Journal of Economic Perspectives—Volume 19, Number 4—Fall 2005—Pages 3–24

The Vulcanization of the Human Brain: A Neural Perspective on Interactions Between Cognition and Emotion

Jonathan D. Cohen

Emotions influence our decisions. They do so in just about every walk of our lives, whether we are aware or unaware of it and whether we acknowledge it or not. In particular, I will argue that emotions may explain inconsistencies in human behavior and forms of behavior that some have deemed irrational, though such behavior may seem more sensible after a discussion of the functions that emotions serve—or may have once served in our evolutionary past.

People do have the capacity to override emotional responses. This capacity relies in large measure on the most recently evolved parts of our brains that support forms of behavior that are more recognizably rational. Neuroscientists are beginning to make headway in identifying the neural mechanisms involved in both emotional responses and higher cognitive processes. This effort offers the promise of a deeper understanding of how and why emotions impact decision making, how this may contribute to behavior that appears to deviate from optimality, and how and when we are able to overcome such emotional responses.

In this article I will review a series of human brain imaging studies of decision-making behavior. I will argue that the human brain is best understood as a confederation of mechanisms that usually act together, but at times may come into competition with one another, favoring different evaluations of similar circumstances. Modern brain imaging methods allow us to measure this competition and relate it to behavioral outcomes. I will describe examples of these processes, drawn from different domains of psychological function. These examples provide evidence that competition among different brain mechanisms can provide insight into the seemingly inconsistent or irrational responses that are of such interest to behavioral scientists. In particular, I will entertain the idea that a broad range of

■ *Jonathan D. Cohen is Eugene Higgins Professor of Psychology and Director, Center for the Study of Brain, Mind and Behavior, both at Princeton University, Princeton, New Jersey. He is also Professor of Psychiatry, University of Pittsburgh, Pittsburgh, Pennsylvania.*

decisions engage evolutionarily old brain mechanisms that have consistently been implicated in emotional processing. While these mechanisms are still of great value in many circumstances, in others they may no longer be the most reliable stewards of our best interests.

I begin with an example concerning moral reasoning—a domain somewhat removed from economics—partly to establish at the outset the generality of the issues and principles I wish to highlight, and partly because of the compelling nature of the findings in this domain. I then provide two examples from economic decision making, followed by findings from an even more remote domain—research on placebo responses to pain—that further underscores the generality of the principles of brain function involved: that most evaluations—from the most primitive to the most sophisticated—engage multiple different brain systems and that when these result in different appraisals, behavior reflects the outcome of the competition among them.

These findings are likely to have increasing relevance to economists and sociologists. I will also argue that ultimately these issues may matter most of all to policymakers, for at least two reasons. First, some of our most important social policies are meant precisely to protect us from emotional responses that may have once served us well, but now can interfere with our welfare. Second, the development of the human capacity for rational thought has given rise to powerful new technologies that greatly increase our ability to control our behavior and our environment. Many of these technologies are becoming almost universally available—but the expression of our ability for rationality and control does not seem to be as universally expressed. This situation may pose important challenges to social and perhaps even ecological stability. A better understanding of how the mechanisms underlying emotional responses interact with those responsible for rational thought and the control of behavior may help guide the development and implementation of better social policies to address these concerns.

Emotions, Optimality and the Brain

Let me start with a central conjecture of standard economic theory: people make decisions that maximize their utility.[1] This assumption holds that people always act rationally in accordance with their long-term goals. Of course, these assumptions are not entirely realistic. Nobel prizes in economics have been awarded for elegant demonstrations of this counterpoint, like those to Herbert Simon in 1978 and Daniel Kahneman in 2002. The interesting question, however, is *how and why* people deviate from optimality, and here several possibilities arise.

[1] This standard definition of optimality in economics invites a number of important questions. Whose utility is being maximized: the individual, the social group or the species? Over what time frame: immediate or a lifetime? What are the relevant dimensions of a good or an opportunity that contribute to its utility? I will return to several of these questions further on.

One possibility is to blame seeming deviations from optimality on a failure of the theorist's imagination, rather than the irrationality of the decisionmaker. From this perspective, social scientists plying their analytic and experimental skills will eventually identify the relevant sources of utility that are being optimized in such cases. Without care, however, this tack could become circular: assuming that people are always optimal, it is by necessity true that what they do is optimal. Nevertheless, it is possible that as yet undiscovered principles of utility—perhaps involving a complex interplay of individual and social values—will someday provide a more principled explanation for how people actually behave. Economists, and their colleagues in psychology and sociology, continue to search for such principles.

At the other extreme, it is possible that people are simply incapable of optimally maximizing their utility for a variety of reasons, including limited access to information (about the past, present or future), limited ability to learn and limited ability to focus and control behavior. That is, people seek to optimize utility subject to constraints. Behavioral economists have devoted considerable effort to identifying and trying to understand these constraints (for example, Kahneman, 2003; Lowenstein, 1996; O'Donoghue and Rabin, 1999). In a related approach, emotions are often blamed for their rude intrusions on rational behavior. However, this explanation begs important questions: What are emotions, how do they work, and why do they not want us to achieve our best?[2] These questions are not new, nor is it likely they will be answered satisfactorily any time soon. However, advances in neuroscience are beginning to bear on these questions and to open up new avenues of inquiry about the relationship of emotions to rational behavior.

Among the most recent and exciting developments in neuroscience has been the introduction of methods for imaging the function of the intact human brain. This in turn has opened up the opportunity to study the involvement of the brain in uniquely human activities, such as reasoning and complex forms of decision making. These studies provide support for a view of the brain as a confederation of systems and behavior as the outcome of an interaction among these. For the most part, these systems work cooperatively to guide behavior. However, when disagreements arise, behavior reflects the outcome of a competition among systems. Such circumstances are illuminating, as a way for identifying the functions of component mechanisms (that ordinarily work seamlessly together and are therefore difficult to distinguish) and by providing a possible explanation for apparent anomalies in behavior. Some of the systems in the brain—including ones that have consistently been associated with emotional processing—appear to rely on mechanisms that are

[2] The nature of emotions has, of course, been the subject of intense inquiry and debate for entire fields of science, which is well beyond the scope of this article. For present purposes, I will use "emotions" to refer to low-level psychological processes that are engaged by stimuli (or memories) with evaluative significance (different for each type of emotion) and elicit strong and stereotyped behavioral responses. By this definition, emotions are a type of valenced automatic response (as will be discussed further below), which means that emotions are automatic processes associated with strong positive or negative utility (for example, fear is a response to an aversive stimulus that leads to withdrawal, whereas visual recognition of a chair is also a rapid, automatic response, but does not carry specific value information nor command a particular behavioral response, unless one is tired and looking for a place to sit).

evolutionarily old, highly conserved (that is, relatively unchanged over evolutionary history) and well-adapted for particular purposes, but are now subject to engagement in ways that do not always serve our best interests. This argument is built on the premise that evolution is an efficient, but also an opportunistic, process. It is efficient in the sense that it often produces solutions that effectively (sometimes even optimally) meet the needs of the organism given the resources it has available (for example, Anderson, 1993; Barlow, 1981; Brenner et al., 2000; Hecht et al., 1942; Holmes et al., 2004). It is opportunistic in the sense that the solutions produced are heavily shaped by the specific circumstances and constraints under which they arise, which include mechanisms inherited from earlier adaptations as well as the current state of the environment.

However, when circumstances change, mechanisms that once were adaptive may no longer be as fully so. That is, evolution can produce mechanisms that may be *locally* but not *universally* optimal. Because evolution tends to be conservative, such mechanisms may persist and continue to operate, including in new circumstances for which they are not as well-adapted (for example, Waldman, 1994; Barkaw, Tooby and Cosmides, 1992; Pinker, 2002). Emotions may reflect such adaptations (for example, Wright, 1994; Tooby and Cosmides, 1990). That is, emotions may reflect the operation of mechanisms that evolved as efficient, perhaps even optimal, responses to circumstances that prevailed during our development as a species, conferring advantages that helped ensure our survival. However, circumstances have now changed. As a result, while our emotional responses are no doubt still of great value in many circumstances, they are not always the reliable stewards of our best interests that they once were.

At the same time, arguments that focus on emotions and deviations from optimality can easily overlook a glaring and intriguing fact: Economists, who devised the standard economic model and define optimal performance against which it is referenced are, well. . . humans. The very fact that economic theory, contrived by the human mind, can describe optimal behavior provides *prima facie* evidence that human beings can conceive of optimal behavior and, therefore, in principle are capable of it. There is good reason to believe that this capability—presumably dependent on higher cognitive faculties such as reasoning, planning and problem solving—relies heavily on the function of a particular set of brain structures, including the prefrontal cortex, as I will discuss shortly. In other words, these structures may be a critical substrate for "homo economicus."

However, as already noted, higher cognitive abilities do not universally govern behavior. One reason for this, suggested by a long legacy of psychological evidence, is that our higher cognitive faculties have limited capacity. For example, we can only contemplate one "thought-provoking" problem at a time and are also seriously limited by the number of factors that can be simultaneously contemplated regarding that problem. The use of these resources for one problem therefore limits their use for others, making them valuable when needed, but less efficient than more automatic mechanisms that can often be engaged simultaneously (like driving a car and carrying on a casual conversation with a passenger). Thus, the confederation

of systems within the brain includes different types of mechanisms: one type, including emotional processes, that is highly specialized for particular purposes, can respond rapidly, automatically and definitively to eliciting events, but is highly stereotyped in its responses and therefore relatively inflexible; and another type that is less rapid in its responses and has limited capacity at any given moment, but is more flexible and able to support behavior consistent with a broader range of goals.

This tradeoff—between speed and specialization on one side and generality on the other side—parallels a similar tradeoff faced by engineers in the design of computers. The standard central processing unit can carry out the full range of computations required for general purpose computing, but is inefficient in carrying out the narrow but frequently used set of functions required for most computer uses, such as displaying a video image on the screen or communicating on a serial port. Dedicated chips are much more efficient at handling these specialized and well-defined operations, but such chips are typically ill-suited for handling the broad range of functions of a general purpose central processing unit. Modern computers have chips of both types that work closely together. Natural evolution seems to have arrived at a similar solution, in the form of highly automatic responses (including emotions) for many critical functions and mechanisms supporting higher cognitive processes that are more flexible and general purpose, but limited in capacity. In many circumstances—including those familiar to humanity's evolutionary ancestors—these different types of mechanisms function synergistically to achieve our goals. However, in the circumstances of modern life, these systems may prescribe different behaviors. In such cases, the outcome of competition between these mechanisms determines behavior.

These ideas are, of course, not new. For example, Minsky (1986) used the term "society of minds" to convey the idea that the human mind can be understood as an aggregate of separate processes, each with its own goals and operating according to its own principles. Furthermore, psychologists have long recognized the distinction between efficient but highly specialized "automatic" processes and less efficient but more general mechanisms involved in "controlled" processing (Cohen, Dunbar and McClelland, 1990; Kahneman and Treisman, 1984; Posner and Snyder, 1975; Shiffrin and Schneider, 1977). This idea has also found its way into the decision-making and economic literatures, where a distinction has been made between System 1 and System 2 mechanisms (Kahneman, 2003; Camerer, Loewenstein and Prelec, 2005; Stanovich, 1999). System 1 corresponds closely to automatic processing; it quickly proposes intuitive answers to problems as they arise. System 2 corresponds closely to controlled processes; it monitors the quality of answers provided by System 1 and, in some situations, corrects or overrides these judgments. What is new, however, is that researchers now have the tools to begin to identify and characterize these systems at the level of their physical implementation in the human brain. Neuroscience gives detailed access to the mechanisms that underlie behavior and thus may allow scientists to answer questions that cannot be answered easily, or at all, by observing behavior alone. For example, neuroscientists have begun to identify and distinguish among pathways that mediate different types of responses to the same emotional stimulus, some of which

involve only lower-level mechanisms and others that engage higher-levels systems, including the prefrontal cortex (LeDoux, 1996).

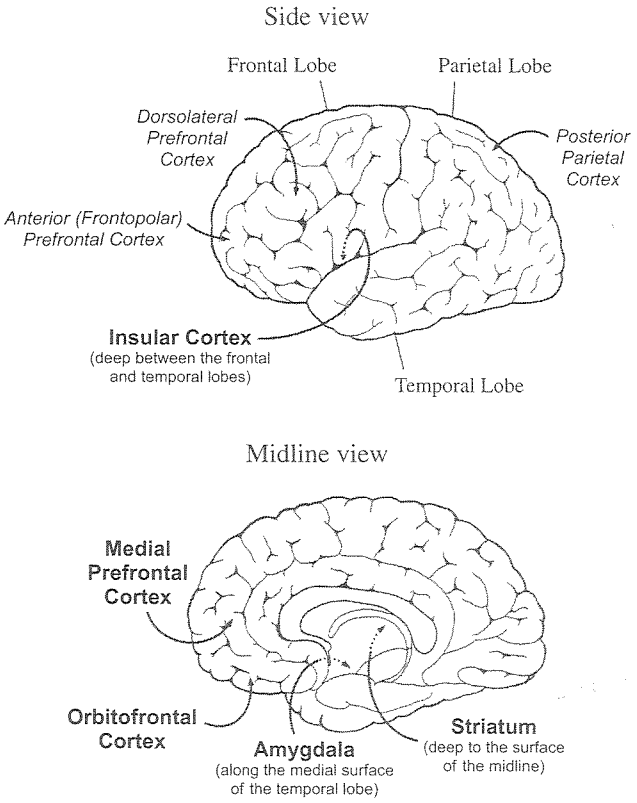
Neuroscience research has profited by a broad array of tools, ranging from the study of patients with brain damage to the direct recording of neural activity in nonhuman species and the construction of sophisticated computational models of neural function. However, only recently has it become possible to track the activity of specific brain areas in normal human subjects while they perform cognitive tasks. This has been made possible by the advent of methods such as positron emission tomography (PET scanning) and functional magnetic resonance imaging (fMRI). Most current studies use fMRI, because it has the advantage of being noninvasive (requiring no injections), can exploit the large installed base of MRI scanners, and provides the best available combination of information about the location and timing of brain activity. However, this method also has important limitations.

First, and most importantly, it does not measure brain activity directly. Rather, it infers brain activity by indexing changes in local blood oxygen concentration (Kwong et al., 1992; Ogawa et al., 1992). Although this measure is highly correlated with changes in neural activity, its indirectness limits fMRI signals to a spatial resolution of millimeters and a temporal resolution of seconds. This scale is far cruder than the most detailed features of brain organization and function. Nevertheless, it is well-suited for certain purposes, in particular the study of neural mechanisms on a scale relevant to many psychological processes.

Perhaps the most important caution to keep in mind about fMRI is that—like many other methods in neuroscience—it is correlative; that is, it reveals a correlation between brain activity and a task manipulation or behavioral response. As such, it cannot definitively establish the causality of the relationship between a pattern of brain activity and a particular psychological function or behavior. Nevertheless, the results often invite inferences about causality that can be tested in subsequent studies. Accordingly, neuroimaging studies—together with the use of more traditional methods in neuroscience—have driven substantial progress in understanding the neural mechanisms underlying emotional and cognitive processes.

At the coarsest level, the brain can be divided into the neocortex—the folded sheet of cells that forms the outer surface of the brain—and deeper, evolutionarily older subcortical structures (below the cortex) that include the striatum (near the brain's core) and the brainstem (at its base). It has long been known that several subcortical structures, particularly those in the brainstem that release the neurotransmitter dopamine and those in the striatum that are influenced by the release of dopamine, respond directly to rewarding events themselves or to their anticipation (for example, Schulz, Dayan and Montague, 1997; Knutson et al., 2001). These structures are believed to be involved in fundamental forms of reinforcement learning and are almost universally affected by drugs of abuse (Montague, Hyman and Cohen, 2004). These, and other subcortical structures responsive to valenced events (that is, events associated with positive or negative utility), make direct connections with several structures within the frontal lobes (the part of the brain just behind the forehead) and temporal lobes (the part of the brain just beneath the temples) that also appear to code value (for example, Craig, 2002; O'Doherty

Figure 1
The Human Cerebrum



Notes: Lateral (side) and medial (midline) views of the human cerebrum, identifying areas critically associated with decision making. Areas in bold have consistently been associated with emotional processing, while areas in italics have consistently been associated with higher level cognitive processes (see text).

et al., 2001; Phelps et al., 2001; Rolls, 2000; Whalen et al., 2001). These cortical areas include medial and orbital regions of frontal cortex (along the inner surfaces and base of the frontal lobes, respectively), the amygdala (along the inner surface of the temporal lobes), and insular cortex (at the junction of the frontal and temporal lobes) as shown in Figure 1. These cortical structures, along with their subcortical counterparts, are classically referred to as the limbic system of the brain, and are thought to be critical to emotional processing (Dalgleish, 2004).

In contrast, a considerable body of evidence indicates that our higher cognitive faculties, including deliberative thought, abstract reasoning, problem solving, planning and language all rely heavily on additional brain structures. One set that is almost universally implicated in such higher-level processes includes anterior and dorsolateral regions of prefrontal cortex, lying along the upper and front most surfaces of the frontal lobes (Duncan, 1986; Koechlin et al., 1999; Miller and

Cohen, 2001; Shallice and Burgess, 1991; Stuss and Benson, 1986).[3] The prefrontal cortex occupies one-third of the neocortex and is one of the brain areas that has expanded most in humans relative to other primate species. Strictly speaking, the prefrontal cortex also partially encompasses some of the emotion-related areas noted above. However, for simplicity, and because areas associated with higher cognitive function comprise the greatest extent of the prefrontal cortex, I will use this term to refer to those regions in the remainder of this article. At the broadest level, two categories of function can be ascribed to prefrontal cortex: reasoning abilities and the capacity for cognitive control—that is, the ability to guide thought and action in association with abstract goals or intentions, especially when this requires overcoming countervailing habits or reflexes (Miller and Cohen, 2001).

By exploiting this type of knowledge about brain organization and function, and determining which brain systems are associated with a particular behavior, researchers may be able to understand better the processes driving the behavior. Here, I illustrate this point with examples of research from my laboratory and that of my colleagues. These examples are drawn from three different domains of psychological function. Together, they provide converging evidence that behavior—in circumstances that seem to elicit the inconsistent or irrational responses of such interest to behavioral scientists—is determined by a competition between brain systems that arrive at different evaluations of the same circumstances.

Moral Reasoning

Philosophers have long been concerned with identifying a rational basis for moral principles. While a rational basis for morality may eventually be discovered by philosophers, it is evident in the meantime that the moral intuitions of ordinary individuals display a number of inconsistencies. The Trolley Dilemma highlights the sort of inconsistencies that are common in moral intuitions (Thomson, 1986). This dilemma is comprised of two scenarios.

In the Switch Scenario, a trolley is heading down a track on which there are five railroad workers, who will be killed if nothing is done. You are an operator at the local switch station. You are too far away to warn the workers; however, if you flip the switch quickly, you can divert the trolley onto a sidetrack. Unfortunately, there is another worker on that track, and he will die, but your action will spare the other five. Is it morally acceptable to flip the switch? The question here is not whether you think you *would* flip the switch but rather whether you *should* or, in any event, whether it is ethically acceptable to do so. When asked this question, most people respond that it is acceptable to flip the switch (Greene et al., 2001).

[3] As with all other mental functions, higher-level cognitive processes such as these rely on interactions distributed across many brain regions, including the temporal and parietal lobes, the cerebellum and parts of the striatum. In this article, I will focus on prefrontal cortex, both for expository purposes and because of its central role in higher cognitive functions. However, the specific functions carried out by prefrontal cortex, the organization of these functions within prefrontal cortex and their interaction with other brain areas also critical to higher cognition remain active and critically important topics for research.

Now consider the Footbridge Scenario. Here again, a trolley is destined to kill five workers. This time, however, you are a bystander on a footbridge that crosses over the track. There is also a very large worker on the footbridge who is close to its edge. You realize that if you push him off the bridge, he will fall directly on the track and block the trolley. Although he will be killed, this action will spare the five workers further down the track. You are certain that you can successfully push him off and that he is big enough to stop the trolley, but you are too slight an individual to stop the trolley if you jumped off the bridge yourself. Is it acceptable to push the worker off the bridge? In this case, most people respond that it would not be morally acceptable to do so (Greene et al., 2001).

The inconsistency here, of course, is that people have different intuitions about two cases that are identical from an "actuarial" point of view. In the Switch Scenario most people seem to feel that it is acceptable to trade one life for five; in the Footbridge Scenario, they do not. Philosophers and ethicists have struggled to explain this apparent inconsistency, focusing on differences between the two situations that may explain the differences in behavior. For example, in the Switch Scenario, the worker's death is *incidental*: It is simple misfortune that he is there. In the Footbridge Scenario, however, the worker's death is *instrumental*. That is, you would be using him as an instrument (a means) to stop the trolley. This suggests that our moral intuitions in this case may rest on a single abstract principle: fellow human beings may be used only as ends, and not as means (Kant, 1785 [1959]).

However, this principle cannot explain the Trolley Dilemma. Consider a slight variant of the Switch Scenario, in which everything is the same except that the sidetrack rejoins the main track just before the location of the five workers. Now, if there is no worker on the sidetrack, the trolley will continue on to kill the five workers. In other words, if you flip the switch the death of the worker on the sidetrack will, in this case, be instrumental—you will be using him as a means to spare the others. Not surprisingly, adding the extra bit of track (connecting the sidetrack to the main track) does not alter people's intuitions. They still reliably indicate that it is morally acceptable to flip the switch. This suggests that instrumentality cannot be the sole criterion by which people judge the morality of such acts. Other factors differentiate the Switch and Footbridge scenarios. To date, however, none have proven able to explain the inconsistencies in people's moral intuitions about these and a large number of related dilemmas. That is, it appears that people's moral intuitions, like their economic behavior, are not easily explained by universal rational principles.

Some philosophers, such as my colleague Josh Greene, as well as social psychologists such as Jonathan Haidt, have offered a different account of these findings. They have suggested that the circumstances of some moral dilemmas may elicit emotional responses while others do not (Greene and Haidt, 2002). This differential engagement of emotion might explain the inconsistencies that are observed. For example, pushing the worker off the bridge to his death may arouse a greater negative emotional response than the prospect of flipping a switch that reroutes the trolley. This negative response may produce an automatic, unreflective judgment that leads them to interpret "feels bad" as "is wrong." A growing body of

behavioral evidence favors this hypothesis (for example, Wheatley and Haidt, 2004).

This hypothesis also makes predictions that can be tested using brain imaging techniques. In one such study, we asked participants to respond to a series of dilemmas while their brain activity was scanned using fMRI (Greene et al., 2001). The dilemmas were designed to fall into one of three categories: "personal" moral dilemmas like the Footbridge Scenario, requiring the infliction of direct, serious bodily harm on another individual to achieve some good; "impersonal" moral dilemmas like the Switch Scenario, involving the infliction of harm in a more removed fashion; and a set of nonmoral dilemmas included as a comparison requiring a similar degree of mental effort, as judged by response times.

Our prediction was that the personal moral dilemmas would elicit emotional responses that influenced participants' moral intuitions. Our findings were consistent with this prediction. When participants contemplated personal moral dilemmas, brains areas were activated that have consistently been associated with emotional processing in other brain imaging studies, such as medial frontal cortex. In contrast, impersonal moral dilemmas elicited activity in dorsolateral areas of the prefrontal cortex (at the upper and outer surface of the frontal lobe, as shown in Figure 1) consistently associated with cognitive processes such as working memory, abstract reasoning and problem solving, but not commonly observed in direct association with emotional processing. Indeed, these areas were virtually identical to those activated by the nonmoral dilemmas, suggesting that both impersonal moral and nonmoral dilemmas engaged similar abstract reasoning processes. Findings from a subsequent study have provided additional evidence that, within the context of moral reasoning tasks, activity in the prefrontal cortex precedes and is directly associated with utilitarian judgments (Greene et al., 2004; Figure 2)—that is, judgments favoring the greater good, even when this competes with a strong aversive response to the prospect of harming another individual.

Why does the apparently rational prefrontal cortex seem to prevail in some cases, like the Switch Scenario, but an emotional response prevails in other cases, like the Footbridge Scenario? One answer to this question considers how such emotions might have evolved. Perhaps emotional aversion to harming other humans evolved as an adaptation that allowed early humans to aggregate more effectively into stable social structures, conferring upon them a competitive advantage (Wrangham and Peterson, 1996; Wilson, 2002). However, this adaptation would have arisen at a time when the scope of aggression was limited literally to a stone's throw; that is, there would not have been strong pressure to develop an emotional aversion to harming other humans at greater distances since this was simply not possible. Indeed, in our studies, an important factor that distinguished the personal from the impersonal moral dilemmas was the directness or proximity with which harm had to be inflicted on a victim to achieve the greater good. The evolution of our emotional apparatus did not anticipate a world in which aggression can be expressed impersonally over large distances. This may explain why many people have a strong emotional response to the prospect of pushing a worker off a bridge, but less so to flipping a switch that will cause a trolley some distance away to kill a worker.

I do not want to place too much emphasis on the particular evolutionary hypothesis I have considered here. Anthropologists and evolutionary biologists are likely to have much more to say about such matters than psychologists or economists. Indeed, even the approach of drawing upon evolutionary explanations to understand neural and psychological mechanisms—sometimes referred to as evolutionary psychology—is highly controversial and has been criticized for offering nothing more than a series of "just so" stories. However, even if scientists are never able to establish with certainty the specific evolutionary or developmental course of a neural or psychological mechanism, the evolutionary perspective can nevertheless be used to generate testable hypotheses that can lead to deeper insights into the nature of the mechanisms involved and the circumstances in which they are likely to be engaged. For example, the hypothesis above might be tested by manipulating proximity—say, by including a condition in which the worker on the footbridge is plunged onto the track by flipping a switch that releases a trap door, rather than directly pushed. Such work is currently in progress.

Research along these lines stands to gain by exploring the large set of emotional processes that may influence moral decision making, how they operate and the circumstances in which they become engaged. This might reveal general principles of emotional engagement (such as the "psychological distance" of an eliciting event; Trope and Liberman, 2003), in addition to specific features attendant to particular emotions. Such insights into the factors that trigger (or fail to trigger) emotional responses offers the promise of a deeper understanding of some seemingly inconsistent or irrational behaviors and perhaps even how they may be addressed.

Economic Decision Making: Fairness and Intertemporal Choice

Behavioral economists have also identified a large number of circumstances in which people seem to behave inconsistently and in which their decisions deviate from what is predicted by the rational principles of the standard economic model. As in the case of moral reasoning, such behavior may be driven by emotional responses. We have explored this possibility in two contexts, one involving the ultimatum game (Camerer, 2003; Guth et al., 1982) and the other involving intertemporal choice (Ainslie, 1975; Frederick et al., 2002).

The Ultimatum Game

In this game, a pair of partners is given an endowment that they must split. One partner proposes how to split the sum, and then the other may accept or reject the offer. If the offer is accepted, each player is allotted the proposed amount. If the split is rejected, neither partner receives anything, and the game is over. Standard economic theory suggests that any split offering a small but nonzero amount should be accepted. After all, for the partner who must decide whether to accept or to reject the offer, something is better than nothing. However, offers of less than about 20 percent of the sum are routinely rejected, even though this means getting nothing. This basic outcome holds across a wide range of cultures

and sizes of endowments, including circumstances in which offers as large as a month's pay are rejected, if they are perceived to be unfair (Henrich et al., 2001). One account of this behavior is that it reflects a commitment to establishing a bargaining position (or one's reputation) in anticipation of future interactions. However, unfair offers are rejected even in "single shot" versions of the game, when concerns over bargaining position and reputation should not matter because participants know they will interact with a given partner only once and in a fully confidential manner (Camerer and Thaler, 1995).

An alternative hypothesis is that rejections of low offers are driven by an emotional response—perhaps anger at the unfairness of the offer or pleasure at the thought of punishing someone who has tried to take advantage. We tested this possibility in a neuroimaging experiment (Sanfey et al., 2003). Participants were introduced to ten partners with whom they were told they would play one round each of the ultimatum game in confidence, always in the role of the recipient of the offer. They were then scanned as they entertained and responded to offers from each of the partners. They were also presented with some offers that they were told were generated by a computer program. Consistent with the results of previous studies, offers of 20 percent or less were typically rejected. Entertaining such offers was reliably associated with brain activity in two areas of interest: a region of dorsolateral prefrontal cortex very similar to the one observed in the moral reasoning study; and the anterior insula, which in previous imaging studies has consistently been associated with negative emotional responses such as physical pain and disgust (Calder et al., 2001; Derbyshire et al., 1997; Wicker et al., 2003). The strength of activity in the anterior insula correlated directly with the likelihood that subjects rejected the offer. When activity in the anterior insula exceeded the level of activity in the prefrontal cortex, the offer was significantly more likely to be rejected than when the reverse held true. These effects were greatest when the offer was thought to come from a human rather than from a computer program.[4] These findings suggest that when participants reject an unfair offer, it is not the result of a deliberative thought process. Rather, it appears to be the product of a strong (seemingly negative) emotional response.

This negative response may seem irrational: after all, it forfeits a present reward without any obvious offsetting future benefit. However, once again, this aversive response may reflect a conserved adaptation to evolutionarily older circumstances. At the time that sociality began to evolve, the size of our social groups was much smaller, and thus the likelihood of repeated interactions with the same individual, or others who had heard about you, was much greater. In such circumstances, a strong negative emotional response to an unfair offer would ensure that people acted in accord with their own longer-term interest—and against the temptation of an immediate reward—by committing them to acts that protected

[4] One might wonder, here, what the fate of such effects will be as people have greater experience with computers. On one hand, one might expect emotional responses to computers to decrease as they become more familiar. On the other hand, as computers become more sophisticated, and anthropomorphic, they may elicit responses that are increasingly similar to those from other humans.

their reputation and discouraged exploitative behavior in the future. That is, the response was locally optimal in a world where people interacted mainly with a small social group, and the need to protect reputation over the long term outweighed modest momentary gains. This may not be as advantageous in a modern world with many more anonymous interactions.

Again, caution about this interpretation is warranted. Some have suggested altruistic punishment—that is, acting to punish certain behaviors even with no immediate gain to oneself—is a quite rational form of behavior when viewed from the broader perspective of social evolution. For example, Fehr (2003; Fehr and Simon, 2000) has suggested that social groups are more likely to remain intact during periods of fitness challenges if they are made up of individuals who execute altruistic punishment. Thus, what appears to be irrational behavior when viewed from the local perspective of the individual may in fact be explained in rational terms at a broader social level.[5] But whatever the story behind the evolution of this mechanism, the neural evidence suggests that the outcome of behavior is determined by the relative degree of engagement of an emotional response and processes mediated by the prefrontal cortex. Such findings provide strong impetus for the construction of economic models that take account of emotional factors (for example, Bolton and Ockenfels, 2000; Fehr and Schmidt, 2000).

Intertemporal Choice

Standard economic theory suggests that intertemporal choice should be consistent over time. For example, if I prefer $10 now to $11 in a week, then I should have the same preference for $10 a year from now over $11 a year and a week from now. That is, barring circumstantial factors, a week's delay should be discounted the same whether it occurs this week or in a year. However, when presented with such choices, people often exhibit inconsistencies, for example preferring the $10 option if that amount can be received immediately, but the $11 option if the two rewards are postponed a year or a year and a week into the future. In general, people act more impatiently when making decisions about shorter-term options than longer-term ones. This tension is apparent in many real world situations. For example, people will often pay extra for express delivery of a desirable (but nonessential) item that is in stock, but not for one that is out of stock and therefore will not be deliverable until later. This phenomenon has also been used to explain addiction behaviors. An alcoholic may rationally plan to avoid walking past the bar on his way home from work, but if he happens to pass a bar on his alternate way home, his preference may reverse.

This pattern suggests that many people do not discount future outcomes at a constant exponential rate, but instead practice hyperbolic-like discounting, with a disproportionate weight on immediate rewards (Herrnstein, 1997; Rachlin, 2000). However, at a deeper level, preference reversal may reflect the operation of two

[5] Even if this interpretation holds—that altruistic punishment is rational at a group level—nevertheless, this behavior may no longer be optimal in a highly structured society that has effective and possibly more efficient social technologies for enforcing fairness at the group level, like laws and taxes.

fundamentally different mechanisms, one of which strongly prefers immediate rewards, while the other adheres to a more ecumenical weighting of preferences across time (Laibson, 1997; Loewenstein, 1996; Metcalfe and Mischel, 1999; Shefrin and Thaler, 1988). That is, intertemporal choice—like moral reasoning and bargaining—may under some circumstances engage different brain systems that operate according to different principles.

My colleagues and I hypothesized that responses to immediate rewards may reflect the engagement of highly conserved mechanisms in what has classically been referred to as the limbic system of the brain. In particular, we predicted the involvement of medial frontal and striatal structures, which receive dense projections from brainstem dopamine neurons. As noted earlier, it has long been recognized that these play a central role in the evaluation of reward and reinforcement learning and are also known to be engaged by most drugs of abuse. At the same time, we conjectured that mechanisms in the prefrontal cortex and associated structures may be more sensitive to the potential value of future rewards.

A recent brain imaging study corroborated these predictions. When people are offered an option for immediate reward (in this case, an Amazon.com gift certificate), limbic structures—including those mentioned above—were activated significantly more than when both options in the choice involved future delivery (two or more weeks from today). In contrast, areas in the prefrontal cortex and associated structures, such as the parietal cortex (including an area believed to be important for numerical comparisons), were engaged by all choices. The activity of these structures was directly associated with choice outcome. For choices involving an option for immediate reward (thus engaging both limbic and prefrontal systems), selection of the *later* option was associated with significantly greater prefrontal cortex activity than choice of the immediate one. Thus, once again, it appears that prefrontal cortex and its associates support more considered forms of evaluation and that a competition between these and limbic structures determined behavior. These studies are currently being extended to other types of rewards, delivered over a broader range of time frames, and the results are being used to guide the development of formal models of the decision-making mechanisms involved.

Placebo Response

The interaction between prefrontal cortex function and lower-level evaluative processes is perhaps most strikingly observed in the context of the placebo effect. Placebo studies find that when people are told to expect that a medication will work in a certain way, they often experience it working—even if the supposed medication is only an inert substance. For example, in a study by Wager et al. (2004), participants were treated with a cream on their arm prior to receiving a series of painful electric shocks. In one condition, participants were told that the cream was a highly effective new topical anesthetic that would reduce the pain. In another condition, they were told that the cream was electrode paste. In actuality, the cream was the same inert substance in both conditions. During brain scanning, a warning

cue was presented indicating the strength of the upcoming shock, followed by a delay prior to its delivery. In this way, brain activity could be measured during the anticipation of the shock and during its delivery. About one-third of participants experienced a placebo response to the alleged anesthetic (which is typical of placebo studies). This response was associated with a significant reduction of activity in pain-responsive regions of the brain during shock delivery.

Most interesting, however, was the pattern of activity during the anticipatory period. Several regions of the prefrontal cortex showed an increase in activity that preceded and predicted both the degree of placebo relief experienced and the reduction of activity in pain responsive areas observed during the later delivery of the shock. Thus, it appears that higher-level representations of expectation (in this case, pain relief) within the prefrontal cortex may be able to control even the most primitive and conserved of evaluative mechanisms: the visceral response to pain.

Discussion

The studies described above illustrate two of the central points of this article. First, a variety of decisions—ranging from moral judgments to economic choices—can engage emotional responses thought to be highly conserved by evolutionary pressure. These mechanisms may be well-adapted to some circumstances, but in others introduce inconsistency and in some cases arguably suboptimal behavior. The second point is that, under these circumstances, the engagement of the prefrontal cortex can override these processes and support more "rational" forms of behavior. Neuroimaging findings have shed light on both of these facets of prefrontal cortex function—control and reasoning—demonstrating the generality of processing domains in which the prefrontal cortex can exert control over behavior (including the most primitive of responses, such as pain) and its close association with higher-level cognitive processes such as reasoning and problem solving. Some studies have gone so far as to suggest that prefrontal cortex function is directly related to measures of general intelligence (Duncan et al., 2000). Needless to say, however, many critical questions remain unanswered. For example, do the functions supported by prefrontal cortex comprise a unitary system or a family of closely related but distinguishable mechanisms as appears to be the case for many automatic processes, such as emotions?

The Times They Are a Changin'

Much remains to be learned about the neural mechanisms supporting both emotions and higher level cognition. Nevertheless, the observations made so far invite an intriguing speculation. The evolution of the prefrontal cortex, and the attendant capacity for reasoning and control that it has conferred upon humans, has given rise to dramatic new technologies and social change. In turn, these technologies and changes have created a variety of physical and social circumstances for which older, more conserved mechanisms may no longer be as well-adapted. Indeed, the evolution of the prefrontal cortex, in creating circumstances some of which have rendered our emotional responses to be less adaptive, may lie

at the heart of important dimensions of social and perhaps even evolutionary instability. That is, the pace of change in our environment engendered by the prefrontal cortex may have outstripped the pace of evolution, introducing challenging, and possibly even dangerous, social and ecological instabilities. At the same time, the development of the prefrontal cortex may hold the key to averting the consequences of these instabilities. Let me consider these claims in slightly more detail.

The first claim—that the emergence of the prefrontal cortex has given rise to modern circumstances that render older, conserved emotional responses maladaptive—runs in two directions. Some circumstances trigger emotional responses that are no longer adaptive, whereas others fail to elicit responses that are appropriate. For example, consider the case of intertemporal choice. Steep discounting may have been highly adaptive when most (if not all) valuable resources were perishable or were difficult to defend given the lack of well-defined and well-enforced property rights. However, with the evolution of the prefrontal cortex, and the concomitant development of technologies from refrigerators to bank accounts, steep discounting and impulsive behaviors are substantially less adaptive, yet they persist. Furthermore, some technologies seemed to have exacerbated such maladaptation with potentially serious social consequences. For instance, the remarkable array of desirable and immediately available products, coupled with clever advertising schemes that appeal to basic impulses and desires, may be important contributors to the persistent failure of people to save adequately for their own retirement. The prefrontal cortex has also helped create large quantities of addictive substances—substances that, ironically, target dopamine neurotransmission in limbic structures, the very mechanism that appears to favor immediate reinforcement.

Conversely, the Trolley Dilemma suggests that technology has literally put distance between our capacity to harm others and the aversive response that this ordinarily triggers. In the Switch Scenario, this distance is beneficial and serves the greater good, allowing five lives to be spared at the cost of one. But in other cases, the combination of emotion and distance may be less salubrious. Consider the situation in which an opportunistic head of state is considering the use of nuclear weapons preemptively against a competing nation. Would he be as likely to proceed if, setting practicalities aside, he had to inflict the same degree of harm first-hand? Indeed, in the heat of the cold war, it was suggested that the "nuclear hot button" be surgically implanted in the chest of the president's closest associate, requiring the president to rip that person's chest open to access it. The point of this procedure would be to leverage an evolutionarily old adaptation (an aversive emotional response to personal violence) in the service of a modern need for caution—a need that technology has introduced but to which our emotional responses may not be fully adapted.

The Vulcanization of the Human Brain

The truly remarkable fact, however, is that the prefrontal cortex seems to provide us with the capacity to manage the very circumstances it has created. In the moral reasoning study, prefrontal cortex activity was associated with utilitarian

decisions that were made in the face of aversive emotional responses, and in the economic decision-making studies prefrontal activity was associated with the choice of more remunerative options in the face of pride or impatience. This outcome is consistent with the evidence that the prefrontal cortex supports a highly flexible and general reasoning capability, including the type of analysis that economists carry out in determining what constitutes rational behavior.

The title of this article was meant to capture this point. Vulcanization is the process of treating a substance (such as rubber) to improve its strength, resiliency and usefulness. Similarly, evolution seems to have vulcanized the human brain through the development of the prefrontal cortex. In the examples we have considered, this outcome is manifest at the level of the individual. However, it can also be seen in technological and social innovations. For example, the specialized training given to doctors and soldiers involves the cultivation of mechanisms for averting or overcoming strong emotional responses that may interfere with their professional functions. These mechanisms may not rely directly on the prefrontal cortex; instead, they may involve the training of other lower-level mechanisms specific to the particular circumstances involved. Importantly, however, the social structures that devised and support the training procedures almost certainly did rely on the prefrontal cortex. Where training is ineffective, social and technological innovations have produced mechanisms of precommitment, such as Social Security and 401(k) plans in the retirement system as well as nicotine gum for smokers and antabuse for alcoholics (a drug that makes the ingestion of alcohol toxic). These measures are all designed specifically to protect us against ourselves, and while automatic payroll deductions and antabuse may not engage the prefrontal cortex, once again their development almost certainly did.

That is the good news. The bad news is that this evolutionary development does not seem to express itself fully or uniformly across our species, at the individual or social levels. Just as individuals do not always exhibit rational behavior, support for precommitment technologies or policies does not always prevail in the social and political arenas. The reasons may be based on individual or cultural differences. There is clearly a genetic component to variations in general intelligence, and it is possible that this is due to variations in the development of prefrontal cortex. However, it is still unclear to what extent intelligence (and prefrontal development) is influenced by individual experience, social, cultural and even economic factors (Noble and Farah, 2005). What does seem clear is that the unequal expression of capabilities afforded by the prefrontal cortex poses a potentially serious problem. It takes relatively few individuals with highly focused rational abilities to produce powerful new technologies, but not much of a prefrontal cortex to abuse them. A small number of physicists and engineers can design and build a nuclear bomb, but even a monkey can press the button to detonate it. It is not news that this situation poses a potential threat to the stability of our society. Behavioral economists, sociologists, anthropologists, behavioral ecologists and others have begun to consider the issues involved when the pressures of individual needs and desires exercised over the near term come into conflict with longer-term consequences for the common good (Elster, 2000).

It is tempting to suggest here that if everyone were fully rational, the world would be better off. This speculation conjures memories of Mr. Spock, the character in Gene Roddenberry's Star Trek series who was half Vulcan—a people whose thought and behavior were governed entirely by logic.[6] As the story goes, the reliance on logic reflected the evolution of Vulcan society in response to the threats posed to social and ecological stability by the erratic and destructive influences of emotions. However, even Roddenberry acknowledged the oversimplification of this notion: Vulcans periodically returned to their home planet for a well-guarded annual ritual in which they yielded to an all-consuming catharsis of emotion, and episodes of Star Trek routinely ended with a conciliatory vindication of human emotion.

Indeed, emotions can be critical in galvanizing action for good, moving people to acts of charity and heroism, and encouraging participation in important social functions such as the political process. Even negative emotions such as anger and fear continue to serve important functions such as protection and motivation. Conversely, in situations when emotions are weak, they may fail to provide an impetus to action. One could argue that rational utilitarianism has proven too weak a motivator for political action on the environment, poverty and a wealth of other social problems (including the failure to support programs that properly value future consequences of current actions, such as retirement savings). That is, even rationally determined actions may require emotional impetus to get executed. David Hume (1739 [1969]) noted this long ago, when he argued that reason must be motivated by passion. Conversely, there may be a stiff price to pay when our cognitive apparatus is overexercised. This system can have its own "appetite" for control which, if unabated, can lead to chronic stress and associated disruptions of emotion and visceral responses (like frustration, anxiety and depression) that can ultimately produce widespread damage to vital bodily organs including the brain and even the prefrontal cortex itself (Sapolsky, 1998; McEwen and Lasley, 2002).

Rather than contemplating how reason can prevail over emotion, a more profitable pursuit—as an exercise for the human prefrontal cortex, and for science as its social instrument—is to understand better the conditions under which particular emotions are engaged and interact productively with higher-level cognitive processes and those in which emotional responses may lead us astray. The empirical studies described above are just a few among a growing number that are moving in this direction (for example, Ochsner and Gross, 2005). Such work must also be coupled with appropriate theory development. Toward this end, neuroscientists have begun to construct formal models of several of the systems discussed in this article, including the dopamine system and the prefrontal cortex (as discussed in Miller and Cohen, 2001; Montague, Hyman and Cohen, 2004), as well as other systems central to interactions between cognition and emotion (Aston-Jones and Cohen, 2005; Yu and Dayan, 2005). These formal models, in turn, may inform the development of more detailed and accurate models of economic decision making.

Broader scale theories of social systems may also benefit by these develop-

[6] As far as I can tell, there is no indication that Roddenberry was playing on the meaning of "vulcanization" suggested earlier, so I must accept full blame for this indulgence.

ments. For example, game theory models built around agents that incorporate more detailed and realistic mechanisms of cognition and emotion could be used to explore how the development and spread of capabilities provided by the prefrontal cortex affect the equilibrium dynamics of a population, and its interactions with the environment. Such analyses might reveal important but as yet unappreciated factors that influence these dynamics. These factors, in turn, may suggest new policy interventions that favor stable solutions over unfavorable outcomes.

Summary and Conclusion

The prefrontal cortex evolved in the context of pre-existing emotional processes and must therefore have developed to interact effectively with such processes. However, the full emergence of the prefrontal cortex, and the technological and social innovations to which it has recently given rise, have begun to produce changes in the physical and social environment that have outpaced evolution, creating circumstances for which the emotional mechanisms in our brains may not be fully adapted. This situation, coupled with the apparent heterogeneity in the development of the prefrontal cortex across members of the human race, introduces a potentially fundamental instability: A world in which the potency of technology introduced by the prefrontal cortex, and perhaps manageable by it, is equally accessible to mechanisms that were not adapted to the use of that technology. Albert Einstein has been quoted as saying: "The world we have made, as a result of the level of thinking we have done thus far, creates problems we cannot solve at the same level of thinking at which we created them" (MacHale, 2002). It remains to be seen whether the prefrontal cortex (and associated structures) has the capacity to meet the challenges that its appearance seems to have introduced, or whether evolution has truly taken a bite of Eden's apple.

■ *I would like to express my gratitude to Angus Deaton, Ernst Fehr, Josh Greene, Jonathan Haidt, George Loewenstein, Bruce McEwen, Leigh Nystrom, Alex Todorov and Robert Wright, as well as the editors and reviewers (James Hines and Michael Waldman) for invaluable comments on earlier drafts of this article. I would also like to acknowledge my collaborators on the studies reviewed (John Darley, Josh Greene, David Laibson, George Loewenstein, Sam McClure, Richie Davidson, Steve Kosslyn, Ed Smith, Jim Rilling, Bob Rose, Alan Sanfey and Tor Wager), as well as my colleagues (Daniel Kahneman, Read Montague, Chris Paxson, Eldar Shaffir and Peter Shizgal) whose ideas, support, guidance and tutelage in matters economic, psychological and neuroscientific I deeply appreciate, but who should not be blamed for any distortions of theory or concepts that I have introduced. The research studies reviewed in this article were conducted under the support of grants from the National Science Foundation, the National Institute of Mental Health, the National Institute of Aging, the Seaver Institute, the Mind, Brain, Body and Health Initiative, as well as institutional support from the Center for Health and Well Being and the Center for the Study of Brain, Mind and Behavior at Princeton University.*

References

Ainslie, George. 1975. "Specious Reward: A Behavioral Theory of Impulsiveness and Impulse Control." *Psychological Bulletin.* 82:4, pp. 463–96.

Anderson, John R. 1993. *Rules of the Mind.* Hillsdale, N.J.: Lawrence Erlbaum Associates.

Aston-Jones, Gary and Jonathan D. Cohen. 2005. "An Integrative Theory of Locus Coeruleus-Norepinephrine Function: Adaptive Gain and Optimal Performance." *Annual Review of Neuroscience.* July, 28, pp. 403–50.

Barkow, Jerome H., Leda Cosmides and John Tooby, eds. 1992. *The Adapted Mind: Evolutionary Psychology and the Generation of Culture.* New York: Oxford.

Barlow, Horace B. 1981. "Critical Limiting Factors in the Design of the Eye and Visual Cortex." *Proceedings of the Royal Society of London, Series B.* 212, pp. 1–34.

Berke, Joshua D. and Steven E. Hyman. 2000. "Addiction, Dopamine, and the Molecular Mechanisms of Memory." *Neuron.* 25:3, pp. 515–32.

Bolton, Gary E. and Axel Ockenfels. 2000. "ERC: A Theory of Equity, Reciprocity, and Competition." *American Economic Review.* 90:1, pp. 166–93.

Brenner, Naama, William Bialek and Rob de Ruyter van Steveninck. 2000. "Adaptive Rescaling Optimizes Information Transmission." *Neuron.* June, 26, pp. 695–702.

Calder, Andrew J., Andrew D. Lawrence and Andrew W. Young. 2001. "Neuropsychology of Fear and Loathing." *Nature Reviews Neuroscience.* 2:5, pp. 352–63.

Camerer, Colin. 2003. "Dictator, Ultimatum, and Trust Games," in *Behavioral Game Theory: Experiments in Strategic Interaction.* New York: Russell-Sage, chapter 2.

Camerer, Colin and Richard H. Thaler. 1995. "Ultimatums, Dictators and Manners." *Journal of Economic Perspectives.* 9:2, pp. 209–19.

Camerer, Colin, George Loewenstein and Drazen Prelec. 2005. "Neuroeconomics: How Neuroscience can Inform Economics." *Journal of Economic Literature.* 43:1, pp. 9–64.

Cohen, Jonathan D., Gary Aston-Jones and Mark S. Gilzenrat. 2004. "A Systems-Level Perspective on Attention and Cognitive Control: Guided Activation, Adaptive Gating, Conflict Monitoring, and Exploitation vs. Exploration," in *Cognitive Neuroscience of Attention.* M. Posner, ed. New York: Guilford Press, pp. 71–90.

Cohen, Jonathan D., Kevin Dunbar and James L. McClelland. 1990. "On the Control of Automatic Processes: A Parallel Distributed Processing Account of the Stroop Effect." *Psychological Review.* 97:3, pp. 332–61.

Craig, A. D. 2002. "How Do You Feel? Interoception: The Sense of the Physiological Condition of the Body." *Nature Reviews Neuroscience.* 3:8, pp. 655–66.

Dalgleish, Tim. 2004. "The Emotional Brain." *Nature Reviews Neuroscience.* 5:7, pp. 583–89.

Davidson, Richard J. and William Irwin. 1999. "The Functional Neuroanatomy of Emotion and Affective Style." *Trends Cognitive Science.* 3:1, pp. 11–21.

Derbyshire, Stuart W., A. K. Jones, F. Gyulai, S. Clark, D. Townsend and L. L. Firestone. 1997. "Pain Processing During Three Levels of Noxious Stimulation Produces Differential Patterns of Central Activity." *Pain.* 73:3, pp. 431–45.

Di Chiara, Gaetano and Assunta Imperato. 1988. "Drugs Abused by Humans Preferentially Increase Synaptic Dopamine Concentrations in the Mesolimbic System of Freely Moving Rats." *Proceedings of the National Academy of Science (USA).* 85:14, pp. 5274–278.

Duncan, John. 1986. "Disorganisation of Behaviour after Frontal Lobe Damage." *Cognitive Neuropsychology.* 3:3, pp. 271–90.

Duncan, John, Rüdiger J. Seitz, Jonathan Kolodny, Daniel Bor, Hans Herzog, Ayesha Ahmed, Fiona N. Newell and Hazel Emslie. 2000. "A Neural Basis for General Intelligence." *Science.* 289:5478, pp. 457–60.

Elster, Jon. 2000. *Ulysses Unbound: Studies in Rationality, Precommitment, and Constraints.* Cambridge, UK: Cambridge University Press.

Fehr, Ernst and Urs Fischbacher. 2003. "The Nature of Human Altruism." *Nature.* 425:6960, pp. 785–91.

Fehr, Ernst and Simon Gächter. 2000. "Fairness and Retaliation: The Economics of Reciprocity." *Journal of Economic Perspectives.* 14:3, pp. 159–81.

Fehr, Ernst and Klaus M. Schmidt. 1999. "Theory of Fairness, Competition, and Cooperation." *Quarterly Journal of Economics.* 114:3, pp. 817–68.

Frederick, Shane, George Loewenstein and Ted O'Donoghue. 2002. "Time Discounting and Time Preference: A Critical Review." *Journal of Economic Literature.* 40:2, pp. 351–401.

Greene, Joshua and Jonathan Haidt. 2002. "How (and Where) does Moral Judgment Work?" *Trends Cognive Science.* 6:12, pp. 517–23.

Greene, Joshua D., Leigh E. Nystrom, Andrew D. Engell, John M. Darley and Jonathan D. Cohen. 2004. "The Neural Bases of Cognitive Conflict and Control in Moral Judgment." *Neuron.* 44:2, pp. 389–400.

Greene, Joshua D., R. Brian Sommerville, Leigh E. Nystrom, John M. Darley and Jonathan D. Cohen. 2001. "An fMRI Investigation of Emotional Engagement in Moral Judgment." *Science.* 293:5537, pp. 2105–108.

Güth, Werner, Rolf Schmittberger and Bernd Schwarze. 1982. "An Experimental Analysis of Ultimatum Games." *Journal of Economic Behavior and Organization.* 3:376, pp. 367–88.

Hecht, Selig, Simon Shlaer and Maurice Henri Pirenne. 1942. "Energy, Quanta, and Vision." *Journal of General Physiology.* 25:6, pp. 819–40.

Henrich, Joseph, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis and Richard McElreath. 2001. "Cooperation, Reciprocity and Punishment in Fifteen Small-Scale Societies." *American Economic Review.* 91:2, pp. 73–78.

Herrnstein, Richard J. 1997. *The Matching Law: Papers in Psychology and Economics.* H. Rachlin and D. I. Laibson, eds. Cambridge, Mass.: Harvard University Press.

Holmes, Phil, Eric Shea-Brown, Jeff Moehlis, Rafal Bogacz, Juan Gao, Gary Aston-Jones, Eric Clayton, Janusz Rajkowski and Jonathan D. Cohen. 2004. "Optimal Decisions: From Neural Spikes, through Stochastic Differential Equations, to Behavior." *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences.* 87:A9, pp. 1–8.

Hume, David. 1969. *A Treatise of Human Nature.* E. C. Mossner, ed. London: Penguin Books.

Kahneman, Daniel A. 2003. "Perspective on Judgment and Choice: Mapping Bounded Rationality." *American Psychologist.* 58:9, pp. 697–720.

Kahneman, Daniel A. and Anne Treisman. 1984. "Changing Views of Attention and Automaticity," in *Varieties of Attention.* R. Parasuraman, D. R. Davies and J. Beatty, eds. New York: Academic Press, Inc., pp. 29–61.

Kant, Immanuel. 1785 [1959]. *Foundation of the Metaphysics of Morals.* Indianapolis, Ind.: Bobbs-Merrill.

Knutson, Brian, Grace W. Fong, Charles M. Adams, Jerald L. Varner and Daniel Hommer. 2001. "Dissociation of Reward Anticipation and Outcome with Event-Related fMRI." *NeuroReport.* 12:17, pp. 3683–837.

Koechlin, Etienne, Gianpaolo Basso, Pietro Pietrini, Seth Panzer and Jordan Grafman. 1999. "The Role of the Anterior Prefrontal Cortex in Human Cognition." *Nature.* 399:6732, pp. 148–51.

Koopmans, Tjalling C. 1960. "Stationary Ordinal Utility and Impatience." *Econometrica.* 28:2, pp. 287–309.

Kwong, K. K., J. W. Belliveau, D. A. Chesler, I. E. Goldberg, R. M. Weisskoff, P. Poncelet, D. N. Kennedy, B. E. Hoppel, C. M. S., R. Turner, H. M. Cheng, T. J. Brady and B. R. Rosen. 1992. "Dynamic Magnetic Resonance Imaging of Human Brain Activity during Primary Sensory Stimulation." *Proceedings of the National Academy of Sciences of the United States of America.* 89:12, pp. 5675–679.

Laibson, David. 1997. "Golden Eggs and Hyperbolic Discounting." *Quarterly Journal of Economics.* 112:2, pp. 443–77.

LeDoux, Joseph E. 1996. *The Emotional Brain.* New York: Simon and Schuster.

Loewenstein, George. 1996. "Out of Control: Visceral Influences on Behavior." *Organizational Behavior and Human Decision Processes.* 65:3, pp. 272–92.

MacHale, Des. 2002. *Wisdom.* London: Mercier Press.

McEwen, Bruce S. and Elizabeth N. Lasley. 2002. *The End of Stress as We Know It.* Washington, DC: Joseph Henry Press.

Metcalfe, Janet and Walter Mischel. 1999. "A Hot/Cool-System Analysis of Delay of Gratification: Dynamics of Willpower." *Psychological Review.* 106:1, pp. 3–19.

Miller, Earl K. and Jonathan D. Cohen. 2001. "An Integrative Theory of Prefrontal Cortex Function." *Annual Review of Neuroscience.* March, 24, pp. 167–202.

Minsky, Marvin. 1986. *The Society of Mind.* New York: Simon and Schuster.

Montague P. Read, Steven E. Hyman and Jonathan D. Cohen. 2004. "Computational Roles for Dopamine in Behavioural Control." *Nature.* 431:7010, pp. 760–67.

Noble, Kimberly G., M. Frank Norman and Martha J. Farah. 2005. "Neurocognitive Correlates of Socioeconomic Status in Kindergarten Children." *Developmental Science.* 8:1, pp. 74–87.

Ochsner, Kevin N. and James J. Gross. 2005. "The Cognitive Control of Emotion." *Trends in Cognitive Science.* 9:5, pp. 242–49.

O'Doherty, John, Morten L. Kringelbach, Edmund T. Rolls, Julia Hornak and Caroline Andrews. 2001. "Abstract Reward and Punishment Representations in the Human Orbitofrontal Cortex." *Nature Neuroscience.* 4:1, pp. 95–102.

O'Donoghue, Ted and Matthew Rabin. 1999. "Doing it Now or Later." *American Economic Review.* 89:1, pp. 103–24.

Ogawa, S., D. W. Tank, D. W. Menon, J. M. Ellermann, S. Kim, H. Merkle and K. Ugurbil. 1992. "Intrinsic Signal Changes Accompanying Sensory Stimulation: Functional Brain Mapping with Magnetic Resonance Imaging." *Proceedings of the National Academy of Sciences of the United States of America.* 89:13, pp. 5951–955.

Phelps, Elizabeth A., Kevin J. O'Connor,

J. Christopher Gatenby, John C. Gore, Christian Grillon and Michael Davis. 2001. "Activation of the Left Amygdala to a Cognitive Representation of Fear." *Nature Neuroscience.* 4:4, pp. 437–41.

Pinker, Steven. 2002. *The Blank Slate: The Modern Denial of Human Nature.* New York: Viking Penguin.

Posner, Michael I. and Charles R. R. Snyder. 1975. "Attention and Cognitive Control," in *Information Processing and Cognition.* R. L. Solso, ed. Hillsdale, N.J.: Lawrence Erlbaum Associates, pp. 55–85.

Rachlin, Howard. 2000. *The Science of Self-Control.* Cambridge, Mass.: Harvard University Press.

Rolls, Edmund T. 2000. "The Orbitofrontal Cortex and Reward." *Cerebral Cortex.* 10:3, pp. 284–94.

Sanfey, Alan G., James K. Rilling, Jessica A. Aronson, Leigh E. Nystrom and Jonathan D. Cohen. 2003. "The Neural Basis of Economic Decision-Making in the Ultimatum Game." *Science.* 300:5626, pp. 1755–757.

Sapolsky, Robert M. 1998. *Why Zebras Don't Get Ulcers: An Updated Guide to Stress, Stress-Related Diseases, and Coping.* New York: W. H. Freeman and Company.

Schultz, Wolfram, Peter Dayan and P. Read Montague. 1997. "A Neural Substrate of Prediction and Reward." *Science.* 275:5306, pp. 1593–599.

Shallice, Tim and Paul W. Burgess. 1991. "Higher-Order Cognitive Impairments and Frontal Lobe Lesions in Man," in *Frontal Lobe Function and Dysfunction.* H. S. Levin, H. M. Eisenberg and A. L. Benton, eds. New York, Oxford University Press, pp. 125–38.

Shefrin, Hersh M. and Richard H. Thaler. 1988. "The Behavioral Life-Cycle Hypothesis." *Economic Inquiry.* 26:4, pp. 609–43.

Shiffrin, Richard M. and Walter Schneider. 1977. "Controlled and Automatic Human Information Processing: II. Perceptual Learning Automaticity, Attending and a General Theory." *Psychological Review.* 84:2, pp. 127–90.

Stanovich, Keith E. 1999. *Who is Rational? Studies of Individual Differences in Reasoning.* Mahwah, N.J.: Lawrence Erlbaum Associates.

Stuss, Donald T. and D. Frank Benson. 1986. *The Frontal Lobes.* New York: Raven Press.

Thomson, Judith J. 1986. "The Trolley Problem," in *Rights, Restitution and Risk.* Cambridge: Harvard University Press, pp. 94–116.

Tooby, John and Leda Cosmides. 1990. "The Past Explains the Present: Emotional Adaptations and the Structure of Ancestral Environments." *Ethology and Sociobiology.* 11:4-5, pp. 375–424.

Tor, D. Wager, James K. Rilling, Edward E. Smith, Alex Sokolik, Kenneth L. Casey, Richard J. Davidson, Stephen M. Kosslyn, Robert M. Rose and Jonathan D. Cohen. 2004. "Placebo-Induced Changes in fMRI in the Anticipation and Experience of Pain." *Science.* 303:5561, pp. 1162–167.

Waldman, Michael. 1994. "Systematic Errors and the Theory of Natural Selection." *American Economic Review.* 84:3, pp. 482–97.

Whalen, P. J., L. M. Shin, S. C. McInerney, H. Fischer, C. I. Wright and S. L. Rauch. 2001. "A Functional MRI Study of Human Amygdala Responses to Facial Expressions of Fear versus Anger." *Emotion.* 1:1, pp. 70–83.

Wheatley, Thalia and Jonathan Haidt. 2004. "The 'Wisdom' of Repugnance: Hypnotically Induced Disgust Makes Moral Judgments More Severe." Unpublished paper.

Wicker, Bruno, Christian Keysers, Jane Plailly, Jean Pierre Royer, Vittorio Gallese and Giacomo Rizzolatti. 2003. "Both of Us Disgusted in My Insula: The Common Neural Basis of Seeing and Feeling Disgust." *Neuron.* 40:3, pp. 655–64.

Wilson, David S. 2002. *Darwin's Cathedral: Evolution, Religion, and the Nature of Society.* Chicago: University of Chicago Press.

Wrangham, Richard W. and Dale Peterson. 1996. *Demonic Males: Apes and the Origins of Human Violence.* New York: Houghton Mifflin Company.

Wright, Robert. 1994. *The Moral Animal: Evolutionary Psychology and Everyday Life.* New York: Pantheon Books.

Yu, Angela J. and Peter Dayan. 2005. "Uncertainty, Neuromodulation, and Attention." *Neuron.* 46:4, pp. 681–92.