goal of measuring time preferences.

Static choice problems with temptation goods generate preference reversals. For example, Read and Van Leeuwen (2001) ask their subjects to choose a snack to be eaten one week later. Subjects tend to choose healthy snacks. One week later, the subjects are told that the researchers lost the paperwork and therefore the subjects must again pick a snack, which will now be eaten immediately. Now preferences shift toward preference for the unhealthy snacks. Qualitatively similar reversals have been documented with other studies (e.g., Loewenstein and Read, 2004; Oster and Scott-Morton 2004).

Economic agents appear to be counterfactually optimistic about their future likelihood of engaging in patient behavior. Dellavigna and Malmendier (2004, 2006) use data on the menu of gym fees (e.g., annual, monthly and per-visit), the frequency of membership terminations, and the frequency of gym visits (measured with swipe-cards) to infer that gym members have an excessively optimistic view of their own likelihood of future exercise.

Economic agents are willing to pay for commitment. Ashraf et al (2004) find that one-quarter of their (rural Indonesian) subjects are willing to put some of their savings in an illiquid account with the same interest rate as an alternative liquid account. Beshears et al (2011) document similar behavior, even when the illiquid account has a slightly lower interest rate than the liquid account. Moreover, Beshears et al (2011) find that laboratory savings accounts attract more (real stakes) deposits the higher the penalty for early withdrawal, holding all else equal. Numerous studies have documented a demand for commitment: e.g., Wertenbroch (1998); Ariely and Wertenbroch (2002); Karlan, Gine, and Zinman (2009); Kauer, Kremer, and Mullainathan (2010); Houser, Schunk, Winter and Xiao (2010); Royer, Stehr, and Sydnor (2011); Alsan, Armstrong, Beshears, Choi, del Rio, Laibson, Madrian, Marconi (2011).

Imputed discount rates are negatively correlated with scores on IQ tests. In both children (Benjamin, Brown, and Shapiro 2006) and adults, (Burk et al 2003, 2006), high scores on tests of intelligence or cognitive function correlate with low rates of measured time discounting (see Shamosh and Gray for a review 2009).

Subjects are less patient when placed under cognitive load. When subjects are asked to remember a relatively long string of digits, their intertemporal choices become more impatient (Shiv and Fedorikhin 1999; Hinson, Jameson, and Whitney 2003). This effect has been produced with both food rewards and monetary rewards.

Subjects are less patient when they are primed with affective cues; likewise, subjects are more patient when they are primed with abstract cues. For example, Rodriguez, Mischel and Shoda (1989) find that children are *less* willing to wait for food rewards when the food is visible. Likewise, children are more willing to wait for food rewards when asked to think about the rewards abstractly (e.g., think of pretzels as logs and marshmallows as clouds). Loewenstein (1996) and Burns et al (2005) review many different visceral/affective manipulations. In a neuroimaging experiment, Albrecht et al (2010) report that subjects choose more patiently and show less affective engagement when (i) they are making choices for themselves that only involve options in the future, or when (ii) they are making choices for someone else.

Subjects show diminished willpower after performing earlier, disimilar tasks that require

willpower. For example, Baumeister and Vohs (2003) show that subjects are less able to sustain pressure on a hand-grip after suppressing the expression of emotion while watching an upsetting video.

The willingness to delay gratification develops from birth to age 20 in parallel with the pre-frontal cortex (PFC). The PFC reaches its approximate terminal anatomical structure more slowly than other brain regions (Green, Fry, and Myerson 1994). Dense neural networks are formed and pruned at least through the early 20's. It is not known whether the association between PFC development and the willingness to delay gratification is causal or correlational, though brain injuries to the PFC provide some support for the causal interpretation ().

The tendency to frequently delay gratification is correlated with cross-species variation in the proportionate size of the pre-frontal cortex. Krietler and Zigler (1990) argue that cross-species variation in domain general patience is well-correlated with the ratio of PFC volume to total brain volume. However, this position has been challenged by research that has argued that high rates of impatience in non-human species are observed in food deprived animals, invalidating comparisons to relatively sated human suspects (e.g., Rosati et al 2007).

The analytic cortex (PFC and parietal cortex) has a low sensitivity to reward delay and the meso-limbic dopamine reward system has a high sensitivity to reward delay. McClure et al (2004, 2007) find that moving a reward further away in time causes the BOLD signal in the analytic cortex to decline relatively little. By contrast, the dopamine reward system displays a much more rapid decline in activation as rewards are delayed. Similar results have been reported by Albrecht et al (2010).

The analytic cortex is more active when a delayed reward is chosen over an immediate reward. McClure et al (2004, 2007) find that the BOLD signal in PFC and parietal cortex is stonger when a delayed reward is chosen relatively to trials in which an immediate reward is chosen. Hare et al (2009) find that left DLPFC is active when subjects reject a good tasting, unhealthy snack in favor of a neutral alternative reward.

The dopamine reward system has a decline in activation that follows a hyperboloid that matches the valuation function implied by choice. Kable and Glimcher (2007) estimate discount functions using choice data and find that the BOLD signal in the mPFC matches the same pattern of decay. McClure et al (2007) find a similar pattern of declining activation in the dopamine reward system.

Exogenously disrupting normal functioning of the lateral pre-frontal cortex (LPFC), causes choices between now-vs-later rewards to shift towards the now option, but does not affect choices between rewards that are both delayed. Figner, Knoch, Johnson, Krosch, Lisanby, Fehr and Weber (2010) show that disruption of left, but not right, LPFC with low-frequency repetitive transcranial magnetic stimulation (rTMS) increased choices of immediate rewards over larger delayed rewards. rTMS did not change choices involving only delayed rewards or valuation *judgments* (in contrast to choices) of immediate and delayed rewards. This paper provides causal evidence for a neural lateral-prefrontal cortex–based self-control mechanism in intertemporal choice.

Multiple-self models with selves that have overlapping periods of control

Three classes of models have been used to organize and explain these findings. We first discuss multiple-self models that have *overlapping periods control*. We then discuss multiple self

models with non-overlapping periods of control. Finally, we discuss unitary self models with dynamically consistent preferences. We emphasize that all of these models have been set up so they make similar qualitative predictions. Hence, they are difficult to distinguish empirically.

Some multiple self models posit the co-existence of multiple neural systems with occasionally conflicting goals/preferences. These systems struggle to control or influence the choices of the decision-maker. Models in this class first gained wide acceptance after Freud argued that human choice is explained by an ongoing conflict among a conscientious superego, a self-interested ego, and a passion-driven id. Related ideas were also advocated by Smith (1775) who drew a distinction between people's "interests" and their "passions." Smith frequently discussed internal struggles between these conflicting sets of preferences. Frederick and Kahneman (2005) and Kahneman (2011) have also developed such models.

In the modern psychology literature dualities are drawn between controlled and automatic cognition (Schneider & Shiffrin, 1977), cold and hot processing (Metcalfe and Mischel, 1979), System 2 and System 1 (Frederick and Kahneman, 2002), deliberative and impulsive decision-making (Frederick, 2002), conscious and unconscious processing (Damasio, Bem), and effortful and effortless systems (Baumeister). Neuroimaing research has led to theories that locate system with impatient proclivities in the meso-limbic dopamine system and a system with relatively patient goals in the PFC and parietal cortex (McClure et al 2004, 2007; Hare et al 2009). These authors argue that the PFC is the seat of self-regulation, self-control and executive function. Individuals with a comprised PFC (e.g., due to cognitive load, lack of cognitive development, willpower exhaustion, injury, or interventions like transcranial magnetic stimulation) are more likely to make relatively impatient choices.

Some economists have also proposed two-system models, including models that contrast "planner" and "doer" systems (Shefrin and Thaler, 1981), patient and myopic systems (Fudenburg and Levine, 2006), and abstract and visceral systems (Loewenstein & O'Donoghue 2006; Bernheim & Rangel, 2003).

Multiple-self models with selves that have non-overlapping periods of control

Researchers have also advocated models with dynamically consistent preferences generated by a unitary self at each point in time. Strotz (1957) was the first to propose such a framework, though his ideas were anticipated by Ramsey (1928) and Samuelson (1931). Strotz's ideas were applied by Laibson (1997) and O'Donoghue and Rabin (1999), who used a model of intergenerational preferences proposed by Phelps and Pollak (1968) to study intra-personal discounting. This model is sometimes referred to as quasi-hyperbolic discounting, present bias, or hyperbolic preferences. In this model, the agent has a well-defined (unitary) set of preferences at each point in time. But the agents' preferences at date *t* conflict with the agent's preferences at all future dates. If the agent anticipates these conflicts, she will attempt to constrain or commit her own future behavior.

More formally, the model posits that the discount function at date t is given by 1 (t=0) and discount function values at times t>1 are $\beta\delta^t$ where β and δ are weakly bounded between 0 and 1. To understand the mechanics of this model, consider the illustrative case $\delta=1$. Now the model implies that current rewards have full weight, and any future reward has weight β. It is easy to see how this framework generates dynamically inconsistent preferences (and therefore a potential taste for commitment). From the perspective of date 0, dates 1 and 2 both have weight β so a unit reward at date 1 is worth just as much as a unit reward at date 2. But from the perspective of date 1, a unit reward at date 1 is worth $1/\beta$ times the value of a unit reward at date 2. Hence, the unitary self at date 0 and the unitary self at date 1 don't agree on the relative value of rewards at

dates 1 and 2.

McClure et al (2004) point out that the β-δ model can *also* be interpreted as a model with multiple *simultaneous* selves. Specifically, posit the existence of two selves. One self is an exponential discounter with discount factor δ. A second self is an exponential discounter with an arbitrarily small discount factor. Suppose that the two selves combine their preferences with weights β and 1-β. Then the aggregate (weighted) preference is 1 for immediate rewards and $\beta\delta^t$ for future rewards).

Unitary-self models

In the last decade, researchers have realized that phenomena like commitment are not necessarily inconsistent with unitary self models that feature dynamically consistent preferences. These models assume that agents have preferences over choice sets. Specifically, agents may prefer not to have an option in their choice set, even if they do *not* pick that alternative. For example, an agent on a diet may find exposure to a tempting food aversive even if that tempting food is in fact not consumed. Dekel et al (2008) and Gul and Pesendorfer (2001) have proposed models in this class. Laibson (2001) and Bernheim and Rangel (2004) propose related models in which menu-based temptation effects are endogenously dependent on past associations between cues/menus and rewards. Such endogenous temptation models are based on the classical conditioning paradigm first proposed by Pavlov (XXX) and application of those principles to associations with environmental cues (e.g., heroin addicts craving when they see former co-users; see Siegel, CITE)

It is not clear how the theoretical literature will develop going forward. Proponents of multiple self models argue that their models closely correspond to psychological/neural evidence. However, Kable and Glimcher (2004) provide neuroimaging evidence that challenges these claims. There is an ongoing debate over the neural foundations of multiple self models.

Proponents of present-bias argue that parsimony and predictive accuracy support their modeling framework. These models make strong predictions that match available data (Laibson, Repetto, and Tobacman 2010) and they provide an empirically validated theory of misforecasting (Della Vigna and Malmendier 2004, 2006). However, present-bias models violate classical welfare assumptions and introduce the possibility of multiple equilibria.

Finally, proponents of dynamically consistent (unitary self) models tend to prefer these models on principle, because these models do not violate classical assumptions and provide a well-defined welfare criterion. However, they rarely make quantitative predictions, so they are difficult to evaluate empirically.

5 The neural circuitry of social preferences

In this section, we review evidence about the neural processes that govern deviations from purely self-interested behavior (i.e., the neural circuitry of social preferences[7]). The evidence is based on neuroeconomic studies that combine noninvasive neuroscience tools – such as fMRI, TMS and tDCS[8]– with behavioral games used in experimental economics. The neuroeconomic approach

[7] This section draws heavily on (and overlaps with) the work of Fehr and Camerer (2007) and Fehr (2009). Readers who are interested in a more detailed account can find details in those papers.
[8] Transcranial direct current stimulation.

aims to provide a micro-foundation of social preferences in terms of the underlying neural networks, which will eventually be achieved with the development of formal models of the underlying brain circuitry showing how the assumptions and parameters of behavioral models of social preferences relate to the empirically verified assumptions and parameters of the brain model. This will lead to a better understanding of the nature of social preferences, and the sources of individual differences in other-regarding behaviors, including pathologies.

Theories of social preferences are based on the concept of decision utility (D. Kahneman, 1994). Decision utility is a utility function that predictions observed decisions. Decision utility can, in principle, be distinguished from (a) experienced utility, which is the hedonic experience associated with the consumption of a good or an event, (b) anticipated utility, which is the anticipation of experienced utility at the time of decision-making, and from (c) remembered utility, which is the experienced utility consumed when remembering past actions and events.

A central question, which recent studies address, is how the brain constructs decision utilities when a person's behavior reflects his or her own rewards but is also governed by competing social preferences such as warm glow altruism, reciprocity, or inequity aversion. This general question implies a host of other important questions such as: Is self-interest a primary motive that appropriate inhibitory machinery needs to constrain? If so, which brain circuitry is involved in these inhibitory processes? To what extent are these processes related to emotion regulation? Do the positive hedonic consequences associated with non-selfish behaviors partially govern deviations from economic self-interest and, if so, are these complex social rewards represented in the striatum and the OFC like primary or monetary rewards (B. Knutson and J. C. Cooper, 2005, P. O'Doherty J, 2004), or do they rely on different neural circuitry?

Social preferences and reward circuitry

Theories of reciprocity and inequity aversion imply that subjects prefer the mutual cooperation outcome over the unilateral defection outcome in the canonical prisoners' dilemma game, even though unilateral defection leads to a higher economic payoff. Although these theories do not make assumptions about the hedonic processes associated with fairness related behaviors (because they rely on decision utilities), a plausible interpretation of these theories is that subjects in fact derive higher hedonic value from the mutual cooperation outcome (J. W. Thibaut and H.H. Kelley, 1959). Therefore, a natural question is whether we can find neural traces of the special reward value stemming from the mutual cooperation outcome. Two neuroimaging studies (J. K. Rilling et al., 2004, James K Rilling et al., 2002) report activation in the ventral striatum when subjects experience mutual cooperation with a human partner compared to mutual cooperation with a computer partner. Given substantial evidence that primary and secondary reward anticipation activates the striatum, these studies suggest that mutual cooperation with a human partner is especially rewarding (holding financial consequences fixed through the computer partner control).

Social preference theories also predict that subjects prefer punishing unfair behavior such as defection in public good and PD games because leaving an unfair act unpunished is associated with higher disutility than bearing the cost of punishing an unfair act. In this view, it is natural to hypothesize that the act of punishing defection involves higher activation of reward circuitry. A PET study (D. DeQuervain et al., 2004) examined this hypothesis in the context of a social dilemma game with a punishment opportunity. This study showed that the dorsal striatum (caudate nucleus) is strongly activated in the contrast between a real punishment condition (in which the assignment of punishment points hurts the defector in economic terms) and a symbolic punishment condition (where the assignment of punishment points did not reduce the defector's economic payoff). In a another study Singer et al. (2006)documented that men (but not women)

who passively observe that a defector in a PD is punished by a third party show reward related activation in the ventral striatum.

Further evidence that decisions involving social preferences are associated with activity in reward circuitry comes from fMRI studies of charitable donations (William T. Harbaugh, Ulrich Mayr and Daniel R. Burghart, 2007, J. Moll et al., 2006), reactions to offers in a take-it-or-leave-it ultimatum bargaining game (Tabibna Golnaz, Ajay B. Satpute and Matthew D. Lieberman, 2007), and from distribution tasks (E. Tricomi et al., 2010). Ventral tegmental (VTA) and striatal areas are both activated by receiving money and by making non-costly donations, indicating that "giving has its own reward" (J. Moll, et al., 2006). Across subjects, those who made more costly donations also had more activity in the striatal reward circuitry. In one study (William T. Harbaugh, Ulrich Mayr and Daniel R. Burghart, 2007) subjects in a forced-donation condition passively observed money being transferred to themselves or to a charity. In a voluntary condition, subjects could *decide* whether to accept these monetary transfers. Subjects reported higher satisfaction in both the forced and the voluntary condition if the charity received a transfer (controlling for the subject's cost of this transfer). Moreover activations in dorsal and ventral striatum in both conditions are positively correlated with the money that goes to the charity. Thus, all else equal, subjects seem to experience charitable donations as rewarding because the very same reward areas that are activated when the subjects themselves receive a monetary transfer are also activated when the subjects make a costly transfer to a charity.

Neural evidence for inequality aversion was reported by Tricomi et al. (2010). In pairs of subjects, one "rich" subject randomly received a $50 endowment at the beginning of a trial (the other "poor" subject did not, but knew the other subject had received the bonus). Both subjects then rated the outcome of additional transfers to "self" and "other" during fMRI. The rich subjects showed a significantly higher activation in reward related areas (e.g. ventral striatum) for transfers to "other" compared to "self", while the poor subjects showed higher neural reward activation for transfers to "self" compared to "other". The authors' interpretation is that the rich subject is rewarded by a reduction in the gap between his or her earnings and the poor subject's earnings, and the poor subject finds an increase in the wealth gap negatively rewarding. Finally, a recent ultimatum game study (Tabibna Golnaz, Ajay B. Satpute and Matthew D. Lieberman, 2007) provides evidence suggesting that the fairness of a bargaining offer – controlling for the absolute size of the monetary gain – is associated with activation in the ventral striatum. The same dollar bargaining offer of, say $5, elicits higher striatal activation if it represents a fair share (say 50%) of the amount which is being bargained over, compared to when that dollar offer represents a small share (only 15%, for example).

The activations observed in these studies and several others indicate that social rewards commonly activate the dorsal or ventral striatum. There is substantial overlap between these areas of activation and activation observed in studies of reinforcement learning or anticipated money reward (E. Fehr, 2009, E. Fehr and C. F. Camerer, 2007). This overlap is consistent with the hypothesis that social preferences are similar to preferences for one's own rewards in terms of neural activation, which is supportive of theories in which decisions reflect a weighted balance between self-interest and the interests of others.

The studies described above use the simplest multiperson paradigms that allocate money between people or entities. These are important building blocks. Some recent studies consider how the neural circuitry of prosocial behaviors and emotions is affected by various factors.

One topic is "social image": How does knowing another person will observe you affect brain activity and choice? Economists have become interested in this topic (e.g. Bernheim and Andreoni, 2009) and it is important since social image could be affected by many details of how information and institutions are organized. An fMRI study showed that activity in bilateral

striatum was stronger when Japanese subjects were being observed making charitable donations, compared to no observation (Izuma, 2008), which is consistent with the hypothesis that reputation derived from charitable donations is rewarding. A follow-up study showed that American autistic adults exhibit no sensitivity to being observed (compared to matched controls; Izuma et al., in prep).

Consistent with a broad concept of inequity-aversion, one study focused on whether knowing that a high-status person suffers a setback produces a positive reward from "schadenfreude". Activity in response to hypothetical scenarios was found in ventral striatum (and BOLD signal correlated with self-rated responses; Takahashi et al., 2009). This result resembles the finding of Singer et al. (2006) mentioned above.

Social preferences and emotions are also likely to play a role in non-economic domains. One neural study exploring this topic presented vignettes based on actual murder cases with "mitigating circumstances", such as a husband murdering his wife to prevent her further suffering. Judges and juries are typically *required* to consider these circumstances during sentencing, even when the guilt of the murderer is established. Yamada et al. (2011) found that insula activity, a known correlate of simpler kinds of empathy, was associated with the strength of sentence reduction.

Do activations in reward circuitry predict choices?

The evidence above is consistent with the view that costly pro-social acts are rewarding. However, the hedonic interpretation of social preference theories also implies that these acts occur *because* they are rewarding. If it could be shown that higher activations in the striatum *imply* a higher willingness to act altruistically, the case for the reward interpretation would be strengthened considerably.

Neuroimaging data do not allow causal inferences. However, it is possible to move towards causality by predicting choice behavior in one treatment ("out of treatment" forecasting) from neural activity in another treatment. For example, individual differences in caudate nucleus activation when punishment is costless for the punisher predicts how much individuals actually pay for punishment when it is costly (D. DeQuervain, et al., 2004). Likewise, individual differences in striatal activity in the condition where donations are forced predicts subjects' willingness to donate money to charities in the condition in which donations are voluntary (William T. Harbaugh, Ulrich Mayr and Daniel R. Burghart, 2007). These results further support the reward interpretation of social preferences, which in turn provides support for the hypothesis of a common neural currency of social rewards and other primary and secondary rewards (P. R. Montague and G. S. Berns, 2002).

The role of the prefrontal cortex (PFC) in decisions involving social preferences

If people have social preferences, the brain must compare social motives and economic self-interest and resolve conflict between them. Several studies indicate that the prefrontal cortex, a brain region that evolved recently (in evolutionary time) plays a decisive role in this conflict resolution. For example, the ventromedial PFC (BA 10, 11) has been implicated (DeQuervain et al. 2004) in the contrast between a costly punishment condition and costless punishment of players who behaved unfairly; this result is consistent with the hypothesis that this area is involved in the integration of separate benefits and costs in the pursuit of behavioral goals (Narender Ramnani and Adrian M. Owen, 2004). In charitable donations (J. Moll, et al., 2006), the contrast between altruistic decisions involving costs and no costs also showed activation of the VMPFC (BA 10, 11, 32) and the dorsal anterior cingulate cortex (ACC). Since the ACC is thought to play a key role in conflict monitoring (M. M. Botvinick et al., 2001), activity in this region is consistent with the existence of a trade-off between self-interest and pro-social motives.

The role of the VMPFC in decisions involving costly altruism is also interesting because of related activation in this region in other studies. The VMPFC is involved in emotional processing and moral judgment (M. Koenigs et al., 2007, J. Moll et al., 2005), in integrating the value of consumer products and their prices (B. Knutson et al., 2007), in the encoding of the willingness to pay for consumer goods, lotteries (V. S. Chib et al., 2009, A. Rangel and T. Hare, 2009), and charitable donations (T. A. Hare et al., 2010). Lesions to VMPFC are also associated with poor choices in various situations (A. Bechara et al., 1997, A. R. Damasio, 1995) which require integrating costs and benefits, and in reduced prosociality (Krajbich et al., 2009). The Hare et al. (2010) study shows that activity in VMPFC is positively correlated with charitable donations consistent with the view that emerged from many other studies (V. S. Chib, et al., 2009, A. Rangel and T. Hare, 2009), that this area of the brain encodes decision utility. In addition, the value signal in the VMPFC is modulated by other signals in the posterior superior temporal cortex which have been shown to be important for overcoming egocentricity bias (pSTC), indicating that VMPFC and pSTC activity are key components of the neural circuitry of social preferences. This does of course not mean that these areas are exclusively dedicated to the processing of social preferences. Rather, in the case of the VMPFC, for example, the studies suggest a general role for this region in integrating emotional feelings about costs and benefits, regardless of whether these choices involve economic consumption goods or "non-economic" goods such as the subjective value of acting altruistically.

The dorsolateral prefrontal cortex (DLPFC) probably also plays an important role in the processing of decisions involving social preferences (Alan G. Sanfey et al., 2003). This study examined the neural circuitry involved in the recipient's behavior in an ultimatum game where the rejection of low positive offers involves a motivational conflict between fairness and economic self-interest. It reports activation of bilateral DLPFC and bilateral anterior insula (AI) in the contrast between "unfair>fair" offers. In addition, the higher the activation of right AI, the more likely a subject is to reject an unfair offer, suggesting that AI activation may be related to the degree of emotional resentment of unfair offers. The DLPFC activation may represent the cognitive control of the emotional impulse to reject unfair offers.

The interpretation that DLPFC activity represents the cognitive control of the impulse to reject implies that interfering or disrupting DLPFC activity reduces the control of the impulse and should, thus, increase the rejection rate. Knoch et al. (D. Knoch et al., 2006) tested this hypothesis by reducing the activation in right and left DLPFC with low-frequency transcranial magnetic stimulation (TMS). Surprisingly, the study found that TMS of right DLPFC increases the *acceptance* rate of unfair offers relative to a placebo stimulation (from 9% to 44%), while TMS of left DLPFC did not affect behavior significantly (relative to a placebo condition). This finding suggests that right DLPFC is *causally* involved in controlling the impulse that pushes subjects

towards accepting unfair offers, i.e., in controlling or weighing economic self-interest. Interestingly, the disruption of right DLPFC only affects subjects' fairness related behaviors but not their fairness judgments, i.e., they still judge low offers to be very unfair, but they nevertheless accept them more frequently and more quickly. A similar dissociation between fairness judgments and fair responder behavior has been observed in Knoch et al (D. Knoch et al., 2007) where the authors down regulate the activity of the right DLPFC with TDCS. Another TMS study (D. Knoch et al., 2009) shows that the right DLPFC is also causally involved in the formation of individual reputations as a trustworthy agent in a repeated trust game, since disruption leads to more untrustworthy behavior which harms reputation. Apparently, when subjects face a trade-off between the short run benefit of cheating their current partner and the long-run benefit of having a good reputation when facing future partners in the trust game, a functioning DLPFC seems to be necessary to enable subjects to decide in favor of their long-run benefit. This role of the DLPFC in overcoming short-run self-interest has also been corroborated in Spitzer et al (2007); this study shows that stronger compliance with a social norm in the face of a possible sanctioning threat is strongly correlated with the strength of DLPFC activity.

In a recent study, Baumgartner et al. (2011) applied TMS *and* fMRI to responders in the ultimatum; they were either stimulated with TMS to the right or the left DLPFC and one control group was not stimulated at all. Subsequently, they played the ultimatum game during fMRI. This combination of methods enables the examination of the causal impact of TMS on behavior and the identification of the neural circuitry that is causally involved in the behavioral change. Interestingly, subjects who received TMS to the left DLPFC or no TMS (i.e., the "normal" subjects) show a much higher rejection rate of unfair offers than subjects who received TMS to the right DLPFC (i.e., the "deviant" subjects). In addition, the normal subjects display significantly higher activity in, and connectivity between, the right DLPFC and the VMPFC when they receive unfairly low offers. These findings are consistent with the view that the activation of right DLPFC and VMPFC, and the connectivity between them, is causally involved in regulating the decision utility of rejecting unfair offers.

One emerging theme of the studies reviewed in this section is that social reward activates circuitry that overlaps circuitry which anticipates and represents other types of rewards to a surprising degree. These studies reinforce the idea that social preferences for donating money, rejecting unfair offers, and punishing those who violate norms, are genuine expressions of preference. The social rewards are traded off with subjects' economic self-interest; the dorsolateral and the ventromedial prefrontal cortex are likely to be crucially involved in the balancing of competing rewards and the computation of decision utilities. Non-invasive brain stimulation can alter these neural processes and subjects' behaviorally expressed social preferences. This establishes the causal relevance of the identified neural computations for subjects' behavior.

However, brain stimulation is not the only way of establishing the causal relevance of fMRI-identified neural circuitry for subjects' behavior. In recent years, several papers indicate that the great potential of pharmacological experiments. Testosteron has been shown to enhance the fairness of bargaining offers in the ultimatum game (C. Eisenegger et al., 2010); the neurohormone oxytocin increases trusting behavior but not trustworthiness (M. Kosfeld et al., 2005); the depletion of the neurotransmitter serotonin increases the rejection rate in the ultimatum game (M. J. Crockett, 2009, M. J. Crockett et al., 2008) and benzodiazepine reduces the rejection rate (K. Gospic et al., 2011). In several cases the pharmacological intervention was combined with fMRI so that the researchers were able to identify the neural circuitry causally involved in the behavioral change (T. Baumgartner et al., 2008, K. Gospic, et al., 2011). While space limits

prevent us from going into the details these studies further confirm the rapid progress that has been made in recent years in this field.

6. Strategic thinking

Game theory started as applied mathematics describing "solutions" to games based on idealized play. Over several decades, game theory grew to include experimental studies, more psychologically realistic models (e.g., Camerer 2003), evolutionary modeling, and design applications. Neuroscience could contribute to game theory by identifying strategic algorithms that are being implemented in the brain. In addition, game theory could be of special use in neuroeconomics by parsing how general reward and learning structures combine with specialized social inference mechanisms (such as "theory of mind") to determine strategic choice.

This section is organized around the bold idea that the neural basis of strategic thinking is likely may have separable components corresponding to the mathematical restrictions imposed in different kinds of game theory. This simplification will surely turn out to be wrong on many details. However, it is certainly likely that different components of strategic thinking and execution require different cognitive capacities that are primarily located in different brain regions (and are differentially developed across species). If these different kinds of cognitive capacities have special value in certain types of games, then there will be some association between brain regions and strategic choices.

For example, a recent study (Martin et al., 2011) showed that chimpanzees make choices in two-strategy matching pennies games which are both closer to (mixed) Nash equilibrium than comparable human choices, and about as statistically independent of past observations as human choices are. The chimps behave more game-theoretically than humans! However, in these games, the main cognitive skill is detecting patterns in choices by others and disguising one's own patterns from others. Chimps are actually better at short term detection and spatial pattern memory than people. This example illustrates how a highly specialized cognitive skill could account for differences in behavior (between species) in a narrow class of games.

We discuss four aspects of strategic thinking and what is known about neural activity during those types of thinking:

1. *Strategic awareness* that outcomes are affected by actions of other players
2. *Beliefs and iterated beliefs* about what other players will do and think;
3. *Learning* about the value of strategies, perhaps by reinforcement or counterfactual "fictive" (model-based) learning
4. *Strategic teaching*, the valuation and adjustment of behavior by anticipating the effects of one's current action on another player's beliefs and future behavior.

The additional topic of social preference (how outcomes other players receive are valued) is discussed in another section of this chapter.

Strategic awareness

The most basic idea in game theory is that people are strategically aware that their outcomes depend on choices by other players. While these seems obviously true for educated adults, strategic awareness may well be absent for human children, other species, in causally

complex environments, and in disorders associated with deficits in social reasoning (such as autism).

Neural evidence: Several studies have shown differential neural activation when playing a game against a human opponent, compared to a computer (e.g., Gallagher et al. 2002; McCabe et al. 2001; Coricelli & Nagel 2009). These papers are methodologically challenging, because it is crucial to control for comparability of the behavior of humans and computers (and particularly its expected reward value) in the presence of feedback. Nonetheless, the studies are highly suggestive that agents used specialized neural processes when playing other humans.

Beliefs, iterated beliefs and strategic choice

If players have some strategic awareness, then what strategic choices do players make if they know they are playing other players? Based on subjective utility theory, a natural theory is that players form beliefs about what other players will do and their strategic choices reveal those beliefs.

The most elegant and prominent assumption in game theory is that beliefs are in (Nash) equilibrium, which is equivalent to mutually rational players having mutual knowledge of one another's strategies. That is, in equilibrium players have somehow correctly figured out what others will do and optimize given their beliefs. However, equilibration is unlikely to come from preplay analysis of a game, and instead is likely to come from experience (as in learning models), evolutionary adaptation, or preplay communication.

It is highly unlikely that the brain would directly compute an equilibrium strategy, let's turn attention to a family of theories which is more neurally plausible—cognitive hierarchy (CH) or level-k theories.

These theories assume that players form beliefs by iterating through steps of thinking (probably 1-3 steps). The iteration starts with a level-0 player who chooses according to a simple heuristic (e.g., randomly, or using perceptual salience). Agents doing one or more steps of thinking compute what lower-level thinkers will do and best-respond or imperfectly "better-respond" using a softmax response.[9]

The **behavioral evidence** in support of these CH theories is that predictions about aggregate choices are typically better approximations of actual human play than equilibrium theories. Importantly, they appear to explain *both* deviations from equilibrium predictions in one-shot play, and also explain when equilibrium predictions are surprisingly accurate (even with no learning; see Camerer, Ho & Chong 2004; Crawford, Costa-Gomes & Iriberri 2010).

Direct cognitive evidence for steps of thinking comes from eyetracking and mouse-based studies. These studies record what information subjects are looking at, and for how long. Then the theory can be tested as a joint hypothesis about information search and choices resulting from that search. For example, level-2 players *must* look at other players' payoffs to choose strategies, but lower level players do not. So the theories predict an association between looking at the payoffs of other players and frequency of higher-level choices, The earliest studies, going back at least two decades, showed approximate conformity of thinking steps to associated predictions of information search by different types (e.g., Camerer et al. 1993; Johnson et al. 2002). More recent studies showed even clearer conformity of imperfect information lookup and choice (Costa-Gomes, Crawford & Broseta 2001; Costa-Gomes & Crawford 2006; Wang, Spezio & Camerer 2010; Brocas et al. 2009).

There is also modest to high intrapersonal reliability across games of an individual's classified level type (although probably lower than levels of reliability for the most stable traits, such as IQ and extraversion). For example, Chong, Ho and Camerer (2005) computed a

[9] See Camerer et al. 1993; Nagel 1995; Stahl and Wilson 1995; Costa-Gomes, Crawford & Broseta 2001; Camerer, Ho & Chong 2004; Crawford et al., 2010.

correlation of +.61 between a subject's average estimated levels in two separate groups of 11 games. There are also modest correlations between estimated thinking levels and both working memory (Devetag & Warglien 2003) and "theory of mind" emotion detection skill (Georganas, Healy & Weber 2010).

Neural evidence: A small number of neuroimaging fMRI studies have explored the neural underpinnings of strategic belief formation and depth of thinking.

Bhatt and Camerer (2005) considered the processes of choice and first and second order belief formation in two-player, dominance-solvable matrix games with 2-4 strategies. In each trial subjects either made a choice in the game, guessed what the other player would do (i.e., stated first order beliefs) or guessed the other player's first order beliefs about their own choice (i.e., stated second order beliefs). In order to isolate the process of reasoning without 'interference' from learning, there was no feedback.

A simple hypothesis consistent with CH modeling is that many subjects will use different reasoning processes in choosing and forming beliefs. For example, level 0 and 1 players may spend no time forming a belief; this could be manifested as substantially greater activity in value-oriented regions during choice than in guessing. Indeed, when subjects' choices and beliefs were *out of equilibrium*[10], the choice task elicited significantly more activity in medial prefrontal cortex (mPFC) and DLPFC (involved in working memory and self-control). However, when subjects' choices and beliefs were *in equilibrium*, activation patterns were not significantly different in choice and guessing trials in a small area of the ventral striatum (probably associated with differential rewards in the two types of trials).

Bhatt and Camerer also defined a measure of "strategic intelligence" (SIQ) based on each player's expected payoffs and belief accuracy. High SIQ subjects had significantly greater activation in the caudate (a reward-related area) and precuneus. Conversely, people with lower SIQ had significantly more activation in the left insular cortex, an area strongly associated with emotional discomfort, and financial risk and uncertainty (e.g., Mohr, Biele & Heekeren 2010). Thus, poor strategic performance seems to reflect high internal strategic uncertainty, as 'felt' by in the insula.

Kuo et al. (2009) did fMRI during play of asymmetric dominance-solvable games and matching games. Games varied in difficulty (corresponding to the number of steps of iterated reasoning necessary to reach Nash equilibrium). Activation in the precuneus scaled with the difficulty of these games. They also studied simple matching games that had the same formats as the dominance solvable games, but in which reward was maximized if you chose the same target as a partner. They found that the middle insula correlated with a measure of how "focal" a game was (and also with expected payoff), as if focality is associated with a bodily "gut feeling" projected to insula.

Coricelli and Nagel (2009) focused on the "p-beauty contest", in which subjects choose numbers in the interval [0,100] and win if their number is the closest to a multiplier p times the average number. Their subjects played a series of games with different values of the multiplier p (and no feedback) against both humans and computers (which chose randomly from all numbers).

They were able to classify people by behavior rather sharply into level-1 thinkers, who choose close to p*50 in most games, and level-2 thinkers who choose $p^2$50. They found significantly more activation in dmPFC (paracingulate) and vmPFC and bilateral tempero-parietal junction (TPJ) (see Figure 1). These are areas that are rather well-established to be part of a candidate "theory of mind" circuit used to compute the intentions, beliefs and desires of others (e.g., Amodio & Frith, 2006). They also find a positive correlation (r = .84) across subjects between activity in dmPFC and how close a subject was to winning.

[10] Being "in equilibrium" is defined behaviorally, as trials in which choices are best responses to beliefs, and both beliefs and second-order beliefs match choices and beliefs of other players.

Yoshida et al. (2008) create a recursive-belief model similar to the cognitive hierarchy approaches and apply it to the game of stag hunt. In their games, two low-value rabbits are present on a two-dimensional grid. A high-value stag is also present. Two players make sequential one-step moves either toward the stag (who also moves) or toward a rabbit. The game ends when either of the players reaches a rabbit target or when the two players end up adjacent to the stag, 'capturing' it.

They formalize a Bayesian notion of steps of recursive anticipation. The model creates trial-by-trial computational regressors. Using fMRI, they find that entropy about opponent thinking steps (strategic uncertainty) activates medial prefrontal cortex (paracingulate) and posterior cingulate. The level of strategy the subject seems to use is correlated with DLPFC as well as frontal eye field and superior parietal lobule. They suggest that paracingulate is activated in mentalizing to determine opponent's strategic thinking type, and DLPFC is involved in implementing planning ahead and working memory during 'deep' strategic thinking' (planning ahead several moves, as in chess, especially given their visual display of the game on a grid).

Learning

Many empirical studies have examined how human (and monkey) agents learn to adjust their strategies in games (see Camerer 2003, ch. 6). While there is a huge literature on the neuroscience of animal and human learning in simple decisions, there is only a small intersection combining estimation of empirical models of human learning and neural observation.

Two popular theories are reinforcement, and belief learning (e.g., fictitious play). In reinforcement learning, strategy values are adjusted by payoffs (or prediction error). In belief learning, beliefs about what others will do are adjusted by observation and then used to compute expected payoffs and guide choice. One popular form of belief learning is weighted fictitious play (WFP), in which beliefs are a weighted average of observed past choices by opponents. Camerer and Ho (1999) noted that learning according to WFP is exactly the same as a general type of reinforcement learning in which strategies that are not chosen are also reinforced according to a foregone payoff, sometimes called "fictive learning". (This kind of learning is sometimes called "model-based" because it requires a model, or understanding of how all possible choices lead to possible payoffs, to compute fictive payoffs.)

From a neural point of view, the observation that WFP is a kind of reinforcement invites consideration of a general model in which strategy values combine both reinforced payoffs and foregone payoffs. In a useful class of models, the fictive weight is δ times the reinforcement weight of one, perhaps because they those value signals are computed differently in the brain and weighted differently in guiding behavior. Empirical estimates from behavior in many games suggest that the fictive learning weight δ is between 0 and 1. These data suggest subjects do use "model-based" information about foregone payoffs, but may not weigh that information as heavily as received rewards.

A plausible hypothesis about locations of neural activity is that reinforced value computations are encoded by prediction error in the midbrain and ventral striatum (as shown by many studies). These are phylogenetically older regions shared by humans and many other species, an anatomical observation that is consistent with the vast array of evidence that reinforcement learning processes are common across species. Some studies indicate that regret signals are encoded in orbitofrontal cortex (Coricelli, Dolan, Sirigu 2007); since fictive learning is typically based on imagined counterfactuals, like those which create regret, it is plausible that these signals would be encoded in OFC and connected areas.

Neural evidence: Available neuroscience studies reject the simple bases case in which there is no fictive learning (i.e., $\delta = 0$) and fictive learning as strong as learning from received rewards ($\delta=1$). Lohrenz et al. show fictive learning signals in VStr similar to prediction error

signals from actual rewards (Lohrenz et al. 2007). Mobbs et al. (2009) show activation in response to rewards earned by similar others, which suggests a more general model in which learning can be both fictive and based on learning from observing others (perhaps depending on "social distance").

Hayden, Pearson, and Platt (2009) also record fictive learning signals from dorsal ACC neurons in rhesus monkeys. They show that the monkeys do respond to fictive rewards (if a high-value target was in a location they didn't choose, they are more likely to choose it next time). The ratio of neural firing rates in response to fictive versus experienced reward is around .70, which suggests a crude estimate of an EWA relative weighting δ parameter.

Fictive learning is a special kind of "model-based" learning in computational neuroscience. In model-based learning, agents use the knowledge of how the values of multiple choice objects are linked—through a "model"— to update assigned values of all objects after receiving a learning signal from one chosen object. Hampton et al. (2008) show clear learning signals corresponding to model-based learning.

Thevarajah et al. (2010) looked for neural correlates of EWA learning in a matching pennies game. In their experiment two rhesus macaques made choices, through eye saccades, against a computerized opponent designed to exploit temporal patterns in the macaques' play. Single-unit electrode recording measured neural firing in intermediate superior colliculus (SCi). SCi is a region that topographically maps saccade sites, and also projects to premotor neurons and also to dopaminergic sites in the midbrain (ventral tegmental area and substantia nigra) so it is a sensible a priori candidate for encoding the value of a saccade (i.e. a strategy choice, given how the game is played). They find a strong correlation between SCi firing rates and EWA strategy values in one monkey, and a modest correlation in the other monkey.

Strategic teaching and influence value

The learning theories described in the last section above are all adaptive; that is, they adjust either estimated strategy value or adjust beliefs in response to previous experience. A further step is "sophistication"—that is, players form beliefs using a model of how other players are learning. There is some evidence that models with sophistication (and learning to be more sophisticated) fit information lookup and choice data better than simple adaptive models (e.g., Stahl, 2000; Camerer, Ho & Chong, 2004).[11]

Sophistication should interact with the nature of repeated matching. When players expect to play together repeatedly, if one player is sophisticated it can pay for her to take actions that deliberately manipulate the learning process of the other player. A common example of this sort of "strategic teaching" is bluffing in poker: Bluffing is betting aggressively to make opponents believe you have a winning hand, so they should quit betting and fold their cards. It is well known that an incentive to "strategically teach" can arise in repeated games, and also in games where a long-run player is matched with a sequence of short-run players (Fudenberg and Levine, 1998).

Hampton, Bossaerts and O'Doherty (2007) did fMRI to study strategic teaching in a two-player "work-shirk" game (a version of asymmetric matching pennies). In early work, Platt and Glimcher recorded neural firing in lateral intraparietal cortext (LIP) and found it associated closely with expected payoffs in this game, for monkeys playing computerized opponents. Simple

[11] Notice that while these theories can be difficult to distinguish using only observed choices, it is easy to distinguish them with cognitive data: Adaptive players *do not* need to look at the payoffs other players get, but sophisticated players *do* need to look at those payoffs. The fact that players usually do attend to payoffs of others players (e.g., Knoepfle, Wang and Camerer, 2009) is evidence for sophistication.

reinforcement learning fits these neural signals well in monkeys (e.g., Seo, Barraclough & Lee, 2009).

The authors fit three models: Reinforcement learning; fictitious play; an "influence model" where players account for the impact of current actions on their own value in the future through its influence on the opponent's reinforcement learning. For example, an employee who chooses W when the employer picked D earns 0. However, if a learning employer is then likely to pick D again in the future, the W choice has an "influence value" because it raises the value of shirking (S).

Hampton et al. found that for about half the subjects choices were better fit by including an influence value term (half were not). They analyzed two areas generally though to be part of the mentalizing circuit, the superior temporal sulcus (STS) and dorsomedial prefrontal cortex (mPFC). They found that these areas correlated with different aspects of the influence model. mPFC activity correlated to predicted reward in the influence model at the time of choice, while the STS correlated to the component of prediction error related to second-order belief, specifically this area correlated with the amount that the model predicted the opponent should adapt his behavior based on your action (when feedback is seen). Notice that both this error signal, and predicted reward, are largest when surprise is involved. Predicted reward in the influence model is largest when the subject switches strategies, i.e. when the subject surprises his opponent. Similarly, the influence update signal is largest when a player's own action is in opposition to his second-order belief (i.e., the player plans to choose a strategy different than what they think the other player expects).

Direct strategic deception is shown by Bhatt, Lohrenz, Camerer and Montague (2010) in bargaining. Two players, a buyer and a seller, play 60 rounds of the game. At the beginning of each round the "buyer" is informed of her private value V, which is an integer drawn with uniform probability between 1 and 10 (Figure 2). She is then asked to "suggest a price" S to the seller, an integer between 1 and 10. The seller sees this suggestion and sets a price P. If $P \leq V$, the trade executes and the seller and buyer earn P and V-P. If $P > V$, the trade does not execute and they get nothing. Importantly, *no feedback* about whether the trade occurred is provided to either player after each round.

By regressing each buyer's suggestions s against their values V, Bhatt et al. could classify buyers into three types. One type showed no strong correlation. A second "incrementalist" type typically had a strong positive correlation (and high R^2), due to deliberate revelation of values (in an effort to increase efficiency). A third "strategist" type used a counterintuitive strategy of sending high S suggestions when they have low values V, and sending low suggestions when they have high values (so S and V are negatively correlated). (This behavior is predicted as level-2 in a modified CH model.) The idea is that naïve level-1 sellers will attempt to make inferences about how "honest" a buyer is by considering the history of suggestions they see in the game. If those sellers see only low values of S they will infer that the buyer is low-balling and will ignore the suggestions[12]. However, if they see a relatively uniform mixture of suggestions, they will think the buyer must be prosocially revealing something about their values to increase gains from trade. They will tend to trust the suggestions, choosing low prices when they see low suggestions and high prices when they see high suggestions. Level-2 strategist buyers will realize this and use low-value rounds, where they don't stand to earn much anyway, to generate credibility so that they can reap all the rewards from very low prices during the high-value rounds.

Bhatt et al. found that during the buyer's price suggestion period, there is stronger activity in the DLPFC for strategists compared to other subjects. This could be interpreted as evidence of active working memory (keeping track of the distribution of recent suggestions in order to make it look honest) or inhibition of a natural instinct to make suggestions which are

[12] Note that the unique Nash equilibrium is for no information to be translated (called "babbling" in game theory jargon).

positively correlated with value. There is also unusually large activity for strategists when they receive a high-value signal (and hence must bluff the most by suggesting a low price) in STS close to the region observed in Hampton et al (2007).

For sellers who are judging how much information is conveyed by a buyer's price suggestion, Bhatt et al. (2011) found that activity in bilateral amygdala was correlated with a seller's "suspicion", as measured by how closely the sellers' price offers matched the buyers' suggestions. A low correlation indicates suspicion and is associated with amygdala activity, consistent with an established role of amygdala in rapid vigilance toward threat (e.g., fear response).

Together, these studies show that there is some match between computations inferred from choices (influence value and "strategizing") and regions thought to be involved in value calculation and mentalizing, and in emotional judgments associated with economic suspicion.

Montague and several colleagues have explored many aspects of a 10-period repeated trust game using fMRI. King-Casas et al. (2005) found signals in the caudate nucleus of the trustee brain in response to positive ("benevolent") reciprocity by the investor. This suggests the brain is computing a rather complex kind of social reward based on an anticipation of future responses. In addition, there is evidence that activity in the caudate region occurs earlier and earlier across later rounds of the experiment, by about 14 seconds, signaling a behavioral "intention to trust" well ahead of the actual behavior.

More recently, Montague's group has used trust games as a tool for doing "computational psychiatry"— that is, exploring how disorders are associated with disruption conventional neural computations that are typically adaptive.

King-Casas et al. (2008) consider behavior and neural activity during the trust game in subjects with borderline personality disorder. Borderline personality disorder (BPD) is characterized by emotional disregulation, including some level of paranoia, often leading to unstable personal relationships. In the King-Casas experiment, subjects with BPD were paired as trustees with healthy investors matched on education, IQ, and socioeconomic status, and played 10 rounds of the trust game.

The major behavioral finding is that pairs that included a BPD subjects earned significantly less money in total than those involving two healthy subjects. This appears to be due to markedly lower levels of investment in the later rounds of the game by investors when playing with a BPD trustee. In healthy pairs, breakdowns of cooperation were often followed by "coaxing" behavior by the trustees: trustees would repay all or most of the money they receive during the trial. This signaled trustworthiness to the investor and often restored a cooperative interaction. Investments appeared to decrease in these pairs because BPD subjects failed to effectively signal their trustworthiness to the investors via this coaxing behavior.

The study found that people with BPD had significantly decreased activation in the anterior insula (aIns) in response to low investments as compared to controls. Activity in aINS has often been linked to subjects experiencing emotional discomfort, perhaps accompanying a violation of social norms (e.g., low offers in the ultimatum game; Sanfey et al., 2003). A lack of activity here when BPD subjects see low investment suggests a failure to interpret those low investments as a lack of trust in response to trustee norm violations. The authors hypothesize that this failure to detect a violation of social norms impairs the ability of the BPDs to respond appropriately with coaxing. In turn this failure to coax leads to decreased cooperation throughout the experiment and fewer returns to both parties.

Chiu et al. (2008) find that autistic subjects had much weaker signals in regions of cingulate specialized to "self" signals about payoffs and actions of oneself.

Discussion of Strategic Neuroscience

As noted in the introduction, the goal of neuroeconomics is *not* to find a special brain area for each task. Quite the opposite: The hope is that common patterns of circuitry will emerge which will inform debates about the computations that are performed, and suggest new theories of behavior and new predictions. Strategic neuroscience is just beginning, but there is some tentative convergence about activity in four regions across studies: mPFC, DLPFC, the precuneus, and the insula. The locations of activity described in this section are identified in three brain "slices" and shown in Figures 2A-C.

mPFC: Activation in dorsal mPFC was found when choices were out of equilibrium (Bhatt & Camerer 2005), among higher-level thinkers (Coricelli & Nagel 2009), when the other player's sophistication is uncertain (Yoshida et al. 2008), and when computing influence value (Hampton et al. 2008) This region is active in many social cognition tasks including self-knowledge and perspective taking (Amodio & Frith 2006; D'Argembeau et al. 2007) and in some non-social tasks which require cognitive control (Ridderinkhof et al. 2004; Li et al. 2006). Amodio and Frith hypothesize that the region is involved with modulating behavior based on anticipated value, with the most posterior areas dealing with simple action values, and representations getting increasingly abstract and complex moving forward toward the frontal pole..

There is very tentative evidence consistent with this hypothesized posterior-anterior value complexity gradient, as measured by the y-coordinate in x-y-z space[13]: The simplest behavior is probably in Bhatt and Camerer (y=36), two-step thinking is a little more complex (Coricelli & Nagel, 2009, y=48) and influence value is rather complex (Hampton et al. 2008, y=63).

DLPFC: The dorsolateral PFC is thought to be involved in working memory (which is necessary for doing "I think he thinks…" types of calculations) and also in inhibition of rapid prepotent responses (such as implementing patient plans, e.g., McClure et al 2004, 2007; resisting tempting foods; Hare, Rangel & Camerer 2009). In the studies in this section, it is seen in Bhatt and Camerer (strategic choice out of equilibrium), Coricelli and Nagel (correlated with higher-level thinking), Yoshida et al. (higher-level thinking), and Bhatt et al. (strategizing price suggestions in bargaining). These results suggest DLPFC may be necessary for a combination of working memory and execuive control required to play strategically at high levels. Importantly, Knoch et al. (2009) found that application of disruptive TMS to right DLPFC reduced the tendency of players to build up reputations in partner-matching repeated trust games (with no such change in anonymous stranger-matching games).

Precuneus: Precuneus activity is seen in Bhatt and Camerer (2005), Kuo et al. (2009), and Bhatt et al. (2010). The precuneus has reciprocal connections with many of the other areas mentioned throughout this chapter including the mPFC, the cingulate including both the ACC and retrosplenial cortices, and the dorsolateral prefrontal cortex.

The precuneus has been implicated in a host of tasks including episodic memory retrieval (Shallice et al. 1994, Fletcher et al. 1995, Lundstrom et al. 2003, Addis et al. 2004), attention guidance and switching (both between objects, and among object features) (Culham et al. 1998; Le, Pardo & Hu 1998; Nagahama et al. 1999; Simon et al. 2002), a variety of imagery tasks (Cavanna & Trimble 2006), and perspective taking (Vogeley et al. 2004; Vogeley et al. 2001; Ruby & Decety 2001). Precuneus is also one of the "default network" areas that are unusually active when subjects are conscious and resting (Raichle et al. 2001).

Our hunch is that it is unlikely that the precuneus plays a special role in strategic thinking. Instead, the activity observed in a few studies is likely to be due to the fact that attentional control and perspective taking are important for complex strategic valuation. A fruitful

[13] A higher positive value of y is further forward, or more anterior, in the brain; more negative values are more posterior toward the back of the brain. Similarly, x values range from the left side (most negative) to the right side (most positive), and z-values range from most negative (the inferior part or bottom of the brain) to the most positive (the superior part or top of the brain).

way to learn more would be to vary a single dimension of games, such as symmetry versus asymmetry, which are designed to require more perspective taking and attentional control, and see if precuneus is actually more active.

Insula: Insula activity appears in Bhatt and Camerer (correlated with low strategic payoff and accuracy) and Kuo et al. (2009) (correlated with focality in matching games). Both studies show activity in the middle insula, around y=0). The insula is thought to be responsible for "interoception", that is, the perception of one's own internal state. It has been proposed that the information received in the posterior insula is processed and re-represented in the anterior insula as subjective emotion, and is also important for a feeling of self (Craig 2002; Critchley 2005; Keysers & Gazzola 2007). It may be that middle insula activity reflects more basic visceral sensations in these games—like intuitive impulses corresponding to generalized strategic uncertainty rather than to more analytical processing.

Summary

Game theory has emerged as a standard language in economics and is the focus of thousands of behavioral experiments. However, only a small number of these studies are focused on measuring non-choice aspects of algorithms that are used to choose strategies. So far, a small number of fMRI studies and several studies using variants of eyetracking are reasonably supportive of cognitive hierarchy-type models, as models of both mental computation and resulting choices.

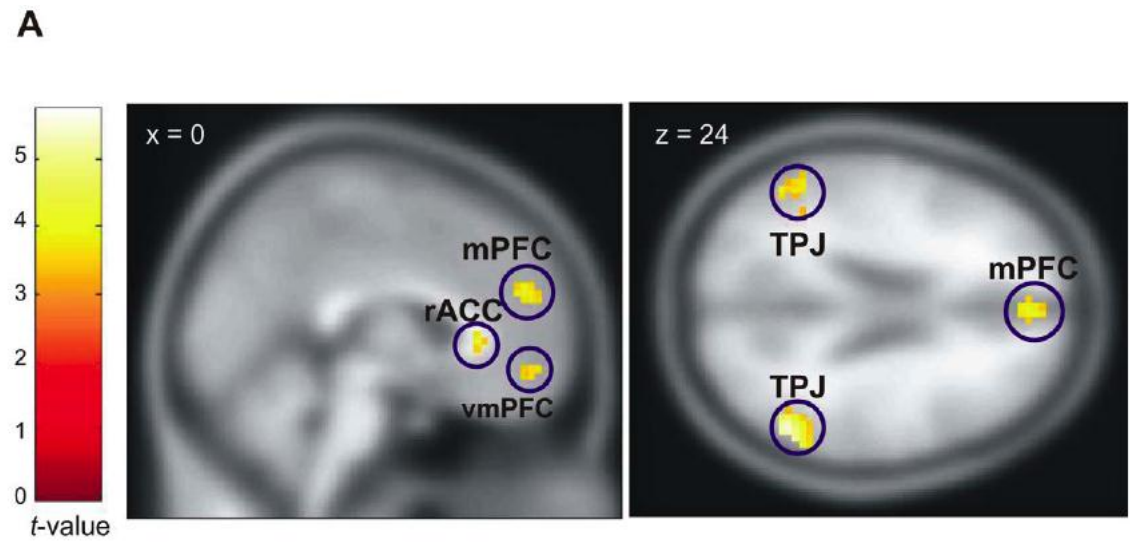
However, given that there is a huge space of possible theories covering strategic thinking, learning and teaching, it may be difficult to rapidly figure out which theories predict best, under what circumstances, without testing both the choice predictions of theories as well as cognitive and biological predictions. Strategic neuroscience could be very useful for making progress. In addition, since many of the candidate regions identified so far in fMRI or close to the scalp (such as TPJ, dmPFC), other tools such as EEG and TMS which record or disrupt electrical activity close to the cortical surface could prove particularly useful in checking robustness of results from fMRI and lesion studies.

Finally, it is useful to ask again—Why care about where? That is, suppose we believed (with more confidence that we have now) that the common areas shown in Figures 2A-C are computing aspects of strategic value or action. What can be done with that information? The answer is that we can couple knowledge of function in these regions with emerging knowledge of how these regions work in different species, develop across the human life cycle (both childhood tissue growth and decline in aging decline), are connected to other regions, and are affected by gene expression, neurotransmitters, and drugs. Combining functional and anatomical knowledge will lead to predictions about the types of animals and people who may behave more or less strategically (as in Figure 1). Predictions can also be made about how activity will be modulated by changes in representations, or simply environmental effects, which either overload or activate these regions.

FIGURE CAPTIONS

Figure 1: Differences in brain activity in response to playing a human versus computer which are, respectively, larger for level-2 players (mPFC, vmPFC, TPJ) and larger for level-1 players (rACC). (Source: Coricelli and Nagel PNAS 2009 [permission TBA]

Figure 2: Regions of activity in various game theoretic and mentalizing tasks. (A) Sagittal slice from back (posterior) to front (anterior) of the brain, x=5. Shows activity in precuneus/posterior cingulate (posterior) and dorsomedial prefrontal cortex (DMPFC) (anterior). (B) Sagittal slice, x=35 shows activity in right insula. (Left insula regions are inverted to opposite right regions for purposes of plotting.) (C) Coronal slice from left to right, y=24. Shows activity in dorsolateral prefrontal cortex (DLPFC). [NOTE: THESE FIGURES ARE IN A SEPARATE FILE HandbookExpEc2CamererSlices.pptx]


A
5
4
3
2
1
0
t-value
x = 0
mPFC
rACC
vmPFC
z = 24
TPJ
mPFC
TPJ

Conclusion

The development of neuroimaging techniques, particularly fMRI, has opened up unprecedented opportunities for research on the human brain. It is now possible to index neural activity associated with human mental faculties, such as language processing, moral reasoning and economic decision making (e.g., Cohen, 2005). However, despite the tremendous opportunity presented by this method, it is fraught with significant limitations that must be understood. These include the noisy nature of the signals being measured, which are only proxies of neural function, as well as the manner in which they are analyzed. Failure to consider these limitations can lead to misleading or erroneous conclusions, and such missteps are certainly present in the rapidly growing literature on fMRI studies. At the same time, there is little doubt that the responsible use of this method has begun, and will continue to fuel a remarkable new period of discovery about the neural mechanisms underlying perception, cognition, and behavior.