**WORK IN PROGRESS — NOT FOR GENERAL CIRCULATION**

# Animal Spirits: Affective and Deliberative Processes in Economic Behavior

George Loewenstein
Department of Social and Decision Sciences
Carnegie Mellon University

Ted O'Donoghue
Department of Economics
Cornell University

Draft: 4/29/2004

**Abstract**

The economic conception of human behavior assumes that a person has a single set of well-defined goals, and that the person's behavior is chosen to best achieve those goals. We develop a model in which a person's behavior is the outcome of an interaction between two systems: a *deliberative system* that assesses options with a broad, goal-based perspective, and an *affective system* that encompasses emotions and motivational drives. Our model provides a framework for understanding many departures from full rationality discussed in the behavioral-economics literature, and captures the familiar feeling of being "of two minds." By focusing on factors that moderate the relative influence of the two systems, our model generates a variety of novel testable predictions.

Acknowledgements: We thank David Laibson, Roland Bénabou, and Antonio Rangel for useful comments, and we thank Christoph Vanberg for valuable research assistance. For financial support, Loewenstein thanks the Integrated Study of the Human Dimensions of Global Change at Carnegie Mellon University (NSF grant # SBR-9521914), and O'Donoghue thanks the National Science Foundation (grant SES-0214043).

At the center of the brain lies a cluster of strange-shaped modules that together are known as the limbic system. This is the powerhouse of the brain — generator of the appetites, urges, emotions and moods that drive our behavior. Our conscious thoughts are mere moderators of the biologically necessary forces that emerge from this unconscious underworld; where thought conflicts with emotion, the latter is designed by the neural circuitry in our brains to win.

Rita Carter, Mapping the Mind. (1998:54)

## I. Introduction

The standard economic conception of human behavior assumes that a person has a single set of well-defined goals, and that the person's behavior is chosen to best achieve those goals (or at least that it is "as if" the person's behavior is determined in this way). Such an approach is intuitive, tractable, and has shed light on a wide range of economic behaviors, from mundane activities such as consumer decision making to more exotic behaviors such as those associated with drug addiction. While this approach seems to suffice in many situations, we show that a more nuanced, "two-system" perspective that takes account of interacting processes within the human brain permits us to better understand a wide range of economic phenomena that are difficult to reconcile with the standard unitary preference approach.

We develop a two-system model in which a person's behavior is the outcome of an interaction between a *deliberative system* that assesses options with a broad, goal-based perspective (roughly along the lines of the standard economic conception), and an *affective system* that encompasses emotions such as anger and fear and motivational drives such as those involving hunger and sex. Our model provides a conceptual framework for understanding many departures from full rationality discussed in the recent behavioral-economics literature. At the same time, it captures the familiar feeling of "being of two minds" — of simultaneously wishing one were behaving one way while actually behaving in a different way. By focusing on factors that moderate the relative influence of the two systems, our model also generates a number of novel testable predictions.

In Section II, we motivate our particular two-system account of behavior. Our distinction between affective and deliberative processes is roughly similar to the variety of dual-system perspectives on the human mind that have been espoused over the years by philosophers and psychologists. It is also roughly consistent with evidence from neuroscience on the different

neural processes taking place in the prefrontal cortex vs. the more primitive brain structures. We also discuss in Section II some insights that emerge from these literatures concerning interactions between the two systems. We describe, for instance, the role of environmental stimuli in activating the two systems, and how the "proximity" of a stimulus plays an especially important role in the degree to which the affective system is activated (relative to the deliberative system). We describe how affect seems to hold a kind of primacy over deliberation. Finally, we discuss evidence on the key concept of *willpower*, which we view as a resource expended by the deliberative system to exert influence over the affective system. In particular, we describe evidence on factors that tend to promote or deplete a person's ability to exert willpower.

In Section III, we develop our general model of interactions between the affective and deliberative systems. Specifically, we assume that each system has an objective function, but neither system has complete control. To formalize the interactions, we assume that the affective system has primary control of behavior, but the deliberative system can influence the affective system's choice by exerting cognitive effort, or willpower. The deliberative system then chooses which behavior to implement by trading off its objectives against the cost of exerting this willpower. Perhaps most importantly, our model endogenizes the relative influence of the two systems via factors that influence the cost of exerting willpower and factors that influence the objectives of the two systems. Our approach is similar to but different from some recent approaches in the economics literature — notably, the planner-doer approach of Shefrin and Thaler (1988) and the hot-mode/cold-mode approach of Bernheim and Rangel (2002, 2003). We discuss these relationships after laying out our model.

To make specific predictions, we must make structured assumptions about the objectives of the two systems and how the two systems respond to specific stimuli, both of which will be domain-specific. To illustrate the potential value of our framework, in Sections IV, V, and VI we apply our model to three specific domains. In Section IV, we apply our model to time preferences. The natural assumptions here are that the affective system cares primarily about short-term outcomes, whereas the deliberative system cares about both short-term and long-term outcomes. In our framework, these assumptions imply that a person will exhibit time discounting even if the deliberative system has no time preference. Moreover, if a person makes a series of interrelated choices, our framework can give rise to several forms of "self-control problems." Indeed, we describe how our framework provides a nice reinterpretation of the

differences between two recent models of self-control: hyperbolic discounting as in Laibson (1997) and O'Donoghue and Rabin (1999), and temptation utility as in Gul and Pesendorfer (2001). Finally, we describe how our model makes predictions about when people should show greater discounting, "steeper" hyperbolic discounting, or stronger temptation (dis)utility.

In Section V, we apply our model to risk preferences. The perspective we suggest is that the objectives of the deliberative system correspond roughly to expected-utility theory, whereas the affective system is more sensitive to outcomes than probabilities. This perspective can explain why people tend to exhibit an S-shaped probability-weighting function and, further, predicts *when* this function should be more or less S-shaped. We also suggest that loss aversion might be driven by the affective system, and enumerate testable hypotheses that would follow from such an account concerning when people should exhibit more or less loss aversion. Finally, we describe how risk preferences are likely to be sensitive to the "proximity" of outcomes — e.g., the vividness with which outcomes are described, or how soon outcomes will be experienced — and provide evidence supporting these predictions.

In Section VI, we apply our model to social preferences, and describe specifically how it can be applied to altruistic preferences. Here, the perspective we suggest is that the deliberative system is driven by moral and ethical principles for how one ought to behave, whereas the affective system can be driven toward behaviors at any point between the extremes of pure self-interest and extreme altruism depending on the degree of empathy that is triggered. Analogous to our treatment of risk preferences, we derive testable predictions concerning when we should expect people to exhibit more or less altruism.

We conclude in Section VII with a discussion of broader implications of our framework.

## II. Motivations

### Dual-system perspectives

Dual-system models of the human mind are ubiquitous in philosophical discussions of human behavior dating back to the ancient Greeks. In the Republic, for example, Plato contrasts the immediacy of desires as short-sighted attractions to particular classes of things, with the broader scope of reason, whose function in the human soul is to "rule with wisdom and forethought on behalf of the entire soul" (Plato, Republic 441e). More recently, in the Theory of

Moral Sentiments, Adam Smith viewed human behavior as a struggle between the "passions" and the "impartial spectator." The passions refer to immediate motivational forces and feeling states, such as hunger, thirst, anger, and sexual lust. The impartial spectator refers to the human ability to take a dispassionate view of one's own conduct — to evaluate one's own behavior as if through the eyes of another person who is unaffected by the passions. Smith viewed the ability to assume the perspective of an impartial spectator as a powerfully moderating force in human behavior, as the source of "self-denial, of self-government, of that command of the passions which subjects all the movements of our nature to what our own dignity and honour, and the propriety of our own conduct, require" (1759:26). Smith recognized, however, that such perspective-taking has constraints, and can be overcome by sufficiently intense passions:

> There are some situations which bear so hard upon human nature that the greatest degree of self-government, which can belong to so imperfect a creature as man, is not able to stifle, altogether, the voice of human weakness, or reduce the violence of the passions to that pitch of moderation, in which the impartial spectator can entirely enter into them. (1759:29)

Dual-process models are also ubiquitous in contemporary psychology (see Chaiken & Trope, 1999 for a broad introduction). Psychologists have proposed a variety of specific dichotomies — e.g., between cognition and emotion, reason and intuition, or consciousness and unconsciousness. And such models have been used to understand much more than behavior; they are also used to explain the formation of attitudes and beliefs, perceptions about others, stereotyping, and so forth. The specific dual-process model that is closest in spirit to our own was proposed by Metcalfe and Mischel (1999). In their approach, there is a "hot emotional system" that is simple, reactive, and fast, and a "cool cognitive system" that is complex, reflective, and slow (and "devoid of emotion"). The person's behavior depends on which system is dominant. Metcalfe and Mischel use this model primarily to explain the diverse results obtained in studies based on Mischel's delay-of-gratification paradigm, and in particular how different control strategies might be useful in helping the cool system to gain dominance (e.g., obscuring the reward stimulus). While our distinction between the deliberative and affective

systems is similar, it is somewhat more nuanced and concrete, and we apply it to a much broader range of behaviors.[1]

**Evidence from neuroscience on the human brain.**

Neuroscientists view the human brain as a complex organ composed of many neural systems, and view behavioral (and other) outcomes as determined from interactions between these systems. Indeed to many neuroscientists the notion of just two processes would certainly seem too few. If, however, one had to pick the most natural neurophysiological division of the brain, it would most likely be between the prefrontal cortex and the more primitive brain structures.[2]

Perhaps the main reason to focus on this particular division is evolutionary. When human evolution departed from that of apes approximately 6 million years ago, our brains were not redesigned anew. Rather, new capabilities — most importantly for our argument, the ability to deliberate about the broader consequences of our actions — were gradually added to the underlying, more primitive brain systems. These new capabilities are primarily centered in the prefrontal cortex, which is also the region of the brain that expanded most dramatically in the course of human evolution (Manuck et al 2003). As Massey (2002:15) comments:

> Emotionality clearly preceded rationality in evolutionary sequence, and as rationality developed it did not replace emotionality as a basis for human interaction. Rather, rational abilities were gradually *added* to preexisting and simultaneously developing emotional capacities. Indeed, the neural anatomy essential for full rationality—the prefrontal cortex—is a very recent evolutionary innovation, emerging only in the last 150,000 years of a 6-million-year existence, representing only about 2.5 percent of humanity's total time on earth.

[1] Our framework also bears a resemblance to Freud's (1924/1968) distinction between the id and the ego. The id, which represents biological forces and is governed by the "pleasure principle," is somewhat close to our affective system. The ego, governed by the "reality principle," is fairly close to our deliberative system.

[2] Another important division is between 'automatic' and 'controlled' processes (Camerer, Loewenstein & Prelec, 2003). Automatic processes such as vision involve massively parallel processing, are not accessible to introspection, and are not associated with a feeling of mental effort. Controlled processes, in contrast, tend to be serial, are often accessible to introspection, and are often associated with a feeling of mental effort. Our division between affect and deliberation is correlated with that between automatic and controlled processes, because affect is more closely associated with automatic processing while deliberation tends to be more controlled, but the association is far from perfect.

The more primitive brain systems have changed little over the course of human evolution and continue to play the same role that they did for our predecessors and do for other mammals (MacLean, 1990). These brain systems evolved to promote survival and reproduction. They incorporate motivational mechanisms — often operating outside of consciousness — that are designed to ensure that we eat when nutritionally deficient, take actions to maintain body temperature, have sex when the situation is propitious for reproduction, and so forth.

The unique human ability to focus on broader goals appears to reside in the prefrontal cortex. The earliest, and perhaps still the best, evidence that the prefrontal cortex plays such a role comes from studies of people with damage to the prefrontal cortex (for an overview, see Damasio 1994). In particular, patients with damage to the ventromedial section of the prefrontal cortex exhibit impaired decision-making abilities. Such people often exhibit no overt limitations in their intellectual abilities, and they are often quite able to predict and verbally describe the future consequences of different behaviors. However, they have trouble deciding on the best course of action. Moreover, while many such patients do formulate plans (or take jobs), they usually fail to implement those plans. Lhermitte (1986) found that, due to their inability to act on long-term goals, the behavior of patients with prefrontal lesions becomes largely a function of immediate contingencies of the environment, a pattern that he aptly describes as an "environmental dependency syndrome."[3]

There is a range of evidence that documents how responses to stimuli are influenced by activity in both the neocortex and lower brain structures. For instance, Joseph LeDoux and his colleagues (summarized in LeDoux, 1996) have demonstrated that both the cortex and the lower brain structures play a role in fear responses. Based on their research using rats, they discovered that there are two neural pathways from the sensory thalamus (a lower-brain structure that performs crude processing of external stimuli) to the amygdala (another lower-brain structure that plays a critical role in fear responses). One pathway goes directly from the sensory thalamus

[3] In one clever study that illustrates the role of the prefrontal cortex in deliberative behavior, Chris Frith and colleagues (see Spence and Frith 1999) scanned subjects' brains while they moved a finger they had been instructed to move in response to a noise. The brain scan revealed activation in the auditory cortex (which does crude processing of sounds) and the motor cortex (the area that controls movement). To localize where free-willed activity happens in the brain, they then added a new component to the task; instead of telling the subjects which finger to lift they left it to them to decide which one to move. With the addition of this new aspect to the task, the prefrontal cortex became activated as well.

to the amygdala and carries relatively crude information about the external stimuli. A second pathway goes first from the sensory thalamus to the neocortex and from there to the amygdala. This second pathway carries more sophisticated information about a stimulus. The fear response depends on both pathways (e.g., if one damages the relevant part of a rat's neocortex, fear responses and fear conditioning is quite different). In one study that is particularly supportive of the perspective offered here, rats were first fear-conditioned to a tone, then "deconditioned" until their overt fear response disappeared. When he then surgically cut the connection between the neocortex and the amygdala, the fear response reappeared, suggesting that the neocortex had been effectively suppressing fear reactions that remained latent in the amygdala.

A second source of evidence on the different activities of the neocortex and lower brain structures comes from split-brain patients. Such patients have had surgery to sever the nerve connections between the two hemispheres of the neocortex (to control severe epilepsy). After this surgery, the two hemispheres cannot communicate, so that if a stimulus is presented to only the right hemisphere, the left hemisphere will be unable to say what the stimulus is. For one particularly interesting patient (reported in Gazzaniga and LeDoux, 1978), while as usual the left hemisphere was unable to say what stimulus had been presented, the left hemisphere was able to accurately judge whether that stimulus was good or bad. Hence, the cognitive processing carried out in the right hemisphere (identifying the stimulus) could not be transferred to the left hemisphere, but the affective processing (valuing the stimulus as good or bad) could be transferred, presumably through the lower-brain structures.

A final piece of neuroscientific evidence comes from work by Kent Berridge (1995) on food reward. Based on studies of rats, Berridge finds two distinct reward processes that he labels "wanting" and "liking". Wanting food corresponds to a disposition to eat or an appetite; liking food corresponds to sensory pleasure or palatability. Most relevant for our purposes, he finds that wanting and liking are mediated by different neural systems. Hence, eating behavior will be an outcome of multiple reward systems operating simultaneously in the brain.

**Affective and deliberative processing.**

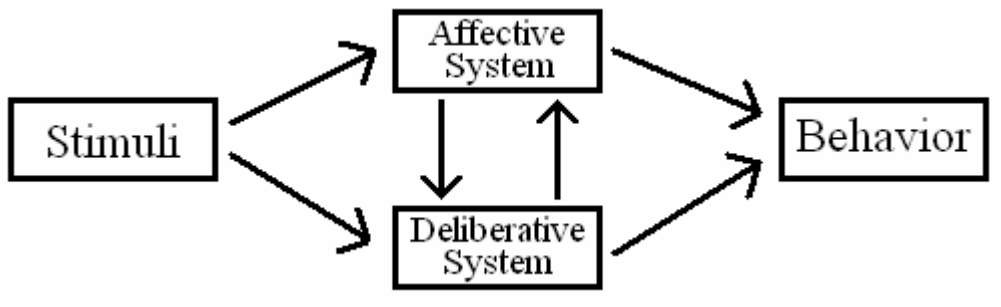
In this paper, we investigate a two-system approach in which there is an *affective system* that is primarily driven by motivational mechanisms and a *deliberative system* that takes into account broader goals. Our decomposition of the brain is motivated by and roughly similar to many of those described above — Adam Smith's passions vs. impartial spectator; Metcalfe and Mischel's hot emotional system vs. cool cognitive system; the neurophysiological division between the more primitive brain structures and the prefrontal cortex; and the more generic distinctions between emotion vs. cognition and unconscious vs. conscious processes. However, it does not exactly correspond to any of these approaches.

Our use of the term "affect" differs from many lay definitions, which tend to focus on the subjective feeling states associated with emotions. In our usage, the central feature of affect is its role in human motivation. All affects carry "action tendencies" (Frijda, 1986) — e.g., anger motivates us to aggress, pain to take steps to ease the pain, and fear to escape (or in some cases to freeze). As the psychologist Robert Zajonc (1998) expresses it, affective processes are those that address "go/no-go" questions — that motivate approach or avoidance behavior. But affect, as we use the term, embodies not only emotions such as anger, fear, and jealousy, but also drive states such as hunger, thirst, and sexual desire, and motivational states such as physical pain, discomfort (e.g., nausea), and drug craving. Buck (1984) refers to these latter influences as "biological affects," which he distinguishes from the more traditional "social affects."[4] Moreover, many of these affective processes occur below the threshold of conscious awareness, and hence are not experienced as an "emotion" or "affect" in the lay sense (LeDoux, 1996).

Figure 1 graphically represents our two-system approach. The main feature is that behavior is the outcome of an interaction between distinct affective and deliberative systems. But Figure 1 also reflects a number of other properties.

[4] Although emotions such as anger and fear might seem qualitatively different than the biological affects, they have more in common that might be supposed. Thus, for example, a recent study showed that hurt feelings activated the same brain regions as would broken bones or other physical injuries (Eisenberger et al 2003). The researchers scanned the brains of subjects using fMRI (functional magnetic resonance imaging) as they played a video game designed to produce a feeling of social rejection. The social snub triggered nerve activity in a part of the brain called the anterior cingulate cortex, which also processes physical pain.

# Figure 1:




**Environmental stimuli and the role of proximity.**

Figure 1 reflects that both brain systems can be activated by environmental stimuli, which might be external — e.g., the sight of food — or internal — e.g., low blood sugar. An environmental stimulus might activate the affective system, as when the sight of a snake induces a motivation to flee, or it might activate the deliberative system, as when the sight of one's wedding ring brings to mind one's spousal responsibilities. In fact, many stimuli activate both systems. Such "bilateral" influences are probably most often complementary and mutually reinforcing, as when the sight of food activates the affective state of hunger and the cognitive state of "it's time to eat." In some instances, however, a stimulus can activate the two systems in competing ways, as when the sight of food activates the affective state of hunger and the cognitive state of "I'm on a diet." It is the latter case for which a dual-process perspective is most useful.

The specific role of stimuli differs across domains of behavior and across individuals. Nevertheless there is one important general effect: The "proximity" of a stimulus plays an especially important role in the degree to which the affective system is activated (relative to the deliberative system). Proximity can be defined on many dimensions — geographic, temporal, visual, social, and so forth. Thus, for example, a tasty morsel is more likely to evoke hunger to the extent that it is nearby, immediately attainable, visible, or being consumed by someone else (in close proximity). Likewise, a person who makes you angry is more likely to evoke anger to the extent that he is geographically close (or likely to be soon) or visible.

Perhaps the best evidence on the role of proximity comes from a series of classic studies conducted by Walter Mischel and colleagues (see for instance Mischel, Ebbesen, and Zeiss 1972; Mischel, Shoda, and Rodriguez 1989; and Mischel, Ayduk, and Mendoza-Denton 2003). Young subjects (ages 4-13) were instructed by an experimenter that they could have a snack immediately or they could have a larger snack if they waited for the experimenter to return. The experimenter then measured how long the subjects were willing and able to wait for the larger snack (with a cap of fifteen minutes). In a baseline treatment, children had the larger delayed snack positioned in front of them as they waited for the experimenter. Relative to this baseline treatment, subjects were able to delay significantly longer when the larger snack was not present, or even when the larger snack was present but covered. Similarly, even when the larger snack is present and uncovered, subjects were able to delay longer if their attention was diverted away from the snack so that they spent less time looking at the larger snack.

The relationships between stimuli and the activation of the affective system often lie outside of conscious awareness. For instance, people can become fear conditioned to subliminal stimuli — e.g., images that are flashed so quickly that the individual is unaware that she has seen anything — and fearful responses can occur in response to visual stimuli not only in the region of awareness but also in the peripheral visual field. Indeed, such subliminal fear conditioning may actually be more powerful exactly because the conscious, deliberative system is unaware of it and hence is less likely to engage in efforts to override it (Anderson et al 2003).

**Bidirectional interactions between affect and deliberation**

Figure 1 also reflects that the affective and deliberative systems interact with each other. Neuroscientists have identified neural connections running in both directions between the more primitive brain systems and the prefrontal cortex (neural pathways run in specific directions). These connections are suggestive of the types of interactions that occur.

The existence of neural connections from the more primitive brain systems to the prefrontal cortex suggests that the affective system can influence the deliberative system. Such input from the affective system may be required for sound deliberative thinking. For instance, input from the affective system may help focus the deliberative system on relevant bodily needs. For example, when the affective system transmits hunger up to the deliberative system, it helps focus the deliberative system on the decision whether to eat. There is, in fact, ample evidence

that affect serves as an essential input into decision-making. One set of studies shows detrimental effects of blocking decision-makers' affective reactions to alternatives (Wilson and Schooler 1991; Wilson et al 1993). Other studies show that damage to the ventromedial prefrontal cortex — the part of the prefrontal cortex that appears to provide the main link between affective and deliberative processing — compromises people's ability to decide on a best course of action, or even to decide at all (ventromedial prefrontal patients often perseverate for hours when it comes to even the most trivial decisions).

At the same time, there is also a large body of research dealing with what psychologists call "motivational" biases on judgment that documents the diverse ways that affect biases cognitive deliberations (Kunda, 1990). For instance, cigarette craving is a neurophysiological brain state that directly motivates smoking but also influences smokers' perceptions of the costs and benefits of smoking (Sayette et al 2001). Because motivational biases undermine the deliberative system's efforts to impartially assess the optimal course of behavior, they provide a second, indirect pathway by which affect influences behavior.

The existence of neural connections in the opposite direction, from the prefrontal cortex to the more primitive brain systems, suggests that the deliberative system can also influence the affective system. Sometimes, this influence might take the form of deliberative thoughts activating emotions in the affective system (as when one fantasizes). More important for economic behavior, however, the deliberative system can attempt to control or override the motivations in the affective system. However, while some degree of emotional self-control is often possible, such attempts to control emotions tire people out emotionally and physically and, in fact, often backfire, exacerbating the very emotion that one is trying to suppress (see for instance Ochsner and Gross 2004; Wegner 1992; Smart and Wegner 1996).

**The primacy of affect.**

Although interactions run in both directions, affect seems to hold a kind of primacy over deliberation. As Adam Smith argued early on, if the deliberative system does not get activated — if it does not attend to a particular choice situation — then behavior will be driven entirely by affective motivations. (Anyone who has ever been put in front of a table of snacks and who has found himself eating without having made any deliberation can appreciate this notion.) Even when both systems are active, affect seems to have greater sway. As Ledoux (1996:19) notes,

"while conscious control over emotions is weak, emotions can flood consciousness. This is so because the wiring of the brain at this point in our evolutionary history is such that connections from the emotional systems to the cognitive systems are stronger than connections from the cognitive systems to the emotional systems."

Affect not only holds greater sway over deliberation than vice-versa, but affective reactions tend to occur first, temporally, with deliberations typically playing a secondary, corrective, role. For instance, in Joseph LeDoux's work on fear responses in rats (discussed above), in addition to discovering the direct and indirect pathways from the sensory thalamus to the amygdala, they also discovered that the direct pathway is about twice as fast as the indirect pathway. As a result, rats can have an affective reaction to a stimulus before their cortex has had the chance to perform more refined processing of the stimulus. Such immediate affective responses provide organisms with a fast but crude assessment of the behavioral options they face which makes it possible to take rapid action. To use LeDoux's example, it is useful to have an immediate defensive reaction to a curved object rather than wait for the cortex to decide whether that object is a coiled snake or a curved stick.

The same pattern can be seen in humans. In a series of seminal papers with titles such as "Feeling and Thinking: Preferences Need No Inferences" (1980), and "On the Primacy of Affect" (1984), Zajonc presented the results of studies which showed that people can often identify their affective reaction to something — whether they like it or not — more rapidly than they can even say what it is, and that their memory for affective reactions can be dissociated from their memory for details of a situation, with the former often being better. People often remember whether we liked or disliked a particular person, book, or movie, without being able to remember any details other than our affective reaction (Bargh, 1984). Similarly, Gilbert and Gill (2000) propose that people are "momentary realists" who initially trust their emotional reactions and only correct them through a comparatively laborious and time-consuming cognitive process. Thus, if the car behind you honks after the light turns green, you are likely to respond with immediate anger, followed, perhaps, by the recognition that if you had been behind someone who was, like you, spaced out at the wheel while eating a sandwich and talking on the cell phone, you might have reacted similarly. As Adam Smith (1759:136) expressed it, "We are angry, for a moment, even at the stone that hurts us. A child beats it, a dog barks at it, a choleric

man is apt to curse it. The least reflection, indeed, corrects this sentiment, and we soon become sensible, that what has no feeling is a very improper object of revenge."

**Willpower.**

Although the affective system seems to hold a kind of primacy, the deliberative system can often override affective responses (at least partially). Extending LeDoux's snake example, even if the cortex identifies the object as a snake, and thus justifying the defensive affective reaction, the cortex can still "will" the person to walk by the snake. However, research by Baumeister and colleagues (for a summary see Baumeister and Vohs, 2003) suggests that attempts by the deliberative system to over-ride affective motivations require an inner exertion of effort, often referred to as "willpower." This research supports many common intuitions about willpower. In particular, it shows that, much like the energy exerted by muscles, willpower is in limited supply (at least in the short term). Baumeister's basic willpower paradigm involves having subjects carry out two successive, unrelated tasks that both (arguably) require willpower, and comparing the behavior on the second task to a control group which had not performed the first task. The general finding is that exerting willpower in one situation tends to undermine people's propensity to use it in a subsequent situation. Thus, in one study, subjects who sat in front of a bowl of cookies without partaking gave up trying to solve a difficult problem more quickly than did subjects who were first tempted by the cookies.

One line of research that is especially important to our argument (summarized in Baumeister and Vohs, 2003) shows that simply making decisions can undermine willpower. In this research, which conformed to the basic paradigm just described, some subjects were asked to make a long series of choices between products while other subjects were simply asked to report on their usage of the same products. Afterward, in an ostensibly new study administered by a new experimenter, they were asked to consume a bad-tasting beverage. Subjects who had made many choices drank a significantly smaller amount of the beverage than did those in the control group.

Decision-making probably has this effect because deliberating involves the prefrontal cortex, which is the same part of the brain involved in self-regulation. It should not be surprising, then, that other cognitive tasks that involve capacities centered in the prefrontal cortex have a similar effect. Another function served by the prefrontal cortex is what is called

"working memory" — the ability to hold small amounts of information, such as a phone number, in mind for short periods of time. Research has shown that having subjects perform such tasks — an intervention labeled "cognitive load" by psychologists — undermines efforts at self-control. In one innovative study, Shiv and Fedorikhin (1999) had half of the subjects memorize a 7-digit number (high cognitive load) and others memorize a 2-digit number (low cognitive load). Subjects in both groups were instructed to walk to another room in the building where they were to report the number they had memorized. On their way, they encountered a table at which they were presented with a choice between a highly caloric slice of cake and a bowl of fruit-salad. The researchers predicted that high cognitive load would undermine self-control leading to choice of the cake, and this is what they found; 59% chose the cake in the high-load condition, but only 37% in the low-load condition.

Another variable which seems to undermine willpower is stress. Several studies have shown, for example, that stress often leads to relapse by abstinent addicts. In one of the most carefully crafted study of this type, Shiffman and Waters (2004) had smokers who had quit carry palm pilots around which beeped at random intervals, then asked them questions. They were also instructed to enter information into the palm pilot if they smoked a cigarette. One of the most important findings from this study, which reinforces findings from numerous other studies employing different methods, was that relapse was often immediately preceded by stressful events.

## III. A Two-System Model of Behavior

In this section, we develop our two-system model of behavior. For expositional clarity, we lay out our model in a static setting in which a person makes a choice at a single point in time. Some additional issues arise in dynamic settings in which an individual makes (interrelated) choices at multiple points in time. We describe these issues when we apply the model to time preferences in Section IV.

Suppose a person must choose an option $x$ out of some choice set $X$. When making this choice, the person is exposed to a vector of environmental stimuli $s$. These stimuli can activate affective states in the brain — e.g., anger, hunger, and fear. We use $a(s)$ to represent the vector of affective states induced by a vector of stimuli $s$. These stimuli can also activate cognitive states, by which we mean memories of broader goals — e.g., "I'm on a diet," "I don't engage in

behavior *x*," and "I'm married." We use the vector *c(s)* to represent the vector of cognitive states induced by a vector of stimuli *s*.

The *affective system* is motivated to engage in certain behaviors, and is primarily driven by affective states that are currently activated. When hungry, for instance, the affective system is motivated to eat. We capture these motivations with a motivational function *M(x,a)*. If the affective system alone were completely in charge of behavior, and if the current vector of affective states were *a*, the affective system would "choose" $x^A \equiv \arg\max_{x \in X} M(x,a)$, which we refer to as the *affective optimum*.

The deliberative system evaluates behavior with a broader and more goal-oriented perspective. We capture the desirability of actions as perceived by the deliberative system by a utility function *U(x,c,a)*. This formulation says that the deliberative system is influenced by both cognitive and affective states: the cognitive states currently in one's mind affect the broader goals, while at the same time emotions transmitted up from the affective system can influence deliberative thinking. If the deliberative system alone were completely in charge of behavior, and if the current vectors of cognitive and affective states were *c* and *a*, the deliberative system would choose $x^D \equiv \arg\max_{x \in X} U(x,c,a)$, which we refer to as the *deliberative optimum*.

The person's behavior is determined by an interaction between the two systems. To formalize the interactions, we assume that the affective system has primary control of behavior, but the deliberative system can influence the affective system's choice by exerting cognitive effort, or willpower. We capture this cognitive effort by assuming that, to induce some behavior $x \neq x^M$, the deliberative system must exert an effort cost, in utility units, of $h(W,\sigma)*[M(x^A,a) - M(x,a)]$. This formulation says that the further the deliberative system moves behavior away from the affective optimum, the more willpower is required. The scaling factor *h(W,σ)* represents the current cost to the deliberative system of mobilizing willpower — i.e., the higher is *h(W,σ)*, the larger is the cognitive effort required to induce a given deviation from the affective optimum (we assume $h(W,\sigma) > 0$ for all *W* and $\sigma$). This cost of willpower will depend on the person's current willpower strength, which we denote by *W*, and on other factors that undermine

or bolster the deliberative system, which we denote by $\sigma$. We describe these factors in more detail below.[5]

In "deciding" how much to influence the affective system, the deliberative system trades off the desirability of actions — as reflected by its utility function — against the willpower effort required to implement actions. Specifically, if the current vector of stimuli is *s*, the deliberative system will choose the action $x \in X$ that maximizes:

$$V(x,s) \equiv U(x,c(s),a(s)) - h(W,\sigma)*\left[M(x^A,a(s)) - M(x,a(s))\right]$$

Our formulation is motivated by the evidence on the primacy of affect, and on how the deliberative system must expend willpower to influence behavior. Hence, we have built a kind of principal-agent model in which the deliberative system (the principal) chooses which behavior to induce subject to the constraint that it must incur a cost (exerting willpower) to get the affective system (the agent) to carry out its chosen behavior.[6] We note, however, that our model of choice is also more directly consistent with the conceptualization in Figure 1 wherein neither the deliberative system nor the affective system has primary control of behavior. Specifically, because $U(x^D,c(s),a(s))$ is unaffected by the person's chosen action *x*, choosing *x* to maximize $V(x,s)$ is equivalent to choosing *x* to minimize:

$$\left[U(x^D,c(s),a(s)) - U(x,c(s),a(s))\right] + h(W,\sigma)*\left[M(x^A,a(s)) - M(x,a(s))\right].$$

Hence, our model is equivalent to thinking of behavior as coming from the minimization of a weighted sum of two costs: a cost to the deliberative system from not getting its optimum $x^D$, and a cost to the affective system from not getting its optimum $x^A$. In this reinterpretation, the scaling factor $h(W,\sigma)$ captures the relative weights of the two systems. As $h(W,\sigma)$ approaches zero, the deliberative system is in complete control of behavior, whereas as $h(W,\sigma)$ gets very large, the

[5] The linear formalization is obviously a simplification, but it is sufficient for the points we make in this paper.

[6] In line with our formulation, the prefrontal cortex is sometimes referred to as performing an "executive function" (Shallice and Burgess 1998): much as a chief executive needs to work through the existing structure and culture of the firm to implement her plans, the deliberative system has to work through the affective system to influence behavior.

affective system is in complete control of behavior. More generally, the model will generate behavior that is somewhere in between the deliberative optimum and the affective optimum (either $x^D \geq x \geq x^A$ or $x^A \geq x \geq x^D$); exactly where behavior falls will depend on the relative strength of the two systems as captured in the cost of willpower $h(W,\sigma)$.[7]

To make specific predictions, we must put more structure on the objectives of the two systems, and how these objectives respond to specific stimuli, both of which will depend on the domain of behavior under consideration. In subsequent sections, we apply our model to specific domains of behavior and derive more detailed implications. In the remainder of this section, we lay out some general implications of our model.

The most basic contribution of our model is to provide a conceptual framework for understanding some of the many departures from the standard economic model that have been identified in the behavioral-economics literature. In particular, the standard economic model can roughly be interpreted as the special case of our model in which the deliberative system is in full control, and, under this interpretation, deviations from the standard model are driven by motivations coming from the affective system.

A second general contribution is that our model captures the familiar experience of simultaneously "being of two minds" — the experience of simultaneously wishing one were doing one thing while actually doing another. Specifically, a natural interpretation of one's "wishes" is that they correspond to the desirability of actions as perceived by the deliberative system (and reflected in the utility function $U$). Hence, whenever the affective system pushes behavior away from the deliberative optimum, the person will "wish" she were behaving differently from what she is doing.

A third general contribution is that our model makes predictions about *when* the affective system is likely to have a larger influence on decisions. One source of such predictions comes from systematic effects of stimuli on cognitive and affective states — i.e., effects operating through $c(s)$ and $a(s)$. As discussed in Section II, while the effects of environmental stimuli are mostly domain-specific, one basic effect seems to operate across most domains: The proximity

[7] There is of course yet another interpretation (that we don't particularly like) wherein the deliberative system has primary control of behavior but the affective system influences the deliberative system by inflicting "pain" whenever behavior deviates from the affective optimum. Under this interpretation, $h(W,\sigma)*[M(x^A,v(s))–M(x,v(s))]$ represents the pain imposed by the affective system, and $h(W,\sigma)$ reflects how susceptible the deliberative system is to such pain.

of stimuli plays a much larger role for affective states than for cognitive states. As a result, proximity will often produce predictable divergences between the two system's inclinations.

A second — and more novel — source of such predictions comes from systematic variations in the cost of willpower $h(W,\sigma)$. Our formulation above assumes that the cost of mobilizing willpower depends on the person's current willpower strength, which we denote by $W$, and on other factors that undermine or bolster the deliberative system, which we denote by $\sigma$. We now describe each of these in more detail.[8]

Research by Baumeister and colleagues (discussed in Section II) suggests that it requires cognitive effort for the deliberative system to influence the affective system, and moreover that this resource is in limited supply. The willpower strength variable $W$ is meant to capture the current stock of this resource; we assume that $h$ is decreasing in $W$, so that, as one's willpower strength is depleted, the deliberative system finds it more difficult (more costly) to influence the affective system. It is also useful to provide a simple formalization of the dynamics of willpower. Let $W_t$ denote the person's willpower strength in period $t$, and let $w_t \equiv M(x_t^A,a_t) - M(x_t,a_t)$ denote the amount of willpower exerted in period $t$. The person's willpower strength in period $t+1$ will depend on a combination of her willpower strength in period $t$ and the amount of willpower she exerted in period $t$, or $W_{t+1} = f(w_t,W_t)$. To reflect that willpower strength is a resource that is used over time, we assume that $W_{t+1}$ is decreasing in $w_t$ and increasing in $W_t$. In words, the more willpower she used last period the smaller will be her current willpower strength, and the more willpower strength she had last period the larger will be her current willpower strength. Our model predicts, therefore, that if a person has exerted willpower in the recent past, her current behavior will be further from the deliberative optimum. Similarly, our model predicts that if a person has been forced to repeatedly make choices and therefore repeatedly expend willpower, her current behavior will be further from the deliberative optimum. This latter prediction implies that the frequency of choice can play a critical role in people's behavior. While not important for our analysis in this paper, we suggest two further assumptions with regard to the dynamics of willpower. First, it seems natural to assume that willpower is replenished over time — e.g., that there is a replenishment rate $r > 0$ such that $f(w,W) > W$ when

[8] Roughly speaking, the latter corresponds to idiosyncratic factors that influence the cost of willpower, whereas the former corresponds to the effects of one's own past experiences.

$w < r$ and $f(w,W) < W$ when $w > r$. Second, it seems natural to assume that there is an upper bound $\overline{W}$ on the stock of the resource, or $f(0,W) < \overline{W}$ for all $W$.[9]

In addition to willpower strength, there are other factors that can undermine or bolster the deliberative system. Two generic factors that undermine willpower are "stress" and "cognitive load." When people are either experiencing stress or high demands on the limited capacity of the deliberative system, the deliberative system is undermined, making it more difficult (costly) to exert willpower. Hence, increased stress or higher cognitive load both move behavior further from the deliberative optimum.

We shall return to these general predictions with regard to willpower strength, stress, and cognitive load when we examine specific domains of behavior. As we shall see, we can use these general predictions in two ways. First, in domains where we have some notion as to what the preferences of the two systems are, these general predictions turn into more specific predictions. For instance, in the realm of intertemporal choice, where it seems reasonable to believe that short-term discounting is mainly the product of the affective system, our model predicts that depleting willpower or putting people under cognitive load should lead to increased discount rates. Second, in domains where we are uncertain as to what the preferences of the two systems are, these general predictions tell us how to discover these preferences (or test conjectures about these preferences), namely, by seeing which way behavior moves when we deplete willpower, increase stress, or increase cognitive load. For instance, in the realm of altruism, where it is unclear exactly how the two systems care about other people, we can learn about the two systems' objectives by examining how behavior changes when we deplete willpower or place people under cognitive load.

A final contribution of our model is that it helps to organize — but unfortunately not resolve — the welfare debate arising out of the recent behavioral-economics literature. The standard revealed-preference approach to welfare analysis used by economists assumes *a priori* that whatever a person does must correspond to what is in her own best interest. A major theme in the behavioral-economics literature, in contrast, is that people may not behave in their own best interest. Indeed, Kahneman (1994) suggests that it might be fruitful to think of there being

[9] A functional form that satisfies these assumptions is

$$f(w,W) = \begin{cases} W - \left(1 - e^{-\alpha(w-r)}\right)W & \text{if } w \geq r \\ W + \left(1 - e^{-\alpha(r-w)/\left(\overline{W}-W\right)}\right)\left(\overline{W} - W\right) & \text{if } w \leq r \end{cases}$$

one utility function that rationalizes behavior ("decision utility") and another utility function that captures welfare ("experienced utility"). However, once one relaxes the revealed-preference paradigm, it is important to have some principled way to decide what is an appropriate measure of welfare, and there has been much debate on this point in the literature. To the extent that our model provides a conceptual framework for departures from the standard economic model, it also provides some guidelines for how to think about welfare analysis.

Our model suggests two natural candidates for measuring welfare: (i) the deliberative system's utility function *U*, which reflects the desirability of actions as perceived by the deliberative system; and (ii) the deliberative system's objective function *V*, which reflects both the desirability of actions as perceived by the deliberative system and the willpower effort required to implement actions. (A third possibility is the affective system's motivational function *M*, but we think everyone would agree that this is inappropriate.) Unfortunately, neither candidate seems unambiguously better than the other.

Perhaps the main argument in favor of using the deliberative system's utility function *U* as the welfare criterion is that it represents how people would "like" to behave, both from a removed perspective and even in the moment. But there are some problems with *U* as a welfare criterion. One major problem arises from the fact that affective states can influence the broader goals of the deliberative system — as reflected by the fact that affective states are an argument in the function *U* — which means that the deliberative system may not have a stable set of broader goals. In other words, if one were to elicit from a person how she would like to behave, the answer is likely not invariant to the current activity in the affective system. A possible response is that the proper way to elicit broader goals is to first put people in an affectively neutral state. But even then, to the extent that deliberative system needs affective inputs to evaluate different options (as suggested by Damasio's research on patients with damage to the prefrontal cortex), we still might not get a "true" measure of the person's broader goals.[10] Moreover, a large body of research on what Loewenstein calls "hot-cold empathy gaps" (e.g., van Boven and Loewenstein 2003; Loewenstein and Angner, 2002) suggests that people tend to give little if any weight to affective states they are not currently experiencing — even though such affective states confer real utility and disutility and hence should normatively be taken into account. It is easy to

[10] In Berridge's (1995) research on multiple reward systems in the brain (on "wanting" vs. "liking"), he argues that people are most likely not consciously aware of what generates pleasure. If so, it seems we ought to be even less willing to take only the deliberative system into account when measuring welfare.

commit to a diet when one is not currently hungry or to decide to go cold turkey right after satiating oneself on a drug, but it is likely that resolutions of this type pay insufficient heed to the miseries that would actually be involved in implementing the decision.

A second major problem arises from the failure of this approach to take any account of willpower effort. Even holding constant behavioral outcomes, if implementing those outcomes required willpower effort, and if that effort was unpleasant, then it seems inappropriate to ignore that unpleasantness in welfare analysis. Put more concretely, it would seem that a policy that did not affect behavioral outcomes but dramatically reduced the willpower effort required to implement those outcomes would make people much better off, but we would conclude otherwise if we use $U$ as the welfare criterion.

So should we instead use the deliberative system's objective function $V$ as the welfare criterion? Doing so would be more in line with the standard revealed-preference approach to welfare — because $V$ represents the "preferences" that rationalize the person's behavior. The major problem with this approach, however, is that it corresponds to a belief that actual behavior maximizes welfare, and there are reasons to believe that actual behavior often reflects an excessive influence of affect. One reason to be wary is that affective states are relatively transient, and, as demonstrated by the research on incidental affect, can often influence behavior even when they are manifestly irrelevant to the decision at hand. An even bigger reason to be wary is that affective states can often be easily manipulated for strategic purposes that promote the interests of the manipulator to the detriment of the interests of the decision-maker. Finally, for intertemporal decisions there is evidence that people underappreciate the effects of future affect, as we discuss in Section IV.

Hence, neither candidate seems superior. Ideally we would like something in-between that recognizes the value of policies that reduce the willpower required to implement outcomes, but also something that takes limited account of affective states, particularly those states that are transient, easily manipulated, and tangential to decisions. At this point, there is no clear welfare measure that has these properties.

Before applying our model to specific domains, we take a moment to reflect on how our approach relates to some other approaches in the economics literature. An early two-system approach in the economics literature is the planner-doer model of Shefrin and Thaler (1988). In their approach, a farsighted "planner" who maximizes long-run utility interacts with a series of

myopic "doers" who maximize short-run satisfaction. Shefrin and Thaler emphasize the devices that the planner uses to influence the doers. In our terms, one can roughly interpret their model as the planner being the deliberative system and the doer being the affective system. However, there are several important differences. First, in their approach, at any point in time, either the planner or the doer is in complete control, whereas in our approach the deliberative and affective systems are both always at work. Second, in their model it is never really specified when the planner vs. doer is in charge; our approach makes predictions about the relative influence of the deliberative and affective systems at different points in time.

A more recent two-system approach in the economics literature comes from Bernheim and Rangel (2002, 2003). In their model, the brain can operate in one of two modes, a "cold mode" or a "hot mode". In the cold mode, the person makes sound, deliberative decisions with a broad, long-term perspective. In the hot mode, the person's decision-making is influenced by emotions and motivational drives — that is, by affect. Which mode is triggered depends (stochastically) on environmental conditions, which in turn might depend on past behavior (e.g., if you choose to go to a party tonight rather than stay home, you increase the likelihood of experiencing a craving for alcohol tonight). To simplify their analysis, Bernheim and Rangel do not explicitly model the source of behavior in the hot mode; rather, they assume that it follows some simple rule (e.g., consume the addictive product in their model of addiction, or consume a proportion of wealth in their model of savings). Their primary focus is on behavior in the cold mode, where they assume that the person performs the usual maximization of discounted expected utility but accounts for the possibility of and outcomes in future hot modes. Finally, in order to investigate policy issues, they need a measure of welfare, and they argue that the person's discounted expected utility — that is, her cold-mode preferences — is the natural measure.

Once again, there are some important differences between their approach and ours. First, their approach is, in a sense, the special case of our approach in which the cost of willpower $h(W,\sigma)$ (stochastically) flips back and forth between zero and infinity — so that the person alternates between the deliberative system having complete control and the affective system having complete control. Although there may be situations in which one system or the other is entirely in control — in particular, people may sometimes go on affective autopilot — we believe that in the most common and interesting cases, both systems are activated

simultaneously. In this sense, our approach is a generalization that permits both systems to influence behavior at the same time, and permits a broader set of possibilities for the relative influence of the two systems. Second, whereas they assume mechanistic behavior in the hot mode, we investigate directly the source of hot-mode behavior. Finally, by taking the person's discounted expected utility to be the natural measure for welfare, they are taking the stance that, from a normative perspective, the deliberative system's "utility" is all that matters. As discussed above, we see reasons to be cautious about such claims.[11]

Finally, our approach is also related to economic models that incorporate affective or visceral influences on decision-making (e.g., Laibson, 2001). Such models also assume that environmental stimuli — both external and internal — can trigger affective states which in turn influence a person's preferences. However, such models remain within the standard conception of the brain as a single, coherent decision-maker.

## IV. Time Preference

In this section, we describe the most obvious and concrete application of our model — to intertemporal choices — decisions that involve tradeoffs between current and future costs and benefits. People are often of two minds when it comes to intertemporal choice; people are often powerfully motivated to take myopic actions, such as eating highly caloric foods, imbibing addictive drugs, eschewing contraception, "flaming" on email, and so on, while recognizing simultaneously that these activities are not in their self-interest. As Adam Smith (1759:227) wrote, seemingly referring to an act of sexual misconduct,

> At the very time of acting, at the moment in which passion mounts the highest, he hesitates and trembles at the thought of what he is about to do: he is secretly conscious to himself that he is breaking through those measures of conduct which, in all his cool hours, he had resolved never to infringe, which he had never seen infringed by others without the highest disapprobation, and the infringement of which, his own mind forebodes, must soon render him the object of the same disagreeable sentiments.

[11] In fact, there is a second assumption required to justify their welfare standard, namely that the deliberative system is not concerned with any willpower effort exerted in any future periods. We address discuss this point more directly in Section IV.

In the realm of time preference, there is a natural starting point for what drives the two systems: The affective system is driven primarily by short-term payoffs, whereas the deliberative system cares about both short-term and longer-term payoffs.[12] In this section, we investigate the implications of our model given these assumptions.

Consider first static intertemporal choices in which people make once-and-for-all intertemporal trade-offs. Suppose a person chooses a current action $x$ that generates an immediate payoff $z_1(x)$ and a future payoff $z_2(x)$. The myopic affective system cares only about the immediate payoff, and so the affective system's motivational function is $M(x) = z_1(x)$. The more far-sighted deliberative system values both payoffs, and so the deliberative system's utility function is $U(x) = z_1(x) + z_2(x)$.[13] Hence, the person will choose her behavior $x$ to maximize:

$$V(x) = [z_1(x) + z_2(x)] - h*[z_1(x^A) - z_1(x)].$$

Since the affective optimum $x^A$ is exogenous to the person's choice, this is equivalent to maximizing:

$$\tilde{V}(x) = z_1(x) + [1/(1+h)]*z_2(x).$$

Because $1/(1+h) < 1$, this two-period example shows that our model gives rise to "discounting" without assuming that the deliberative system has a literal time preference — that is, even though the deliberative system may weigh different time periods relatively evenly, the affective system's focus on near-term payoffs will lead to behavior that puts higher weight on near-term payoffs than on future payoffs.

Our model also makes predictions for how the various factors that influence the relative strength of the two systems should impact the discount rates revealed by simple intertemporal

[12] Emotions sometimes drive far-sighted behavior, as when fear and anxiety cause people to "save for a rainy day" (see, e.g., Loewenstein, 1987). However, these far-sighted emotions can be interpreted as instances of the deliberative system's attempts to influence behavior by manipulating affect in the affective system. Indeed, there is little evidence that animals other than humans experience such far-sighted emotions, and humans are qualitatively different from other animals in the length of their time perspective.

[13] For notational simplicity, we suppress the arguments for affective and cognitive states in $M$ and $U$ and the arguments for willpower strength and cognitive load in $h$ whenever they are not crucial to our analysis.

trade-offs — particularly the *x* now vs. *y* in the future trade-offs that are often studied. For instance, our model predicts that if we deplete willpower or put people under cognitive load, then their elicited discount rates should be larger. Indeed, the Shiv and Fedorikhin (1999) study reported in Section II seems consistent with this prediction. Our model also predicts that the proximity of immediate outcomes should play a large role in elicited discount rates. Thus, for example, the extent that an immediate reward can be seen or smelled (assuming that the appearance and smell are attractive) will affect the magnitude of discount rates that people's behavior reveals, which is consistent with the research by Mischel and colleagues described in Section II. Finally, our model predicts stimulus-specific discounting. The sight of food might lead to increased discounting for food but not for sex, while the sight of an attractive potential sexual partner might lead to increased discounting for sex but not for food.

Consider next dynamic intertemporal choices in which people make repeated (and interrelated) intertemporal trade-offs over time. Specifically, suppose that a person chooses an action $x_1$ in period 1 and an action $x_2$ in period 2, and that these actions generate a period-1 payoff $z_1(x_1)$, a period-2 payoff $z_2(x_1,x_2)$, and a future payoff $z_3(x_1,x_2)$. For simplicity, our analysis assumes that there is no mechanism available for committing to future behavior.[14]

Because the period-2 perspective is much like the static case, let's assume that behavior in period 2 is determined as above. In other words, period-2 behavior $x_2$ will maximize

$$V^2(x_1,x_2) = [z_2(x_1,x_2) + z_3(x_1,x_2)] - h*[z_2(x_1,x_2^A) - z_2(x_1,x_2)]$$

or equivalently

$$\tilde{V}^2(x_1,x_2) = z_2(x_1,x_2) + [1/(1+h)]* z_3(x_1,x_2).$$

This problem will generate a period-2 behavior that is a function of the already taken (and therefore fixed) period-1 behavior.

To analyze the period-1 perspective, we must address some additional issues. The most interesting and novel issue is the question of how the deliberative system cares about future payoffs, and in particular whether the deliberative system incorporates future willpower effort into its decision. To make this issue precise, let's again assume that the affective system cares only about the immediate payoff, whereas the deliberative system cares about the payoff in all

[14] If in period 1 the person fully commits to a path of behavior, then it is essentially a static intertemporal choice. Similarly, if there is no link between the two decisions — if the desirability of current actions is independent of actions in other periods (as when $z_2$ and $z_3$ are both additively separable in $x_1$ and $x_2$) — then it is essentially a sequence of static intertemporal choices.

three periods — so the desirability of actions as perceived by the deliberative system is $z_1(x_1) + z_2(x_1,x_2) + z_3(x_1,x_2) \equiv U^*(x_1,x_2)$. The issue is whether the deliberative system cares only about the desirability of actions, or whether it also incorporates the willpower effort it expects to exert in period 2.

Consider first the extreme in which the deliberative system is not at all concerned with future willpower effort, and so its utility function is merely $U(x_1,x_2) = U^*(x_1,x_2)$. In this case, the deliberative system will choose period-1 behavior to maximize

$$V^1(x_1,x_2) = [z_1(x) + z_2(x_1,x_2) + z_3(x_1,x_2)] - h^*[\, z_1(x_1^A) - z_1(x_1)\,].$$

As above, this is equivalent to choosing period-1 behavior to maximize

$$\tilde{V}^1(x_1,x_2) = z_1(x_1) + [1/(1+h)]^*[\, z_2(x_1,x_2) + z_3(x_1,x_2)].$$

Recalling that $x_2(x_1)$ maximizes

$$\tilde{V}^2(x_1,x_2) = z_2(x_1,x_2) + [1/(1+h)]^* z_3(x_1,x_2),$$

one can see that this model is equivalent to a model of hyperbolic discounting as in Laibson (1997) and O'Donoghue and Rabin (1999).[15] In other words, our model provides a reinterpretation of such models, specifically, that the source of the preference for immediate gratification comes from the motivation of the affective system, while at the same time the person gives no weight to in-the-moment willpower effort incurred in the future. The result is a time inconsistency in the "preferences" that rationalize behavior.

Consider next the alternative extreme in which the deliberative system gives full weight to future willpower effort exerted, and so its utility function is $U(x_1,x_2) = U^*(x_1,x_2) - h^*[\, z_2(x_1,x_2^A) - z_2(x_1,x_2)\,]$. In this case (and with a little manipulation), the deliberative system will choose period-1 behavior to maximize

$$\begin{aligned} V^1(x_1,x_2) = \quad & z_1(x_1) \quad - \quad h^*[\, z_1(x_1^A) - z_1(x_1)\,] \\ + \quad & z_2(x_1,x_2) \quad - \quad h^*[\, z_2(x_1,x_2^A) - z_2(x_1,x_2)\,] \\ + \quad & z_3(x_1,x_2). \end{aligned}$$

Recalling that $x_2(x_1)$ maximizes

$$\begin{aligned} V^2(x_1,x_2) = \quad & z_2(x_1,x_2) \quad - \quad h^*[\, z_2(x_1,x_2^A) - z_2(x_1,x_2)\,] \\ + \quad & z_3(x_1,x_2), \end{aligned}$$

[15] Specifically, the $(\beta,\delta)$ preferences used in those papers are equivalent to the preferences in the text when $\beta = 1/(1+h)$ and $\delta = 1$.

one can see that this model is much like a model of temptation utility as in Gul and Pesendorfer (2001). Once again, our model provides a reinterpretation of such models, suggesting specifically that the source of temptation (dis)utility is the willpower effort that the deliberative system must exert to control affective motivations, and that individuals fully take into account future temptation (dis)utilities. The result is that the "preferences" that rationalize behavior are time consistent, and yet the individual still has a preference for commitments that reduce future temptations.[16]

Hence, our model helps to identify similarities and differences between two different approaches to self-control problems that have appeared in the literature. The source of both is the motivation coming from the affective system; the difference is in whether the deliberative system cares about future willpower costs. Our formulation also highlights the fact that the two models represent extreme cases, suggesting that it might be fruitful to investigate the in-between case in which the deliberative system gives partial weight to future willpower effort.[17] At the same time, our model differs from existing models of hyperbolic discounting and temptation utility in an important way: Much as for our simple predictions with regard to the degree of discounting, our model predicts that the degree of hyperbolic discounting or temptation disutility will vary over time in systematic ways — depending on such things as willpower depletion, cognitive load, the proximity of stimuli, and the type of stimuli.

Our discussion above glosses over a second important issue with regard to period-1 behavior: How does the person form expectations about future behavior? A natural assumption (at least for economists) might be that the person has correct expectations about future behavior. However, this assumption relies on a more primitive assumption that the person accurately accounts for the effects of future affective motivations, and there is evidence that people tend to underestimate the influence of future affective motivations when predicting future behavior (see

[16] Bénabou and Pycia (2002) provide a similar reinterpretation of the Gul and Pesendorfer model. Motivated by Shefrin and Thaler's planner-doer model, they show that Gul and Pesendorfer's representation of preferences can be interpreted as coming from an intrapersonal conflict between two subselves, one short-sighted and the other long-lived. However, much as in Bernheim and Rangel's framework, they assume that one of the subselves will have complete control, but which one is ex ante uncertain.

[17] Formally, a person might put weight $\gamma \in (0,1)$ on future willpower costs, in which case her period-1 behavior will maximize $V^1(x_1,x_2) = z_1(x_1) - h^*[z_1(x_1^A) - z_1(x_1)] + z_2(x_1,x_2) - \gamma * h^*[z_2(x_1,x_2^A) - z_2(x_1,x_2)] + z_3(x_1,x_2)$.

for instance Loewenstein 1996).[18] Hence, it might make more sense to assume that the deliberative system has some perception $\hat{\gamma} \in [0,1]$ of how much the affective system will influence future decisions, and so the deliberative system predicts that period-2 behavior will maximize

$$\hat{\tilde{V}}^2(x_1,x_2) = z_2(x_1,x_2) + [1/(1+\hat{\gamma}h)]* z_3(x_1,x_2)$$

whereas actual period-2 behavior will maximize

$$\tilde{V}^2(x_1,x_2) = z_2(x_1,x_2) + [1/(1+h)]* z_3(x_1,x_2).$$

This issue is much like the distinction between sophistication ($\hat{\gamma} = 1$) and naivete ($\hat{\gamma} = 0$) in the literature on hyperbolic discounting. The difference here is that it makes precise the source of naivete — namely the failure to appreciate the influence of the affective system in future periods. Moreover, our discussion makes clear that this source of naivete is not something inherent to hyperbolic discounting, as it could equally well arise under temptation utility.[19]

We conclude our discussion of time preferences with a few comments on how to measure welfare, and in particular the implications of the two extreme approaches that we outlined in Section III. The extreme of the using the desirability of actions as perceived by the deliberative system means using the deliberative system's utility function $U(x_1,x_2)$. The alternative extreme of giving full weight to any willpower effort exerted corresponds to using $V^1(x_1,x_2)$. Thus, when the literature on hyperbolic discounting uses "long-run utility" as a welfare criterion (as in O'Donoghue and Rabin, 1999), it effectively takes the stance that we ought not to incorporate willpower effort when measuring welfare.[20] In contrast, when Gul and Pesendorfer (2001) apply a standard revealed-preference welfare analysis to their temptation preferences, they are

[18] If it is an environment that the person has experienced many times, an alternative justification for correct expectations is that the person has learned from past experience how she is likely to behave even if she doesn't fully understand what drives that behavior.

[19] For the sake of brevity, our analysis also glosses over a number of other issues with regard to applying our model to specific questions. For instance, in some environments the choice in period 1 might influence the affective optimum $x_2^A$ for period 2 (and therefore $z_2(x_1,x_2^A)$), in which case we would need to ask how well does the person account for such effects.

[20] Similarly, Bernheim and Rangel's (2002, 2003) claim that the cold-mode utility function is the appropriate welfare criterion is also taking the stance that we ought not to incorporate willpower effort when measuring welfare.

effectively taking the stance that willpower effort ought to be given full weight when measuring welfare.[21]

## V. Risk Preferences

A second natural application of our model is to risk preferences. Much as for time preferences, people are often of two minds when it comes to risks. We drive — or wish we were driving as we grip the airplane seat-divider with white-knuckles — even when we know at a cognitive level that it is safer to fly. We fear terrorism even when we know red meat poses a much greater risk of mortality. Perhaps the most dramatic illustration, however, comes from the *phobias* in which people are unable to face risks that they recognize, objectively, to be harmless. Indeed, the fact that people pay for therapy to deal with their fears, or take drugs (including alcohol) to overcome them, suggests that people's deliberative selves are not at peace with their affective reactions to risks.

The standard economic approach to risk preferences assumes that people choose between risky prospects based on their expected utility. However, there is a great deal of evidence that expected-utility theory is not a good descriptive theory of risk preferences, and a variety of alternative theories have been proposed (for a recent review see Starmer 2000). In this section, we describe how observed risk preferences may be influenced by the interaction between the deliberative and affective systems.

To apply our two-system approach, we must address the question of how the two systems respond to risks. For the deliberative system, a natural assumption is that risks are evaluated according to their expected utility (or perhaps expected value). Indeed, most researchers, as well as knowledgeable lay people, agree that expected-utility theory is the appropriate prescriptive theory to use for evaluating risks. It is less obvious what drives the affective system. Rather than speculate, we reverse-engineer what might be driving the affective system given what we know about successful descriptive models of risk preferences. In particular, there are two

[21] One interpretation of Gul and Pesendorfer's analysis is that it represents a *normative* analysis for how one ought to behave in the presence of forces from the affective system (under a belief that willpower effort ought to be given full weight when measuring welfare). From a *descriptive* perspective, however, we suspect that people fail to fully account for future willpower effects, both when predicting future behavior and when choosing current behavior.

features that show up in many descriptive theories of risk preferences: non-linear probability weighting and loss aversion.

Whereas expected-utility theory assumes that probabilities are weighed linearly, many successful descriptive theories of risk preferences assume that people transform the probabilities into decision weights. The most common form of probability weighting is the S-shaped probability-weighting function, wherein low probabilities are overweighted and high probabilities are underweighted. Our model suggests a simple explanation for why such a probability-weighting function might emerge: While the deliberative system may weight probabilities linearly, the affective system is more sensitive to outcomes than to probabilities. To illustrate this point, consider the extreme case where the deliberative system uses expected value and the affective system weights the value of all outcomes equally. More precisely, suppose a person must choose a lottery from some set, and suppose that affective optimum from this set is $\ell^A \equiv (x_1^A, p_1^A; ...; x_M^A, p_M^A)$. According to our model, the person will evaluate a lottery $\ell \equiv (x_1, p_1; ...; x_N, p_N)$ according to

$$[p_1 x_1 + ... + p_N x_N] - h * [(x_1^A + ... + x_M^A) - (x_1 + ... + x_N)].$$

Because the affective optimum is independent of the person's choice, this is equivalent to evaluating lottery $\ell \equiv (x_1, p_1; ...; x_N, p_N)$ according to

$$\left(\frac{p_1 + h}{1 + Nh}\right) * x_1 + ... + \left(\frac{p_N + h}{1 + Nh}\right) * x_N .$$

Hence, if one were to observe behavior generated by this model and use it to estimate a probability-weighting function, one would conclude that small probabilities are overweighted and large probabilities are underweighted (because $(p+h)/(1+Nh) > p$ for $p < 1/N$ and $(p+h)/(1+Nh) < p$ for $p > 1/N$).[22]

There is, in fact, evidence that supports this interpretation. For instance, studies that measure fear by means of physiological responses such as changes in heart rate and skin conductance — which primarily reflect activity in the affective system — find that reactions to an uncertain impending shock depend on the expected intensity of the shock but not the likelihood of receiving it (except if it is zero) (Deane 1969; Bankart and Elliott 1974; Elliott

[22] While this simple example doesn't generate the full S-shape, this is merely due to the linearity of the example.

1975; Monat, Averill, and Lazarus 1972; Snortum and Wilding 1971). Other evidence shows that emotional responses result largely from *mental images* of outcomes (Damasio, 1994). Because such images are largely invariant with respect to probability — one's mental image of winning a lottery, for example, depends a lot on how much one wins but not that much on one's chance of winning — emotional responses tend to be insensitive to probabilities.

Our model does more than provide an ex-post interpretation of the S-shaped probability weighting function. It predicts when the probability weighting function should become more S-shaped, namely when the affective system is playing a stronger role in decision making. Specifically, if a person's willpower is depleted, or if she is under cognitive load or stress, then she should exhibit a more S-shaped probability weighting function. While we know of no existing evidence on these dimensions, there is some closely related evidence that compares "affect-rich" decisions to "affect-neutral" decisions. Consistent with the prediction from our model, Rottenstreich and Hsee (2001) find that probability-weighting for affect-rich outcomes such as kisses, electric shocks, and vacations is more S-shaped — sensitive to departures from impossibility and certainty but insensitive to changes in intermediate probabilities — than probability-weighting of affect-poor outcomes such as money.

A second feature that shows up in many descriptive theories of risk preferences is loss aversion, which is the tendency to weight losses more heavily than gains. A possible interpretation suggested by our model is that loss aversion is a product of the affective system. To illustrate the implications of this interpretation, suppose again that the deliberative system uses expected value, But now suppose the affective system weighs losses more heavily than gains (and, for simplicity, weights probabilities linearly). Specifically, suppose the affective system evaluates a lottery $\ell \equiv (x_1, p_1; ...; x_N, p_N)$ according to $\sum_{i=1}^{N} p_i v(x_i)$ where

$$v(x_i) = \begin{cases} x_i & \text{if } x_i \geq 0 \\ \lambda x_i & \text{if } x_i \leq 0. \end{cases}$$

If a person must choose a lottery from some set, and if the affective optimum from this set is $\ell^A \equiv (x_1^A, p_1^A; ...; x_M^A, p_M^A)$, then according to our model the person will evaluate a lottery $\ell \equiv (x_1, p_1; ...; x_N, p_N)$ according to

$$[p_1 x_1 + ... + p_N x_N] - h * [(p_1 v(x_1^A) + ... + p_M v(x_M^A)) - (p_1 v(x_1) + ... + p_N v(x_N))].$$

Because the affective optimum is independent of the person's choice, this is equivalent to evaluating lottery $\ell \equiv (x_1, p_1; \ldots; x_N, p_N)$ according to $\sum_{i=1}^{N} p_i \hat{v}(x_i)$ where

$$\hat{v}(x_i) = \begin{cases} x_i & \text{if } x_i \geq 0 \\ \left(\frac{1+h\lambda}{1+h}\right) x_i & \text{if } x_i \leq 0 \end{cases}$$

Hence, under our interpretation of loss aversion coming from the affective system, if either the affective system feels stronger loss aversion (higher $\lambda$) or the affective system has more influence (higher $h$), the person should exhibit increased loss aversion. Once again, there is evidence that supports this interpretation.

One source of evidence examines the role of affect for the "endowment effect" — the tendency to value an object more highly when one owns it, a tendency that is commonly attributed to loss aversion (e.g., Tversky and Kahneman, 1991). Analogous to the finding that probability weighting is more S-shaped for affect-rich outcomes, considerable research suggests that loss aversion is much more pronounced for emotional outcomes such as changes in health status (see for instance Thaler, 1980). In one meta-analysis (Horowitz and McConnell, 2002), whereas the median ratio of willingness to accept relative to willingness to pay for ordinary private goods was found to be about 2.9, the mean ratio for goods involving health and safety was 10. A somewhat different form of evidence comes from recent research by Lerner, Small, and Loewenstein (forthcoming), who show that fairly subtle manipulations of affect — specifically inductions of sadness and disgust — have dramatic effects on the endowment effect.

A second source of evidence comes from patients with brain lesions. Shiv et al (2003) compared normal people, patients with brain lesions in regions related to emotional processing (they were normal on most cognitive tests, including tests of intelligence), and patients with lesions in regions unrelated to emotion. Subjects made 20 rounds of investment decisions, where in each round they were given a dollar and made a choice between keeping it or wagering it on a 50-50 chance of losing it or winning $2.50. Patients with emotion-related lesions invested more often than other subjects — that is, they exhibited less loss aversion — and ultimately earned more money. Moreover, whereas normal people were influenced by their outcomes in previous rounds, patients with emotion-related lesions were not.

Our two-system interpretation also predicts that the proximity of uncertain outcomes will play an important role in observed risk preferences. One dimension of proximity is the vividness

of outcomes, and Nisbett and Ross (1980) demonstrate that affective reactions to risks can depend on how vividly those risks are described. For instance, people react differently when a car accident is described simply as a fatal accident vs. when details can be provided — e.g., "the truck's wheel ran over the driver's head" —and in particular, providing details greatly increases emotional reactions without having much effect on cognitive evaluations. More generally, the effect of vividness on risk-taking behavior seems quite evident from real-world observations. For example, it is notorious that people slow down immediately after viewing a graphic wreck on the highway, until the vivid image of the wreck recedes from their mind.

A second dimension of proximity is the temporal proximity of uncertain outcomes. There is a great deal of evidence that temporal proximity is an important determinant of fear responses. As the prospect of an uncertain aversive event approaches in time, fear tends to increase even when cognitive assessments of the probability or likely severity of the event remain constant (Loewenstein, 1987; Roth et al. 1996). Similarly, after the moment of peak risk recedes into the past (e.g., after a near-accident), fear lingers for some period, but dissipates over time.[23] Evidence that temporal proximity can influence risk behaviors comes from studies that document "chickening out" wherein people initially agree to do various embarrassing things in front of other people (mime, sing, tell jokes, or dance) in exchange for payment, but then later change their minds (Van Boven et al 2002). Moreover, consistent with changes in the affective state of fear being the cause, subjects who were shown a film clip designed to induce fear (from Kubrick's "The Shining") right before they made their initial decision were much less likely to choose to perform, and hence much less likely to change their minds when the time came.[24]

What are the normative implications of our model for risk preferences? For instance, if people are aware that flying is safer, but they experience less fear when they drive, would they be better off if they were induced to fly? On the one hand, the fear people experience when they fly is real, and should be taken into account to some degree. On the other hand, we know that people are less likely to die if they fly, and we may worry that the motivational impact of fear of

[23] Such a temporal pattern of fear is highly adaptive; an organism that felt similar fear toward distant and immediate risks would be unlikely to survive long in a hostile environment. Indeed, one of the characteristics of certain types of stress disorders that clinical psychologists treat is the tendency to ruminate over risks that are remote in time (e.g., Nolen-Hoechsema, 1990; Sapolsky, 1994) or to continue to experience fear toward no-longer threatening events that happened in the past (e.g., Barlow, 1988)

[24] The example of chickening out illustrates how a dynamic inconsistency can arise in risk preferences due to changing affective states over time. Here, as temporal proximity changes, so do risk attitudes.

flying exceeds the actual utility loss associated with opting for the safer mode of transportation. Again, the answer is unclear, and depends on how much weight one feels we ought to place on affective motivations.

## VI. Social Preferences

It cannot be controversial to anyone but a Vulcan that social preferences are powerfully influenced by affect. Humans experience a wide range of social emotions, from powerful empathic responses such as sympathy and sadness to more negative emotions such as anger and envy. To give a flavor for how our two-system perspective can more generally be applied to social preferences, in this section we apply our model to one specific social motive — altruism — and its associated affect — empathy.

Modeling social preferences in our framework is somewhat more complicated than modeling time preferences or risk preferences because there is no widely agreed-upon economic model of social preferences that could be used as a natural starting point for the deliberative system. Hence, our discussion of social preferences is even more speculative than our discussion in the previous two sections. The perspective we suggest is that the deliberative system is driven by moral and ethical principles for how one ought to behave, whereas the affective system is driven toward anything between pure self-interest and extreme altruism depending on the degree of empathy that is triggered.

To motivate this perspective, we begin with studies of other-regarding behavior in animals, which reveal something about the affective — "animal" — system of the human brain. A number of studies show that animals, including monkeys and rats, can be powerfully moved by the plight of others (for an overview, see Preston and de Waal 2002). For example, rats who view a distressed fellow rat suspended in air by a harness will press a bar to lower the rat back to safe ground (Rice and Gainer 1962). A more recent study demonstrates that rats can have such powerful empathic reactions to others that they become debilitated — specifically, when another rat is administered electric shocks, the focal rat may retreat to a corner and crouch there motionless (Preston and de Waal 2002). In another remarkable study (Masserman et al 1964), hungry rhesus monkeys were trained to pull two chains, one of which delivered half as much food as the other. The situation was then altered so that pulling the chain with the larger reward caused a monkey in sight of the subject to receive an electric shock. After witnessing such a

shock, two-thirds of the monkeys preferred the non-shock chain and, of the remaining third, one monkey stopped pulling either chain for 5 days and another for 12 days after witnessing another being shocked.

At the same time, other-regarding behavior is not always observed in animals. In the primate studies, for instance, self-starvation to avoid shocking another animal was induced more by visual than auditory cues (i.e., seeing as opposed to hearing the distress of the other animal), was more likely in animals that had experienced shock themselves, was enhanced by familiarity with the shocked individual, was less when the shock recipient was an albino, and was nonexistent when it was a different species of animal. Perhaps stretching the terminology we introduced in Section II, we can view all of these factors as dimensions of proximity.

Humans, like animals, are capable of remarkable depths of empathy toward others in some circumstances, and remarkable empathic indifference in other circumstances. In the Theory of Moral Sentiments, Adam Smith provides a chilling account of the latter:

> Let us suppose that the great empire of China, with all its myriads of inhabitants, was suddenly swallowed up by an earthquake, and let us consider how a man of humanity in Europe, who had no sort of connection with that part of the world, would be affected upon receiving intelligence of this dreadful calamity. He would, I imagine, first of all express very strongly his sorrow for the misfortune of that unhappy people, he would make many melancholy reflections upon the precariousness of human life, and the vanity of all the labours of man, which could thus be annihilated in a moment… And when all this fine philosophy was over,.. he would pursue his business or his pleasure, take his repose or his diversion, with the same ease and tranquility as if no such accident had happened. The most frivolous disaster which could befall himself would occasion a more real disturbance. If he was to lose his little finger to morrow, he would not sleep to-night; but, provided he never saw them, he will snore with the most profound security over the ruin of a hundred millions of his brethren.

If people based their behavior toward other people solely on their affective — empathic — reactions to them, then, sympathetic beggars would be millionaires and United Way would go out of business. Given this remarkable lack of connection between our empathy toward others and the gravity of their plight or need for assistance, how is it that humans behave in an at all sensible way toward fellow humans? According to Adam Smith, the answer is “reason, principle, conscience, the inhabitant of the breast, the man within, the great judge and arbiter of

our conduct." In other words, it is the fact that the deliberative system moderates the empathic reactions of the affective system.

It is unclear what exactly the goals of the deliberative system are, but evidence suggests that it is not pure self-interest. People seem to have well-defined notions of what would be a fair or reasonable allocation of resources between two unknown people (Yaari and Bar-Hillel 1984). Moreover, these well-defined notions also influence people's choices in simple allocation (dictator) games in which they allocate resources between themselves and anonymous others — situations that should evoke relatively little empathy (Andreoni and Miller 2002, Charness and Rabin 2002). Hence, it seems likely that the goals of the deliberative system reflect some combination of moral and ethical principles for how one ought to behave. Indeed, philosophers have long discussed how people's behavior can be influenced by sophisticated reasoning about ethical principles (e.g., Kant, 1785/1991).[25]

To better illustrate our perspective, consider a two-person model of altruistic preferences (the extension to more people is straightforward). Let $(\pi_1,\pi_2)$ denote the material (monetary) payoffs for person 1 and person 2, respectively, and consider person 1's altruistic preferences. Suppose person 1's deliberative system puts some objective weight $\sigma$ on the person 2's material payoff — so the deliberative system's utility function is $\pi_1 + \sigma\pi_2$. At the same time, the weight that person 1's affective system puts on player 2's material payoff depends on the degree of empathy that person 1 currently feels toward person 2, which we denote by $e_{12}$ — so the affective system's motivational function is $\pi_1 + e_{12}\pi_2$. Now, if person 1 must choose a payoff vector from some budget set, and if the affective optimum from this set is $(\pi_1^A,\pi_2^A)$, then according to our model the person will evaluate a payoff vector $(\pi_1,\pi_2)$ according to

$$[\ \pi_1 + \sigma\pi_2\ ] - h^*[\ (\pi_1^A + e_{12}\pi_2^A) - (\pi_1 + e_{12}\pi_2)\ ].$$

Because the affective optimum is independent of the person's choice, this is equivalent to evaluating payoff vector $(\pi_1,\pi_2)$ according to

$$\pi_1 + \left(\frac{\sigma + he_{12}}{1+h}\right)\pi_2.$$

[25] Further information about the interactions between the affective and deliberative systems in altruistic behavior comes from considering certain abnormal populations. For instance, psychopaths and sociopaths, who tend not experience empathy, are purely Machiavellian and self-interested (Cleckley 1976; Lykken 1995).

This model illustrates a number of implications of our perspective. First, unlike for time preferences and risk preferences where the affective system moves behavior away from the deliberative optimum in one systematic direction, here the affective system could push behavior towards more or less altruism relative to the deliberative optimum. In situations where there is very little empathy triggered in the affective system, the affective system will push behavior closer to pure self-interest — as reflected by $e_{12} < \sigma$ implying $\left(\frac{\sigma + he_{12}}{1+h}\right) < \sigma$. In contrast, in situations where there are very high levels of empathy triggered in the affective system, the affective system will push behavior towards more altruism — as reflected by $e_{12} > \sigma$ implying $\left(\frac{\sigma + he_{12}}{1+h}\right) > \sigma$. When a person passes a sympathetic beggar on the street, the person may find herself giving money to the beggar when she thinks she ought to give that money to the United Way. At the same time, when a person is at home and not experiencing any empathic reactions, she may find herself not giving to the United Way when she thinks she ought to (it often seems to require "effort" to write that check to the United Way).

A closely related implication of our perspective is that the effect of willpower depletion or cognitive load on altruistic preferences depends on the degree of empathy triggered. When a person experiences little or no empathy — e.g., when deciding whether to donate to the United Way — our model predicts that willpower depletion or cognitive load should reduce the likelihood of the act — as reflected by $e_{12} < \sigma$ implying $\left(\frac{\sigma + he_{12}}{1+h}\right)$ is decreasing in $h$. In contrast, when a person experiences high empathy — e.g., when deciding whether to pay for a sympathetic beggar's dinner — our model predicts that willpower depletion or cognitive load should increase the likelihood of the act — as reflected by $e_{12} > \sigma$ implying $\left(\frac{\sigma + he_{12}}{1+h}\right)$ is increasing in $h$.

A study by Skitka et al (2002) provides limited support for these implications. Subjects were shown a number case studies of people who had contracted AIDS in different ways, and different case studies made the victim appear more or less responsible (e.g., sexual contact versus a blood transfusion). For each case study, subjects were asked whether the individual should be given subsidized access to drug treatment, and filled out measures of blame and responsibility. In addition, subjects were asked their political orientations. The key manipulation for our perspective is that half of the subjects made their judgments and allocation decisions while also engaged in a tone-tracking task that has been commonly used to induce cognitive load. The study found that subjects were less likely to advocate subsidized treatment under conditions of high load, which we would interpret as evidence that deliberative reactions are more sympathetic than affective reactions to AIDS victims. More interestingly, under conditions of high load, both liberals and conservatives were less likely to provide subsidized treatment to those deemed responsible (relative to those deemed not responsible), whereas under conditions of low load, liberals treated both groups equally whereas conservatives continued to favor groups who were seen as less responsible for contracting the disease. These findings are consistent with our framework if affective and deliberative reactions were consistent for conservatives — so cognitive load has no effect — but conflicting for liberals.

Much as for other domains, there are predictable effects of proximity on altruism. For example, while many people believe at an intellectual level that all people should have a similar claim on their sympathies (besides, perhaps, family members), in fact there is a natural human tendency to have stronger empathic reactions to those who are close to one geographically (e.g., in the same country), or similar to one in terms of ethnicity, age, gender, social class and so on.[26] Indeed, humans can be powerfully moved by even relatively minor misfortunes to those in close proximity — including those in books and movies occurring to fictitious characters — but unmoved by real human misery on a massive scale, even when they know that the latter is far more deserving of attention. This was graphically illustrated, recently, in a New York Times Magazine article (Greene, 2002) about AIDS orphans in Ethiopia in which the author contrasted

[26] Similar patterns can be seen with respect to other social emotions such as hatred and envy — although, logically, it might seem that everyone should envy Bill Gates, the reality is that most people reserve intense envy for those who are similar to them and/or in close physical proximity.

her intense emotional reaction to the death of the father of a classmate of her daughter to her lack of concern about AIDS orphans in Africa — until she actually visited Africa and witnessed the problem first-hand.

The role of vividness for the affective system may help to explain why people treat statistical deaths differently than identifiable ones, since foreknowledge of who will die (or which group deaths will come from) creates a more concrete — and evocative — image of the consequences (see, Schelling 1968; Bohnet and Frey 1999; and Small and Loewenstein, 2003 for an experimental demonstration). The impact of identifiability on affect may help to explain an anomalous tendency of altruists to contribute more to specific instances of a problem than to appeals addressing the entire problem, and more to specific victims than to multiple victims, even when the latter dominates the former in terms of total help rendered (Kogut and Ritov, 2003). Requests for donations to medical research which base their requests on the testimony of a single "poster child" rather than general descriptions of the affliction or its prevalence seem to exploit this phenomenon.[27]

Finally, our model may help clarify some of the welfare debates with regard to altruistic preferences. One question that has been debated in the literature is whether altruism should be incorporated into welfare analysis. A common argument against incorporating altruism is that altruistic preferences seem to be somewhat transient and subject to framing effects, and hence should not be treated as a "real" preference. Our model permits a reinterpretation of this argument. It is the altruistic motives generated by the affective system that are transient (at least to some extent) and subject to framing. However, to the extent that some altruistic motives come from the more stable deliberative system, these motives perhaps should be incorporated into welfare analysis. Indeed, if one takes the perspective that the deliberative system's utility function is the appropriate standard for welfare analysis, this is exactly what one would conclude.

A second question sometimes debated in the literature is the more philosophical question of whether a person who engages in some altruistic act is actually behaving in a "selfish" manner. Once again, while our model doesn't answer this question, it perhaps suggests a useful parsing. Specifically, if altruistic behavior is being driven primarily by the affective system and

[27] Identifiability is important for other social emotions besides altruism. For instance, identified perpetrators of criminal acts also elicit more intense emotional reactions than unidentified, statistical, perpetrators.

the desire to alleviate empathic emotions, then there is a sense in which the act really is a selfish act. But if an altruistic act is being driven primarily by the deliberative system and in particular by a feeling for what's moral or ethical, then perhaps the act should be viewed as genuinely altruistic.

## VII. Discussion

There is a great deal of evidence that people's decisions (and judgments and attitudes and so forth) are influenced by both affective and deliberative processes. The standard economic model focuses, in a sense, exclusively on deliberative processes. Our main contribution in this paper has been to develop a framework to incorporate affective processes into economic analyses, and to analyze the interactions between the two systems. We conclude by discussing the broader implications of our framework.

Economics needs to incorporate the effects of affective processes. We have already demonstrated how the introduction of affective processes can be useful for understanding (i) time preferences and the factors that influence the degree of time preferences (and the degree of self-control problems); (ii) risk preferences and people's tendencies to take (and avoid) the "wrong" risks; and (iii) altruistic preferences and people's tendencies to help the "wrong" victims. But our framework can be useful for a much broader set of applications. For instance, affective processes would seem to be important for understanding advertising, especially advertising that conveys no evident information; behavior in negotiations, and in particular why bargaining often breaks down into a mutually destructive, affect-driven morass (e.g., nasty divorces); behavior in financial markets and especially reactions to news events; and political preferences, and specifically how people seem to respond to political candidates, political parties, and issues at an affective level. Moreover, even in the domains we have discussed, a more detailed application of our framework would have the potential to explain many complex real-world behaviors.

There are a number of directions in which to further expand upon our framework. Perhaps the most important is to more fully explore the dynamics of willpower, which is the most novel aspect of our model. While we have described the effects of short-term changes in willpower strength, the long-term dynamics of willpower may be more important. For instance, our model suggests an alternative explanation for why poor people might be more prone to engage in risky behaviors such as smoking, unsafe sex, and so forth. Existing explanations

usually take the form of poor people's benefits being larger than rich people's, their costs being lower (poor people have less to live for), or the assumption that they are more impatient (short-sighted). Our model implies that if poor people are constantly required to exert willpower to live within their means (i.e., to constantly forgo enticing purchases), then they will have relatively little willpower strength remaining to resist inexpensive temptations like cigarettes or a willing sexual partner.

The dynamics of willpower might be even more important in making sense of the complicated patterns of self-control behavior (or lack thereof) that have been documented in the literatures on addiction, dieting, and sexual risk-taking. To illustrate, consider the difference between rats and humans in drug consumption. Connect a rat to an IV line that, upon the press of a lever, administers a sufficiently rich dose of cocaine or other powerfully reinforcing drug, and one will observe a strikingly simple sequence of behavior. The rat will press the lever repeatedly until it passes out from exhaustion, and when it comes to will resume pressing the lever, eventually to the point of death. Human addicts are far more complicated: they binge, go cold turkey, enter rehab programs, flush their drugs down the toilet, and relapse. The difference clearly comes from our ability to deliberate about the broader consequences of our behavior; but to understand these behavioral complexities, we need to incorporate the dynamics of the battle between the two systems.[28]

There are even more nuanced willpower dynamics. For instance, some, albeit preliminary, studies have found support for the idea that, in addition to being depleted in the short-term by exertion, willpower, like a muscle, may become strengthened in the long-term through repeated use (Muraven, Baumeister, and Tice 1999). More importantly, people's behavior might also reflect their attempts to manage their use of willpower. There is in fact experimental evidence in (a modified version of) the Baumeister paradigm that people do have some awareness of the dynamic properties of willpower and take these into account in a strategic fashion (Muraven 1998). Specifically, people who were aware that there would be multiple

[28] Similar behavioral complexities arise in other domains as well. In the realm of dieting, for instance, the frequent bouts of overeating that most dieters are subject to can be triggered by such disparate events as having eaten something that "breaks" one's diet, thinking that one has done this (whether the food eaten was truly high calorie or not); feeling anxious, depressed or otherwise dysphoric; drinking alcohol; eating with someone else who overeats; smelling and thinking about attractive foods; or being deprived of a favorite food (for an overview, see Herman and Polivy 2003).

willpower tasks seemed to conserve willpower on the earlier task (relative to subjects who were unaware), and in fact those subjects were able to exert more willpower on the later task.

A second direction in which to expand our framework is to incorporate cross-domain aspects of the affective system. While the deliberative system may be well aware of which considerations are relevant for the decision at hand, the affective system may be influenced by a variety of irrelevant factors. In particular, the activation of the affective system often persists even when the source of that activation is no longer present, and hence a stimulus can influence seemingly unrelated behaviors. Indeed, a tremendous amount of research in psychology documents how emotions elicited during a first phase of an experiment can influence subjects' judgments or behavior in a second phase even when those emotions are irrelevant. For instance, Lerner, Small, and Loewenstein (2004) show that which of three film clips — chosen to elicit different emotions — that a subject views has a dramatic impact on subsequent buying or selling prices for an object. Psychologists refer to these affective influences that are patently unrelated to a decision at hand as "incidental affect" (Bodenhausen 1993, 1994). These carry-over effects have important implications for behavior. For instance, the deliberative system may be perfectly aware that it makes no sense to take out frustrations from work on one's spouse; but if the negative feelings generated at work carry over into the home, such cognitive awareness may make little difference (Loewenstein and Lerner 2003).

A third direction in which to expand our framework is to people's assessments of their own behaviors. Because such assessments are an inherently cognitive task, they will naturally tend to exaggerate the role played by deliberation. In effect one could say that the deliberative self egocentrically views itself as in control and commensurately underestimates the influence of affect (see Wegner & Wheatley, 1999).[29] This failure to appreciate the role of affect in behavior can have a negative impact on efforts at self-control. Perhaps the most important form of self-control is not willpower, but rather "self-management" — the ability to avoid getting into, or to remove oneself from, a situation that is likely to engender self-destructive affective motivation. Dieters may steer clear of banquets, drug addicts of places and persons associated with drug-taking, smokers of smoky bars, and alcoholics of bars and parties. To the extent that people are

[29] There are a number of studies in which subjects are "manipulated" into behaving in certain ways and then asked to explain that behavior, and people invariably come up with plausible reasons for why the behavior was purposeful (see for instance Brasil-Neto, Pascual-Leone, Valls-Sole, Cohen, and Hallett, 1992).

unaware of, or underappreciate, the impact of affect on their own behavior, they are likely to underutilize such strategies of self-control.

A second implication of failing to appreciate the role of affect is that people will exaggerate the importance of willpower as a determinant of self-control. People who are thin often believe they are thin due to willpower, and that those who are less fortunate exhibit a lack of willpower. However, it is far more likely that those who are thin are blessed (at least in times of plentiful food) with a high metabolism or a well-functioning ventromedial hypothalamus (which regulates hunger and satiation). Indeed, obese people who go to the extraordinary length of stapling their stomach to lose weight often report that they have a sudden experience of "willpower" despite the obvious fact that stapling one's stomach affects hunger rather than willpower (Gawande 2001). It is easy and natural for those who lack drives and impulses for drugs, food, and sex to condemn, and hence to be excessively judgmental and punitive, toward those who are subject to them — to assume that these behaviors result from a generalized character deficit, a deficiency in willpower. Similarly, the rich, who are not confronted with the constant task of reigning in their desires, are likely to judge the short-sighted behaviors of the poor too harshly.

More speculatively, the deliberative system may have an ability to "train" the affective system over time to experience certain emotions — notably guilt and satisfaction — in a way that serves long-run goals. In particular, the deliberative system might train the affective system to experience guilt reactions in response to certain undesirable behaviors, and to experience feelings of satisfaction in response to certain desirable behaviors. For instance, many academics train themselves to experience guilt when they aren't working. More relevant for economics, many people seem to train themselves to experience satisfaction whenever they transfer money into their savings account and guilt whenever they transfer money out of that account. Of course, such training can have undesirable long-run effects. Someone who successfully creates affective reactions to savings behavior may find that those affective reactions persist even when it is logically time to start consuming those savings. Indeed, when these emotions become sufficiently intense and ingrained, they can actually drive behavior farther than the deliberative

system wanted, producing disorders of excessive preoccupation with the future such as workaholism and tightwaddism.[30]

An important issue is whether the influence of affective processes might be less important for high-stakes decisions. On one hand, it is true that people might be prone to deliberate more on high-stakes decisions. At the same time, however, high stakes could have exactly the opposite effect because high-stakes decisions tend to elicit intense affect.

Another important issue is whether market settings render affective processes irrelevant. Nothing could be further from the truth. Entire industries, such as gambling and sexual merchandising, are devoted to satisfying affective drives — and triggering them in the first place. Other industries exist to facilitate self-control such as dieting, smoking cessation, psychotherapy for phobias, and so on. Affective processes can also be important at a more macroeconomic level because affective reactions are often correlated across individuals. World events such as wars, terrorist attacks, natural catastrophes, and even economic gyrations have powerful affective consequences, instilling emotions and often inducing willpower-undermining stress. Writing in the New York Times Magazine about the economic consequences of the 9/11 terrorist attacks, Paul Krugman (2001:38) commented that "…the reason to be concerned about the economic effects of terrorism is not the actual damage but the possibility that nervous consumers and investors will stop spending." What, then, is the ultimate determinant of consumer demand according to Krugman? His answer is very similar to that which we offer in this paper: "If you ask how much consumers will consume and investors invest over the next few months, the answer is determined largely by feelings — what John Maynard Keynes called 'animal spirits.'"

Affect has a long, honorable, place in the history of economics. While we have emphasized Adam Smith's contributions, affect has a much deeper connection to economics. When the foundations of neoclassical economic theory were being put into place in the late-19th century, for example, economists of the time were acutely aware of the important role played by emotions in economics and wrote extensively of emotional influences on behavior. They tended, however, to view affect as an erratic, unpredictable force that was too complicated to incorporate

[30] In a casual survey of visitors to an airport, Loewenstein and Prelec (cite?) found that the majority of people perceived that spending too little rather than spending too much was their greater problem. Similarly, from survey data of TIAA-CREF participants, Ameriks et al (2003) find that, while many people perceive themselves to have a problem of over-consumption, many other people perceive their problem to be under-consumption.

into the mathematical models of behavior they were so anxious to formulate. As economics became increasingly mathematized, the appreciation of affect waned commensurately. Even so, many prominent economists continued to write about the role of affect in economics — Francis Edgeworth, Eugene von Böhm-Bawerk, Irving Fisher, Tibor Scitovsky, George Katona, Herbert Simon, Harvy Leibenstein, and, perhaps most famously, Keynes. But due to the difficulty in integrating affect into formal models of behavior, discussions of affect have been segregated from formal economic models — much like Adam Smith's segregation of his psychology and economics into two separate volumes. With the aid of recent developments in psychology and neuroscience that, to a remarkable degree, vindicate Smith's early insights, we hope to accelerate the process of reintegrating affect in formal economic analyses.

# References


Ameriks, John, Andrew Caplin, John Leahy, and Tom Tyler (2003). "Measuring Self Control."Mimeo, New York University.

Anderson, A.K., K. Christoff, D. Panitz, E. De Rosa, and J.D. Gabrieli (2003). "Neural correlates of the automatic processing of threat facial signals." *Journal of Neuroscience*, **23**(13, July 2), 5627-5633.

Andreoni, James and John Miller (2002). "Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism." *Econometrica*, **70**, 737-753.

Bankart, C. Peter and R. Elliott (1974). "Heart rate and skin conductance in anticipation of shocks with varying probability of occurrence." *Psychophysiology*, **11**, 160-174.

Bargh, J. (1984). "Automatic and conscious processing of social information," in R. S. Wyer and T. K. Srull, eds., *Handbook of social cognition (Vol 3)*. Hillsdale, NJ: Erlbaum, 1-43.

Barlow, D. H. (1988). *Anxiety and its disorders: The nature and treatment of anxiety and panic*. New York, NY: Guilford Press.

Baumeister, R. F., T. F. Heatherton, and D. Tice (1994). *Losing Control: How and Why People Fail at Self-Regulation*. San Diego: Academic Press.

Baumeister, R. F. and K. D. Vohs (2003). "Willpower, Choice, and Self-Control," in G. Loewenstein et al, eds., *Time and Decision*. New York: Russel Sage Foundation, 201-216.

Bénabou, Roland and Marek Pycia (2002). "Dynamic Inconsistency and Self-Control: A Planner-Doer Interpretation." *Economics Letters*, **77**, 419-424.

Bernheim, D. and A. Rangel (2002). "Addiction and Cue-Conditioned Cognitive Processes." NBER working paper #9329.

Bernheim, D. and A. Rangel (2003). "Emotions, Cognition, and Savings: Theory and Policy." Mimeo, Stanford University.

Berridge, K. (1995). "Food Reward: Brain Substrates of Wanting and Liking." *Neuroscience and Biobehavioral Reviews*, **20**(1), 1-25.

Bodenhausen, G. V. (1993). "Emotions, arousal, and stereotypic judgments: A heuristic model of affect and stereotyping," in Mackie, D. M. and D. L. Hamilton, *Affect, cognition, and stereotyping: Interactive processes in group perception*. San Diego, CA: Academic Press, 13-37.

Bodenhausen, G. V., L. Sheppard, et al. (1994). "Negative affect and social judgment: The different impact of anger and sadness." *European Journal of Social Psychology*, **24**, 45-62.

Bodenhausen, G. V., G. P. Kramer, et al. (1994). "Happiness and stereotypic thinking in social judgment." *Journal of Personality and Social Psychology*, **66**, 621-632.

Bohnet, Iris and Bruno Frey (1999). "The Sound of Silence in Prisoner's Dilemma and Dictator Games." *Journal of Economic Behavior & Organization*, **38**, 43-57.

Brasil-Neto, J. P., Pascual-Leone, A., Valls-Solé, J., Cohen, L. G. and Hallett, M. (1992). "Focal transcranial magnetic stimulation and response bias in a forced-choice task." *Journal of Neurology, Neurosurgery and Psychiatry*, **55**, 964-966.

Buck, R. (1984). *The communication of emotion*. New York: Guilford Press.

Camerer, C., Loewenstein, G., and Prelec, D. (2003). "Neuroeconomics: How Neuroscience Can Inform Economics." Mimeo, Carnegie Mellon University.

Carter, Rita (1998). *Mapping the Mind*. London: Wiedenfeld & Nicolson.

Chaiken, S. and Y. Trope, Eds. (1999). *Dual-process theories in social psychology*. New York: Guilford Press.

Charness and Rabin (2002). "Social Preferences: Some Simple Tests and a New Model." *Quarterly Journal of Economics*, **117**, 817-869.

Cleckley, Hervey Milton (1976). *The mask of sanity: an attempt to clarify some issues about the so-called psychopathic personality*. St. Louis: Mosby.

Damasio, A. R. (1994). *Descartes' error: Emotion, reason, and the human brain.* New York: G.P. Putnam.

Deane, G. E. (1969). "Cardiac activity during experimentally induced anxiety." *Psychophysiology*, **6**, 17-30.

Eisenberger, Naomi I., Matthew D. Lieberman, and Kipling D. Williams (2003). "Does Rejection Hurt? An fMRI Study of Social Exclusion." *Science*, **302** (October 10, 2003), 290-292.

Elliott, R. (1975). "Heart rate in anticipation of shocks which have different probabilities of occurrences." *Psychological Report*s, **36**, 923–931.

Freud, Sigmund (1924/1962). *The ego and the id.* Translated by Joan Riviere. New York: W. W. Norton.

Frijda, Nico H. (1986). *The Emotions*. Cambridge: Cambridge University Press.

Gawande, Atul (2001). "The Man Who Couldn't Stop Eating." *New York Times Magazine*, July 9, 2001.

Gazzaniga, M. S. and J. E. LeDoux (1978). *The Integrated Mind*. New York: Plenum.

Gilbert, Daniel T. and Michael Gill. 2000. "The momentary realist." *Psychological Science*, **11**, 394-398.

Greene, Melissa Fay (2002). "What Will Become of Africa's AIDS Orphans?" *New York Times Magazine* (December 22, 2002).

Greenwald, A.G. and M.R. Banaji (1995). "Implicit social cognition: Attitudes, self-esteem, and stereotypes." *Psychological Review*. **102**(1), 4-27.

Gul, F. and W. Pesendorfer (2001). "Temptation and Self-Control." *Econometrica*, **69**, 1403-1435.

Herman, C. Peter and Janet Polivy (2003). "Dieting as an Exercise in Behavioral Economics," in G. F. Loewenstein, D. Read and R. Baumeister, eds., *Time and decision: Economic and psychological perspectives on intertemporal choice*. New York: Russell Sage Foundation, 459-489.

Horowitz, John and McConnell, Kenneth (2002). "A Review of WTA-WTP Studies." *Journal of Environmental Economics and Management*, **44** (3), 426-447

Johnson, E. J., J. Hershey, et al. (1993). "Framing, Probability Distortions, and Insurance Decisions." *Journal of Risk and Uncertainty*, **7**, 35-51.

Kahneman, Daniel (1994). "New Challenges to the Rationality Assumption." *Journal of Institutional and Theoretical Economics* **150**(1), 18-36.

Kant, I. (1900). *Critique of pure reason*. New York. J. M. D. Meiklejohn. Willey, 1781.

Keynes, John Maynard (1936). *The general theory of employment, interest and money*. New York: Harcourt, Brace.

Kogut, Tehila and Ritov, Ilana (2003) "The 'identified victim' effect: an identified group, or just a single individual?" Paper presented at the 19th 'Biannual Conference on Subjective Probability, Utility, and Decision-Making," Zurich, Switzerland, 25–27 August 2003.

Krugman, Paul (2001). "Fear itself." *New York Times Magazine*, Sept 30, 2001.

Kunda, Ziva (1990). "The case for motivated reasoning." *Psychological Bulletin*, **108**, 480-498.

Laibson, D. (1997). "Golden eggs and hyperbolic discounting." *Quarterly Journal of Economics*, **112**, 443-477.

Laibson, D. (2001). "A Cue-Theory of Consumption." *Quarterly Journal of Economics*, **116**, 81-119.

Lang, P. (1995). "The emotion probe: Studies of motivation and attention." *American Psychologist*, **50**, 372-385.

Lang, P. (1988). "Fear, anxiety, and panic: Context, cognition, and visceral arousal," in S. Rachman and J. D. Maser, eds., *Panic: Psychological perspectives*. Hillsdale, NJ: Lawrence Erlbaum Associates, 219-236.

LeDoux, J. E. (1996). *The emotional brain: The mysterious underpinnings of emotional life*. New York, NY: Simon & Schuster.

Lerner, J. S., D. A. Small, and G. Loewenstein (2004). "Heart strings and purse strings: Carry-over effects of emotions on economic transactions." *Psychological Science*.

Lhermitte, F. (1986). "Human Autonomy and the Frontal Lobes. 2. Patient Behavior in Complex and Social Situations — the Environmental Dependency Syndrome." *Annals of Neurology*, **19** (4), 335-343.

Loewenstein, G. (1987). "Anticipation and the Valuation of Delayed Consumption." *Economic Journal* **97**, 666-684.

Loewenstein, G. (1996). "Out of control: Visceral influences on behavior." *Organizational Behavior and Human Decision Processes*, **65,** 272-292.

Loewenstein, George and Adler, Daniel (1995). A bias in the prediction of tastes. *Economic Journal*, **105**, 929-937.

Loewenstein, George and Erik Angner (2003). "Predicting and Indulging Changing Preferences," in George Loewenstein, Daniel Read, and Roy Baumeister, eds., *Time and decision: Economic and psychological perspectives on intertemporal choice*. New York: Russell Sage Foundation, 351-391.

Loewenstein, George & Lerner, Jennifer (2003). The role of emotion in decision making. In. R.J. Davidson, H.H, Goldsmith & K.R. Scherer, Handbook of Affective Science. Oxford, England: Oxford University Press.

Loewenstein, G. and D. Prelec — airport survey.

Lykken, David T. (1995). *The antisocial personalities*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

MacLean, P. D. (1990). *The Triune Brain in Evolution: Role in Paleocerebral Function*. New York: Plenum.

Manuck, S. B., J. Flory, M. Muldoon and R. E. Ferrell (2003). "A neurobiology of intertemporal choice," in G. F. Loewenstein, D. Read and R. Baumeister, eds., *Time and decision: Economic and psychological perspectives on intertemporal choice*. New York: Russell Sage Foundation, 139-172.

Masserman, Jules Hyman, Stanley Wechkin, and William Terris (1964). "`Altruistic' behavior in rhesus monkeys." *American Journal of Psychiatry* **121**, 584-585.

Massey, Douglas S. (2002) "A Brief History of Human Society: The Origin and Role of Emotion in Social Life." *American Sociological Review*, **67**, (1), 1-29.

McConnell, A. R. and J. M. Leibold (2001). "Relations among the Implicit Association Test, discriminatory behavior and explicit measures of racial attitudes." *Journal of Experimental Social Psychology*, **37**(5), 435-442.

Metcalfe, J. and W. Mischel (1999). "A hot/cool-system analysis of delay of gratification: Dynamics of willpower." *Psychological Review* **106**(1), 3-19.

Mischel, Walter, Ebbe B. Ebbesen, and Antonette Zeiss (1972). "Cognitive and Attentional Mechanisms in Delay of Gratification." *Journal of Personality and Social Pychology*, **21**(2), 204-218.

Mischel, Walter, Ozlem Ayduk, and Rodolfo Mendoza-Denton (2003). "Sustaining Delay of Gratification over Time: A Hot-Cool Systems Perspective," in G. F. Loewenstein, D. Read and R. Baumeister, eds., *Time and decision: Economic and psychological perspectives on intertemporal choice*. New York: Russell Sage Foundation, 175-200.

Mischel, Walter, Yuichi Shoda, and Monica L. Rodriguez (1989). "Delay of Gratification in Children." *Science*, **244**(4907), 933-938.

Monat, A., J. R. Averil, et al. (1972). "Anticipatory stress and copying reactions under various conditions of uncertainty." *Journal of Personality and Social Psychology* **24**, 237-253.

Muraven, M. (1998). *Mechanisms of Self-Control Failure: Motivation and Limited Resource*. Ph.D. diss., Case Western Reserve University.

Muraven, M., R. F. Baumeister, and D. M. Tice (1999). "Longitudinal Improvement of Self-Regulation Through Practice: Building Self-Control Strength Through Repeated Exercise." *Journal of Social Psychology* **139,** 446-57.

Nisbett, R. E. and L. Ross (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs: Prentice-Hall.

Nolen-Hoeksema, S. (1990). *Sex Differences in Depression*. Stanford, CA: Stanford University Press.

Ochsner, Kevin N. and James J. Gross (2004). "Thinking makes it so: A social cognitive neuroscience approach to emotion regulation," in Roy F. Baumeister and Kathleen D. Vohs, eds., *Handbook of self-regulation: Research, theory, and applications*. New York: Guilford Press, 229-255.

O'Donoghue, T. and M. Rabin (1999). "Doing it Now or Later." *American Economic Review*, **89**, 103-124.

Phelps, E. A., K. J. O'Connor, W. A. Cunningham, E. S. Funayama, J. C. Gatenby, J. C. Gore, and M. R. Banaji (2000). "Performance on indirect measures of race evaluation predicts amygdala activity." *Journal of Cognitive Neuroscience*, **12**(5), 729-38.

Plato, Republic. trans. G.M.A. Grube, rev. C.D.C. Reeve (indeanapolis and Cambridge: Hackett Publishing Company, 1992).

Preston, S. and F. de Waal (2002). "Empathy: Its ultimate and proximate bases." *Behavioral and Brain Sciences*, **25**(1), 1-71.

Rice, G. E. Jr. and P. Gainer (1962). "'Altruism' in the Albino Rat." *Journal of Comparative and Physiological Psychology*, **55**(1), 123-125.

Ritov, I. and J. Baron (1990). "Reluctance to Vaccinate: Omission Bias and Ambiguity." *Journal of Behavioral Decision Making*, **3,** 263-277.

Roth, W. T., G. Breivik, et al. (1996). "Activation in novice and expert parachutists while jumping." *Psychophysiology* **33**, 63-72.

Rottenstreich, Y. and C.K. Hsee (2001). "Money, kisses, and electric shocks: On the affective psychology of risk." *Psychological Science*, **12**, 185-190.

Sapolsky, R. M. (1994). *Why zebras don't get ulcers*. New York: W. H. Freeman and Co.

Sayette, Michael A., Christopher S. Martin, Joan M. Wertz, and others (2001). "A multi-dimensional analysis of cue-elicited craving in heavy smokers and tobacco chippers." *Addiction* **96**(10, Oct 2001), 1419-1432.

Schelling, T. C. (1968). "The life you save may be your own," in S. B. Chase, ed., *Problems in Public Expenditure Analysis.* Washington, DC: The Brookings Institute.

Shallice, Tim and Paul Burgess (1998). "The domain of supervisory processes and the temporal organization of behaviour," in A. C. Roberts, T. W. Robbins and L. Weiskrantz, eds., *The Prefrontal Cortex: Executive and Cognitive Functions*. Oxford University Press.

Shefrin, H. M. and R. H. Thaler (1988). "The Behavioral Life-Cycle Hypothesis." *Economic Inquiry* **26**, 609-643.

Shiffman, Saul and Andrew J. Waters (2004). "Negative Affect and Smoking Lapses: A Prospective Analysis." *Journal of Consulting and Clinical Psychology*, **72**(2), 192-201.

Shiv, B. and A. Fedorikhin (1999). "Heart and Mind in Conflict: The Interplay of Affect and Cognition in Consumer Decision Making." *Journal of Consumer Research*, **26**, 278-292.

Shiv, Baba, George Loewenstein, Antoine Bechara, Hanna Damasio, and Antonio Damasio (2003). "Investment Behavior and the Dark Side of Emotion." Mimeo, University of Iowa.

Skitka, Linda J., Elizabeth Mullen, Thomas Griffin, S. Hutchinson, and B. Chamberlin (2002). "Dispositions, Ideological Scripts, or Motivated Correction? Understanding Ideological Differences in Attributions for Social Problems." *Journal of Personality and Social Psychology*, **83**, 470-487.

Small, D. A. and G. Loewenstein (2003). "Helping *the* victim or helping *a* victim: Altruism and Identifiability." *Journal of Risk and Uncertainty*, **26**(1), 5-16.

Smart, Laura and Daniel M. Wegner (1996). "Strength of Will." *Psychological Inquiry*, **7**(1), 79-83.

Smith, Adam (2002). *The Theory of Moral Sentiments*, edited by Knud Haakonssen. Cambridge: Cambridge University Press.

Snortum, J. R. and F.W. Wilding (1971). "Temporal estimation of heart rate as a function of repression-sensitization score and probability of shock." *Journal of Consulting and Clinical Psycholog*y, **37**, 417–422.

Spence, Sean A. and Chris D. Frith (1999). "Towards a functional anatomy of volition." *Journal of Consciousness Studies*, **6**(8-9, Aug-Sep 1999), 11-29.

Starmer, C. (2000). "Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk." *Journal Of Economic Literature*, **38**, 332-382.

Thaler, Richard (1980). "Toward a positive theory of consumer choice." *Journal of Economic Behavior and Organization*. **1**, 39-60.

Tversky, A. and D. Kahneman (1991). "Loss aversion in riskless choice: A reference dependent model." *Quarterly Journal of Economics*, **106**, 1039-1061.

VanBoven, Leaf, Dunning, David, and Loewenstein, George (2000) Egocentric empathy gaps between owners and buyers. *Journal of Personality and Social Psychology*, **79**, 66-76.

Van Boven, Leaf and George Loewenstein (2003). "Social Projection of Transient Visceral Feelings." *Personality and Social Psychology Bulletin*, **29**(09), 1159-1168.

Van Boven, Leaf, George Loewenstein, Edward Welch, and David Dunning (2002). "The Illusion of Courage: Underestimating the Impact of Fear of Embarrassment on the Self." Working paper, Dept. of Social and Decision Sciences, CMU.

Wegner, D. M. and T. Wheatley (1999). "Apparent mental causation: Sources of the experience of will." *American Psychologist*, **54**(7): 480-492.

Wegner, Daniel M. (1992). "You can't always think what you want: Problems in the suppression of unwanted thoughts," in Mark Zanna, ed., *Advances in experimental social psychology* (Vol. 25). San Diego: Academic Press, 193-225.

Wilson, Timothy. D., D. J. Lisle, Jonathan W. Schooler, Sara D. Hodges, K.J. Klaaren, and S.J. LaFleur (1993). "Introspecting about reasons can reduce post-choice satisfaction." *Personality and Social Psychology Bulletin*, **19**, 331-339.

Wilson, Timothy D. and Jonathan W. Schooler (1991). "Thinking too much: Introspection can reduce the quality of preferences and decisions." *Journal of Personality and Social Psychology*, **60**, 181-192.

Yaari, M. and M. Bar-Hillel (1984). "On Dividing Justly," *Social Choice and Welfare*, **1**, 1-24.

Zajonc, R. B. (1980). "Feeling and thinking: Preferences need no inferences." *American Psychologist*, **35**(2), 151-175.

Zajonc, R. B. (1984). "On the primacy of affect." *American Psychologist*, **39**(2), 117-123.

Zajonc, R. B. (1998). "Emotions," in D. Gilbert, S. Fiske and G. Lindzey, eds., *The handbook of social psychology*. New York: Oxford University Press, 591-632.