Précis of Breakdown of Will

Final draft of the précis that appeared in *Behavioral and Brain Sciences* 28, 635-673, 2005

George Ainslie
151 Department of Veterans Affairs Medical Center
Coatesville, PA 19320
George.Ainslie@va.gov
Picoeconomics.org

Based on:
BREAKDOWN OF WILL
Cambridge, 2001,258pp. IBSN: 0-521-59694-7

File Bkdn for BBS12,05

Short Abstract

Behavioral science has long been puzzled by the experience of temptation, the resulting impulsiveness, and the variably successful control of this impulsiveness. *Breakdown of Will* (Ainslie, 2001) presents evidence that contradicts rational models in which discounting the value of future events at a constant rate keeps preference consistent. Both people and nonhuman animals discount the value of expected events in a curve where value is divided approximately by expected delay, a *hyperbolic* form that is more bowed than the rational, exponential curve. This finding implies that conflicting reward-seeking processes will arise spontaneously to get incompatible goals available at different times, that in humans these processes will in effect bargain with each other, and that this bargaining can create ego functions like willpower from the bottom up. Motivation-based models of classical conditioning, compulsiveness, empathy, and the social construction of belief become possible.

Long Abstract

Behavioral science has long been puzzled by the experience of temptation, the resulting impulsiveness, and the variably successful control of this impulsiveness. In conventional theories a governing faculty like the ego evaluates future choices consistently over time, discounting their value for delay exponentially, that is, by a constant rate; impulses arise when this ego is confronted by a conditioned appetite. *Breakdown of Will* (Ainslie, 2001) presents evidence that contradicts this model. Both people and nonhuman animals spontaneously discount the value of expected events in a curve where value is divided approximately by expected delay, a *hyperbolic* form that is more bowed than the rational, exponential curve.

With hyperbolic discounting, options that pay off quickly will be temporarily preferred to richer but slower-paying alternatives, a phenomenon that, over times on the order of days, can account for impulsive behaviors, and over periods of fractional seconds can account for involuntary behaviors. Contradictory reward-getting processes can in effect bargain with each other, and stable preferences can be established by the perception of recurrent choices as test cases (precedents) in recurrent *intertemporal* prisoner's dilemmas. The resulting motivational pattern resembles traditional descriptions of the will, as well as of compulsive phenomena that can now be seen as side-effects of will: overconcern with precedent, intractable but circumscribed failures of self-control, a motivated (“dynamic”) unconscious, and an inability to exploit emotional rewards. Hyperbolic curves also suggest a means of reducing classical conditioning to motivated choice, the last necessary step for modeling many involuntary processes like emotion and appetite as reward-seeking behaviors; such modeling in turn provides a rationale for empathic reward and the "construction" of reality.

Key words: Altruism, appetite, behavioral economics, compulsions, classical conditioning, dynamic inconsistency, emotions, empathy, freedom of will, hyperbolic discounting, impulsiveness, intertemporal bargaining, self-control, social construction, volition, weakness of will

1. Introduction

In a prosperous society most misery is self-inflicted. We smoke, eat and drink to excess, and become addicted to drugs, gambling, credit card abuse, destructive emotional relationships, and simple procrastination, usually while attempting not to. The human bent for defeating our own plans has puzzled writers since antiquity. From Plato's idea that the better part of the self-- reason-- could be overwhelmed by passion, there evolved the concept of a faculty, will, that lent reason the kind of force that could confront passion and defeat it. The construct of the will and its power became unfashionable in twentieth century science, but the puzzle of self-defeating behavior-- what Aristotle called *akrasia*-- and its sometime control has not been solved. With the help of new experimental findings, and conceptual tools from economics, game theory and the philosophy of mind, it is possible to form a hypothesis about the nature of will that does not violate the conventions of science.

In this précis I have followed the outline of *Breakdown of Will* (Ainslie, 2001) in three sections and twelve chapters, but have necessarily been selective in what I describe in detail. In "Breakdowns of Will" I criticize the two main conventional approaches to impulsiveness and self-control (chapter 2), present experimental evidence that vertebrates' evaluation of future options is basically hyperbolic, rather than exponential as conventionally assumed (chapter 3), and argue that the hyperbolic form offers an alternative to classical conditioning as a mechanism for involuntary behaviors (chapter 4). In "A Breakdown of the Will" I argue that hyperbolically-based uncertainty about our own future choices leads us to see current choices as test cases (chapter 5), that this perception establishes willpower through an intertemporal version of the repeated prisoner's dilemma (chapter 6), that this model fits common experiences of will (chapter 7), and that substantial evidence favors the bargaining model over other models of willpower (chapter 8). In "The Ultimate Breakdown of Will" I describe how intertemporal bargaining leads to compulsive side effects (chapter 9) and how a hyperbolically based impulse toward premature satiation of appetite gives emotions their quasi-voluntary quality (chapter 10), and motivates the social construction of facts, the quest for vicarious experience, and indirect approaches to goals (chapter 11). I summarize the conclusions of these arguments in chapter 12.

BREAKDOWNS OF WILL:
THE PUZZLE OF AKRASIA

2. The Dichotomy At The Root Of Decision Science: Do We Make Choices By Judgments Or By Desires?

The puzzle of self-defeating behavior has provoked two kinds of explanation, neither of which has been adequate. Cognitive theories have stayed close to introspective experiences of will and its failure, using familiar concepts like strength, (*e.g.* Baumeister & Heatherton, 1996); but they have not offered systematic causal hypotheses. Utility-based theories have assumed a

comprehensive internal marketplace of desires that compete on the basis of the expected value of their goals, discounted exponentially for delay—that is, by a fixed percentage per unit of time:

$$\text{Value} = \text{Value at no delay} \times (1 - \text{Discount rate})^{\text{Delay}}$$

But discounting the future *per se* doesn't imply impulsiveness-- The most rational planners devalue delayed outcomes. On the contrary, the implication of exponential discounting is stability of preference; the preferred of a set of alternatives does not change based on the individual's proximity to the alternatives (figure 1A). Utility theories have accounted well for many properties of choice, but predict neither self-defeating behavior nor any faculty to prevent it. Hypotheses to reconcile self-defeating behavior with a decision-making process that maximizes utility have cited lack of experience with the consequences (e.g. Herrnstein & Prelec's "primrose path" to addiction, 1992), short time horizons (e.g. Becker & Murphy, 1988), conditioned cravings (e.g. Loewenstein, 1996), and recent discoveries about the neurophysiological process of reward (e.g. Ho et.al., 1998), but all of these explanations can be shown to be incomplete on experimental or logical grounds: Experienced addicts often re-addict themselves after becoming drug-free; short time horizons do not predict people's plans to avoid temptations when they face them from a distance; there is no reason that conditioned cravings should operate differently from other appetites, all of which have conditioned elements; and while studies of brain physiology reveal the sites of powerful rewards, they do not suggest how people come to avoid some of these rewards.

3. The Warp In How We Evaluate The Future

Quantitative research over the past three decades has given utility theory a rationale for the conflict between impulses and controls: The assumed exponential discount curve for discounting the value of expected events is not basic. There is extensive evidence that both people and nonhuman animals spontaneously value future events in inverse proportion to their expected delays (Green, Fry, et.al., 1994; Kirby, 1997; Mazur, 1997). The resulting hyperbolic discount curve is seen over all time ranges, from seconds to decades (Harvey, 1994). A variant of Herrnstein's matching law as applied to delay (Chung & Herrnstein, 1967), this curve is adequately described by Mazur's (1987) simple formula:

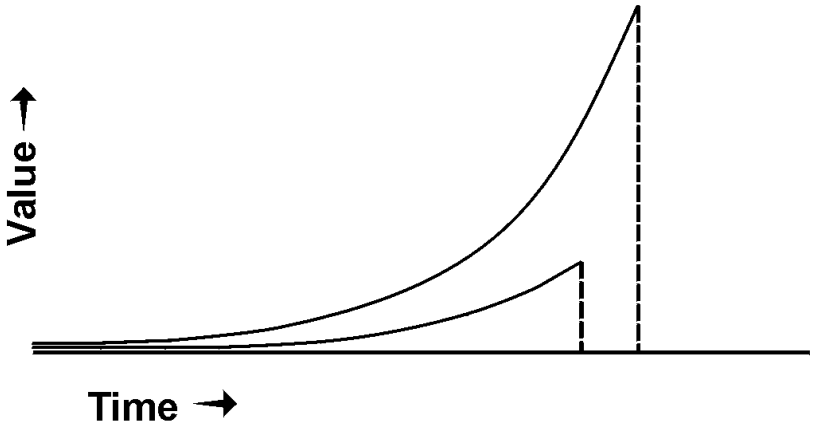
$$\text{Value} = \frac{\text{Value at no delay}}{[\text{Constant} + (\text{Impatience factor} \times \text{Delay})]}$$

The constant is a small number-- Mazur proposed an invariant "1"-- which describes the failure of values to approach infinity as delays approach zero. By varying only one element—the impatience factor-- investigators have been able to produce substantially better fits to choices among delayed rewards than have been possible with the exponential curves that most utility theories rely upon. Data include a number of animal studies (Grace, 1994; Mazur, 1997) and human experiments with both hypothetical (Kirby & Marakovic, 1995; Vuchinich & Simpson, 1998) and real (Green, Fry et.al., 1994; Kirby, 1997) money. Investigators sometimes report that their data fit still better if the denominator is raised to a power (Grace, 1994; Myerson & Green, 1995), but this power is usually close to 1.0, and in any case doesn't change the crucial

implication of this formula: that the elementary discount curve produces a basic tendency to prefer smaller rewards over larger ones *temporarily*, when the smaller reward is imminently available (figure 1B).

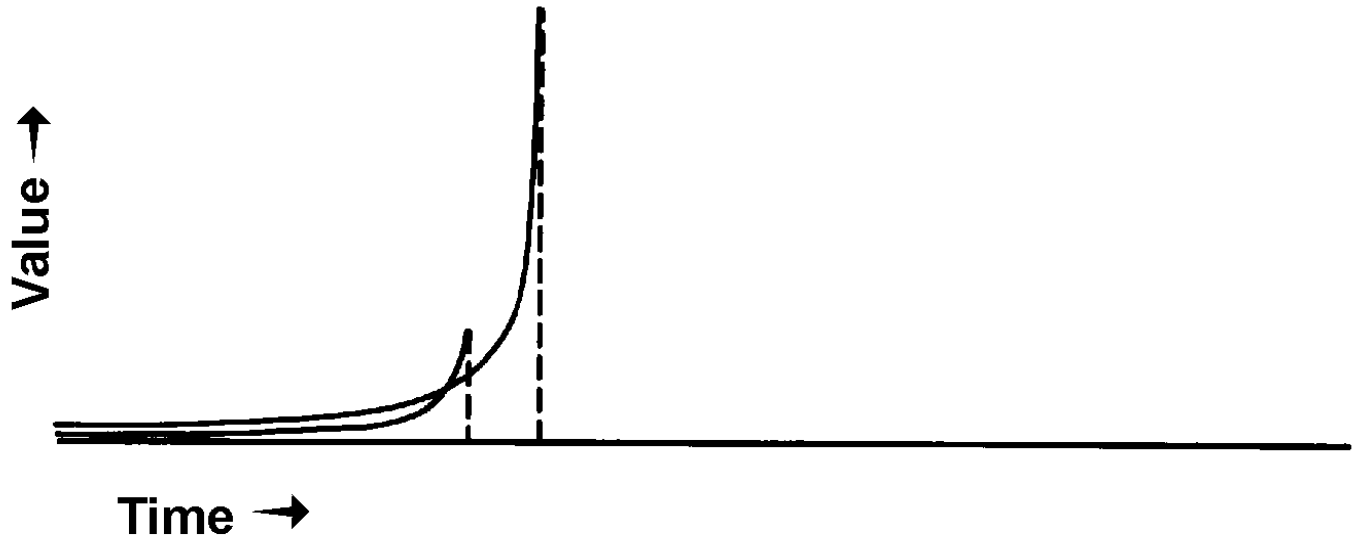
In contrast to exponential curves, hyperbolic discount curves depict a strong but temporary tendency to prefer smaller, sooner (SS) rewards to larger, later (LL) ones, in the period just before an SS reward is due. This change in preference as a function of only elapsing time has also been widely observed—in animals (Ainslie & Herrnstein, 1981; Green et.al., 1981) and in people's choices between sensual rewards like fruit juice (Forzano & Logue, 1992), process rewards like access to video games (Millar & Navarick 1984), negative reinforcers like relief from noxious noise (Navarick, 1982; Solnick et.al., 1980) and token rewards like money, both hypothetical and real (Ainslie & Haendel, 1983; Green, Fristoe, et.al., 1994; Kirby, 1997). The animal findings are important, for they let us be sure that the phenomenon is not the product of cultural expectations or experimenter suggestion.

Figure 1A



Conventional (exponential) discount curves from a smaller-sooner (SS) and a larger-later (LL) reward. At every point their heights stay proportional to their values at the time that the SS reward is due.

Figure 1B



Hyperbolic discount curves from an SS and an LL reward.. The smaller reward is temporarily preferred for a period just before it's available, as shown by the portion of its curve that projects above that from the later, larger reward.

3.1. The self as a population

Hyperbolic discounting offers utility theory a rationale for why people should so frequently have impulses that contradict their own recognized best interests. These highly bowed curves shift the main problem. We're no longer at a loss to explain choices that are short-sighted and temporary; now we have to account for how people learn the self-control that lets them adapt to a competitive world. How does an internal marketplace that disproportionately values immediate rewards grow into what can be mistaken for the long range reward-maximizer of conventional utility theory?

We can no longer regard people as having unitary preferences. Rather people may have a variety of contradictory preferences that become dominant at different points because of their timing. The orderly internal marketplace pictured by conventional utility theory becomes a bazaar of partially incompatible factions, where in order to prevail an option has not only to promise more than its competitors, but to act strategically to keep the competitors from later undermining it. The behaviors that are shaped by the competing rewards must deal not only with obstacles to getting their reward if chosen, but with the danger of being un-chosen in favor of imminent alternatives.

An agent who discounts reward hyperbolically is not the straightforward value estimator that an exponential discounter is supposed to be. Rather she will be a succession of estimators whose conclusions differ; as time elapses these estimators shift their relationship with one another from cooperation on a common goal to competition for mutually exclusive goals. Ulysses planning for the Sirens must treat Ulysses hearing them as a separate person, whom he must influence if possible and forestall if not. If what you do in a situation regularly gets undone later, you'll learn to stop doing it in the first place-- but not out of agreement with the later self that undoes it, only out of realism. Meanwhile, you'll look for steps toward getting what you want from the earlier vantage point, steps that won't get undone, because they forestall a future self who will try to undo them. You'll be like a group of people rather than a single individual; subjectively, however, the results of learning to do this may feel like no more than having to plan for self-control.

This lability of preference in turn predicts that a population of conflicting reward-getting processes will grow and survive within the individual, sometimes leading to choices that are harmful to her in the long run (first elaborated in Ainslie, 1975; detailed in Ainslie, 1992, pp. 123-227). I will call the processes selected for by a particular kind of reward the person's *interest* in that reward: Interests based on rewards within the person should be very like interests based on goals within a society, those factions that are rewarded by ("have an interest in") the goal that names them (e.g. a sobriety interest or drinking interest within the person, like "the petroleum interest," or "the arts interest" within a society). Since a person's purposes should still be coherent where conflicting rewards don’t dominate at successive times, it makes sense to name an interest only in cases of conflict. I wouldn't be said to have separate chocolate and vanilla ice cream interests, even though they're often alternatives, because at the time when I prefer chocolate I don't increase my prospective reward by forestalling a possible switch to vanilla. But I may have an ice cream interest and a diet interest, such that each increases prospective reward in its own time range by reducing the likelihood of the other's subsequent dominance. Put another way, I don't increase my prospective reward in either the long or short range by defending my choice of chocolate against the possibility that I may change to vanilla; but I increase my prospective long range reward by defending my diet against ice cream, and I increase my prospective short range reward by finding evasions of my diet for the sake of ice cream. Whichever faction promises the greatest discounted reward at a given moment gets to decide my move at that moment; the sequence of moves over time determines which faction ultimately gets its way.

Where the alternative rewards are available at different times, each will build its own interest. Such interests are not options chosen by an overarching ego, the *top down* model assumed by holistic theorists, but rather function as quasi-independent agents that have grown to exploit particular sources of reward over particular time ranges. In this *bottom up* model, an interest survives by realizing more expected, discounted reward than rival interests, which sometimes entails finding ways to actively forestall rival interests that would otherwise turn the tables when they became dominant in the future. If my diet interest can arrange for me not to get too close to ice cream, the discounted prospect of ice cream may never rise above the discounted prospect of the rewards for dieting, and the diet interest will effectively have won. However, whenever the value of ice cream spikes above that of dieting, the ice cream interest may undo the effect of many days of restraint.

The ultimate determinant of a person's choice is not her simple preference, any more than the determinant of whether a piece of legislation becomes law is simple voting strength in a legislature; in both situations, strategy is the critical factor. Analysis of this kind of strategy will require an economics of the internal marketplace, a micro-microeconomics (thus, picoeconomics—Ainslie, 1986, 1992) that evaluates the game-theoretic value of the options available to each interest. The target book lays out the rudiments of such an economics.

4. The Warp Can Create Involuntary Behaviors: Pains, Hungers, Emotions

Since we try to identify a set of consistent behaviors as "our own," we will be uncomfortable with the perception that our preferences intrinsically change. The least deniable change occurs with the impulsive actions that could be called deliberate. When we go on a binge or spending spree or even when we have a brief lapse in an intention not to smoke—preference reversals that last from seconds to days—we experience them as decisions. Even here, however, we may not feel fully responsible. An alcoholic learns that she is "helpless against alcohol," and impulses are often personified as alien forces: "The devil made me do it." Thus it is natural to ask whether preferences that have other durations, longer or shorter than that of the deliberate lapse, might underlie processes that are experienced as involuntary. The discussion in the rest of this chapter is not necessary for examining the mechanism of will *per se*, but will be important in our subsequent examination of the will's limitations.

There are long-lasting preferences that nevertheless feel like prisons-- anorexia nervosa, obsessive-compulsive personality disorder, and narrowness of character generally, which are complex in that they are themselves enforced by some kind of self-control; I'll discuss them in chapter 9. At the other end of the scale of durations, there are processes that usually feel involuntary but nevertheless have incentive value, positive or negative: brief "irresistible" urges like tics, emotions-- including the emotion-like component of pain that makes it aversive (Melzack & Casey, 1970)-- hungers, and much of what directs our attention. Many of these processes are innately programmed, so that a given stimulus leads to an invariant response. Even with relatively malleable processes like emotions, a person is not a Lockean blank slate, but has inborn dispositions to respond to particular stimuli in particular ways, for instance with fear to the appearance of being at a great height (Rader et al., 1980). There are, as it were, grooves in the slate, into which the chalk of behavior tends to fall.

However, predisposed responses can still be modified. A neutral stimulus that precedes being at a great height may come to induce fear, or repeated experience with being at a great height may cause it to stop inducing fear. Since these changes are usually involuntary, conventional theory attributes their selection not to the same kind of reward that selects voluntary choices but to an altogether different selective principle, "classical" conditioning. If a stimulus that can call up a slate with particular grooves regularly follows a new stimulus, that new stimulus is said to acquire the ability to call up the same slate. The trouble with this theory is that they are not exactly the same set of grooves—on close examination "conditioned" responses differ in detail from their parent responses (Siegal, 1983), so they must be shaped by some additional selective principle, a third one if it is not the same one that governs choice. And the gist of later conditioning research has been that conditioning does not control responses at all; the pairing of

stimuli connects only the stimuli, not responses (Rescorla, 1988). Conditioning theory is awkward also in several other ways that there is not room to discuss (see Ainslie, 2001, pp. 100 - 114, 1992, pp. 19-22). Since all stimuli that can cause conditioning also have an incentive value (Gerall & Obrist, 1962; Miller, 1969) and conditioning has been successfully modeled on computers as incentive-dependent (Donahoe et al., 1993, 1997), it is worth asking whether this second selective principle can be boiled down to the first—reward. That is, those involuntary responses that are malleable might be modified not by being transferred to new sets of grooves, but by being drawn out of the original grooves by reward whenever the reward is strong enough and the groove shallow enough. In different imagery, these responses would not be *pushed* by trigger stimuli but *pulled* by incentives.

There has always been one massive obstacle to this suggestion—not that these choices are mostly unconscious, for unconscious shaping of behavior is well known (even during sleep—Granda & Hammack, 1961)—but that reward is thought of as attracting only desirable behaviors. How would we be pulled into experiences that we don't want? And if we can't be pulled, we must have to be pushed, presumably by conditioning.

Hyperbolic discount curves have already provided a way around this obstacle for the case of impulses, that is, in the middle of the continuum of preference durations: When a reward precedes a longer period of nonreward, it is often preferred when up close but avoided at a distance. The same kind of cycle can be discerned in itch-like activities, but the cycle length is shorter. Minor itches will abate if never scratched, and the motive to scratch them gets described as an urge rather than a desire, as does the motive to bite your nails, use speech mannerisms, and emit tics. These are voluntary behaviors and may be subject to strong momentary motivation, but people avoid them at a distance and often seek preventive treatments. This is the kind of behavior that Berridge and Robinson have described as "wanted" but not "liked" (1998), the exemplar of which is the electrical brain self-stimulation that a rat will perform to exhaustion once it has begun, but which it will not cross a cage floor to begin again. Berridge and co-workers have catalogued a number of these behaviors in people as well, including brain self-stimulation patients. They think of them as "nonhedonic," classically conditioned, even though these behaviors use muscles that are usually under voluntary control; however, a conditioning mechanism is unnecessary. Forty years ago the same pattern was created just by varying the rate of reward: Pigeons were shown to actively avoid being offered the option of doing poorly rewarded work for food, instead of simply not doing the work when offered (Zimmerman & Ferster, 1964, among others). The mere chance to work for food became aversive, even though the subjects did the work when it was offered—or rather *because* they did the work when it was offered. They avoided being pulled into an undesirable pattern of responses by short term rewards.

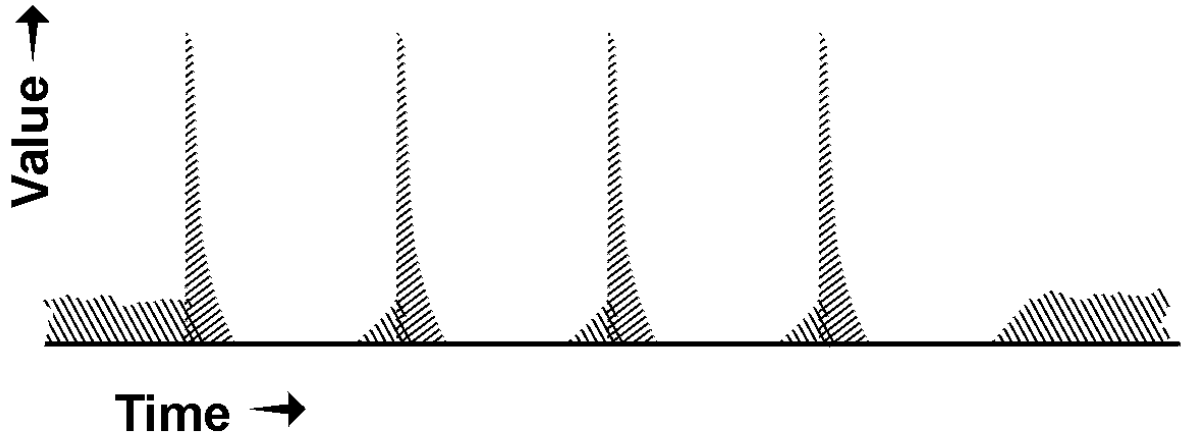
4.1. The problem of pain

An extension of the same cyclic mechanism may explain involuntary behaviors generally. Pain and painful emotions attract attention but deter approach. Pain can't be the simple opposite of reward that is often assumed, because it could not then oblige people to attend to it. The traditional solution to this problem is to treat pain like a reflex and fear like a conditioned reflex, processes that motivate but are not themselves motivated. But in addition to the difficulties just

mentioned with conditioning as a separate principle of selection, there are many indications that emotions and even the emotional part of pain are not automatic, but have to compete with rewarded activities for a person's participation. Granted that emotions are usually *occasioned* by events outside of your voluntary control; the theory that they are *governed* by such events runs afoul of the widespread acknowledgment that they are trainable: You can "swallow" your anger or "nurse" it, and learn to inhibit your phobic anxiety (Marks & Tobena, 1990), panic (Clum et.al., 1993; Kilic et.al., 1997) or grief (Ramsay, 1997). Pain itself registers in consciousness but is less apt to cause emotional aversion during the distraction of intense sports competition or battle than during daily life (Beecher, 1959, pp. 157-190), and less during daily life than when you're trying to go to sleep. Techniques to avoid aversion by distracting yourself are commonly taught for dental procedures and childbirth (Licklider, 1959), and may even cover major surgery in people with strong attention-focusing skills ("good hypnotic subjects"—Hilgard & Hilgard, 1994, pp. 86-165). Techniques to foster or inhibit emotions in everyday life have been described (Parrott, 1991), as has their use in preparing yourself for particular tasks (Parrott, 1993). Most schools of acting teach an ability to summon emotion deliberately (e.g. Strasberg, 1988; Downs, 1995), because even in actors actual emotion is more convincing than feigned emotion (Gosselin et.al., 1998). The frequent philosophical assertion that emotions have a moral quality—good or bad (e.g. Hume as presented by Baier, 1991)—implies motivated participation; some philosophers have gone so far as to call the passions voluntary (e.g. Sartre, 1939/1948). In sum, emotions show signs of being goal-directed processes that are ultimately selected by their consequences, not just their antecedents. That is, they are at least partially in the realm of motivated behaviors, not conditioned responses; they are *pulled* by incentives rather than *pushed* by stimuli. Even pain itself and "negative" emotions like fear and grief seem to be urges that lure you into participating in them, rather than being automatically imposed states.

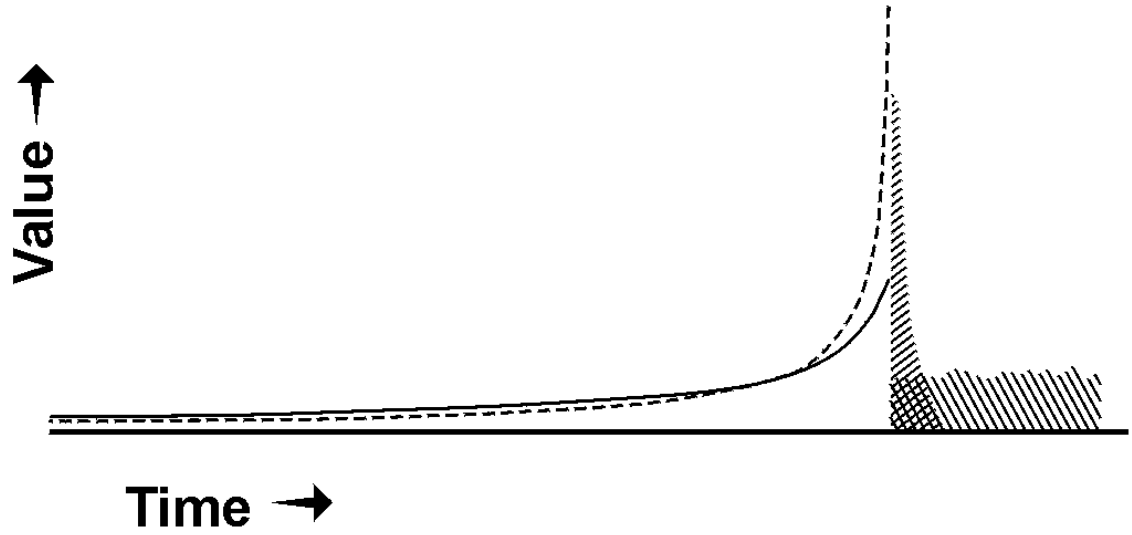
But we just saw that a cycle of reward and subsequent unreward can draw you into an activity which, at even a fairly slight distance, is aversive. A faster version of this cycle provides a model of how mental processes can be involuntary and still be reward-dependent, even if their overall pattern is aversive: If an itch is a fast addiction, maybe a pain is a fast itch. That is, perhaps the vividness but aversiveness of pain and negative emotions is a pattern of repeating, brief, intense reward, the occurrence of which causes an otherwise continuous nonreward (Figure 2A). Each reward is dominant so briefly that it can command only attention, not a motor response (Figure 2B), and the overall pattern motivates avoidance. Of course, for these two elements to fuse in perception, the cycle duration would have to be a fraction of a second.

Figure 2A



Aversion as a cycle of brief, intense reward (rightward hatching) that interrupts an ongoing baseline reward (leftward hatching) for a relatively longer time.

Figure 2B



Hyperbolic discount curves drawn from a single spike in an aversion sequence such as that in figure 2A. (Each curve is the sum of the curves from each moment of reward—see figure 4.) The spike has less area than the baseline reward to which it is an alternative; but because it's taller it will be preferred just before it's available.

In this way hyperbolic discounting has the power, in theory at least, to unite along a common dimension not only Berridge's liking and wanting but even action and passion.[1] This does require, however, that we strip "reward" of its connotations of pleasure, and leave it with a basic functional definition: "that which increases the likelihood that the processes it follows will recur." In return we are freed from dealing with two different selective principles for responses, which involve the same set of stimuli, but which differ in that one (seen as using classical conditioning) selects for both positive and negative processes and the other (seen as using reward) selects for the positive and against the negative. Rather, pain, emotions, and other "conditionable" processes—probably including appetite-- must all pay off quickly and repeatedly to attract participation; but great variance in the rewardingness of the longer phases between these payoffs determines how negative or positive their valence will be.

If emotions and similar processes are reward-dependent behaviors, a problem arises converse to the problem of pain: What keeps you from emitting the positive ones *ad lib*, in effect coining unlimited reward? I will address this problem in chapter 10.

A BREAKDOWN OF WILL: THE COMPONENTS OF INTERTEMPORAL BARGAINING

5. The Elementary Interaction of Interests

An interest that has survived in the marketplace of reward-getting strategies needs to have ways to forestall incompatible interests, at least well enough to sometimes get the reward on which this interest is based. This need accounts for the examples of self-committing tactics that have long puzzled utility theorists, who depict the person as a unitary reward maximizer with no reason to restrict her own freedom. Three kinds of tactic are straightforward: 1) finding constraints or influences outside of your psyche, sometimes physical devices like pills that spoil an appetite, or illiquid investments (Laibson, 1997), but more often the influence of other people; 2) keeping your attention off temptations, either consciously (Metcalfe & Mischel, 1999) or in the Freudian defense mechanisms of suppression, repression, or denial; and 3) cultivating or inhibiting emotions, either consciously (Mischel & Mischel, 1983) or in the defense mechanisms of isolation or reversal of affect. If an underlying, universal discount curve is hyperbolic in shape, a motive to self-commit should also be observable in nonhuman animals; and in fact it is: Given choices between SS (smaller, sooner) rewards and LL (larger, later) ones, nonhuman subjects will sometimes choose an option available in advance that prevents the SS alternative from becoming available (Ainslie, 1974; Hayes *et.al*, 1981). The converse is true of punishments—Rats will press a bar committing them to get .5 sec of shock 40 seconds later instead of 5 seconds of shock 45 seconds later, rather than leave the choice open and subsequently fail (almost always) to choose .5 seconds of imminent shock over 5 seconds of shock 5 seconds later (Deluty *et.al*, 1983).

However, these tactics are less adaptable, and often less available, than what is usually called willpower. Willpower represents a fourth tactic, which seems to be at once the strongest and most versatile, but which has hitherto been mysterious. What is there about "making a resolution" that adds anything to your power to resist changing motivations? When people have given up smoking or climbed out of debt they mostly say they "just did it." Words like volition,

personal rules, character, intention, and resolve are often applied, but don't suggest how people have learned to resist temporary preferences for shortsighted options.

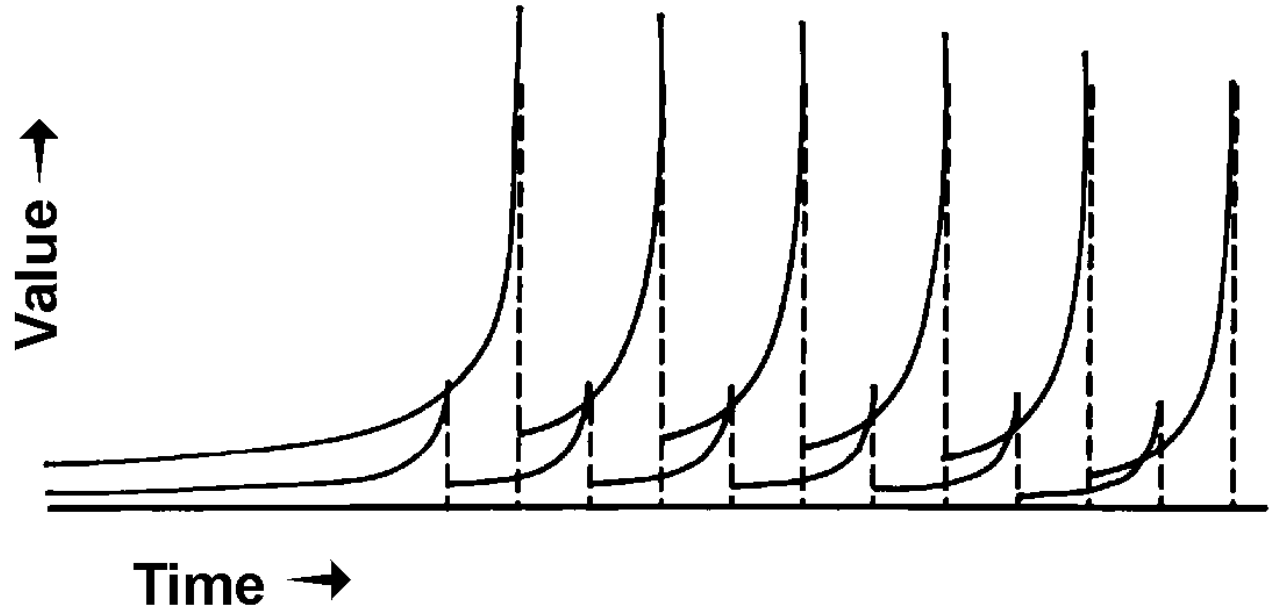
The specific property that has most often been attributed to the will is the perception of individual choices as referable to a larger principle. Writers since antiquity have recommended that impulses could be controlled by deciding according to principle, that is, deciding in categories containing a number of choices rather than just the choice at hand. Aristotle said that incontinence (*akrasia*) is the result of choosing according to "particulars" instead of "universals" (*Nichomachean Ethics* 1147a24-28); Kant said that the highest kind of decision-making involves making all choices as if they defined universal rules (the "categorical imperative," 1993/1960, pp. 15-49); the Victorian psychologist Sully said that will consists of uniting "particular actions... under a common rule" so that "they are viewed as members of a class of actions subserving one comprehensive end" (1884, p. 663). In recent years behavioral psychologists Heyman (1996) and Rachlin (1995) have both suggested that choosing in an "overall" or "molar" pattern (respectively) will approach reward-maximizing more than a "local" or "molecular" one.

Hyperbolic discounting suggests a workable rationale for choosing according to principle, albeit one that requires a degree of self-awareness probably unavailable to nonhumans: Insofar as you interpret your current choice as information predicting your own future choices between similar rewards, the incentives bearing on your current choice will to some extent include the bundle of future rewards that this choice predicts. That is, the current choice of a larger, later (LL) reward over a smaller sooner (SS) reward, if perceived as a *test case*, will come to predict a whole bundle of LL rewards in the future, and thus be valued more than it would be by itself. There is experimental evidence in animals showing that the hyperbolically discounted effects of each reward in a series simply add (analyzed in Mazur, 1997). More importantly, because hyperbolic curves are relatively high at long delays, bundling rewards together predicts an increase in the hyperbolically discounted value of the LL rewards relative to the hyperbolically discounted value of the SS rewards. Thus a bundle of LL rewards may be consistently worth more than a bundle of SS ones even where the discounted value of the most imminent smaller reward greatly exceeds the discounted value of its LL alternative (figure 3A).

Experiments in both humans and rats have verified the predicted anti-impulsive effect of bundling choices together. Kirby and Guastello reported that students who faced five weekly choices of a SS amount of money immediately or a LL amount one week later picked the LL amounts substantially more if they had to choose for all five weeks at once than if they chose individually each week (2001). The authors reported an even greater effect for SS vs. LL amounts of pizza. Ainslie and Monterosso reported that rats made more LL choices when they chose for a bundle three trials all at once than when they chose between the same SS vs. LL contingencies on each separate trial (2003). The effect of such bundling of choices is predicted by hyperbolic but not exponential curves: Exponentially discounted prospects do not change their relative values however many are summed together (Figure 3B); by contrast, hyperbolically discounted SS rewards, although disproportionately valued as they draw near, lose much of this differential value when choices are bundled into series. In Figure 3A,the schooner-like picture of the summed discount curves from series of rewards, the "sails" get gradually lower as the choice point moves later in the series, for they comprise a decreasing number of curves added together. The last pair of sails are the same as a lone pair. However, if the series has no foreseeable end,

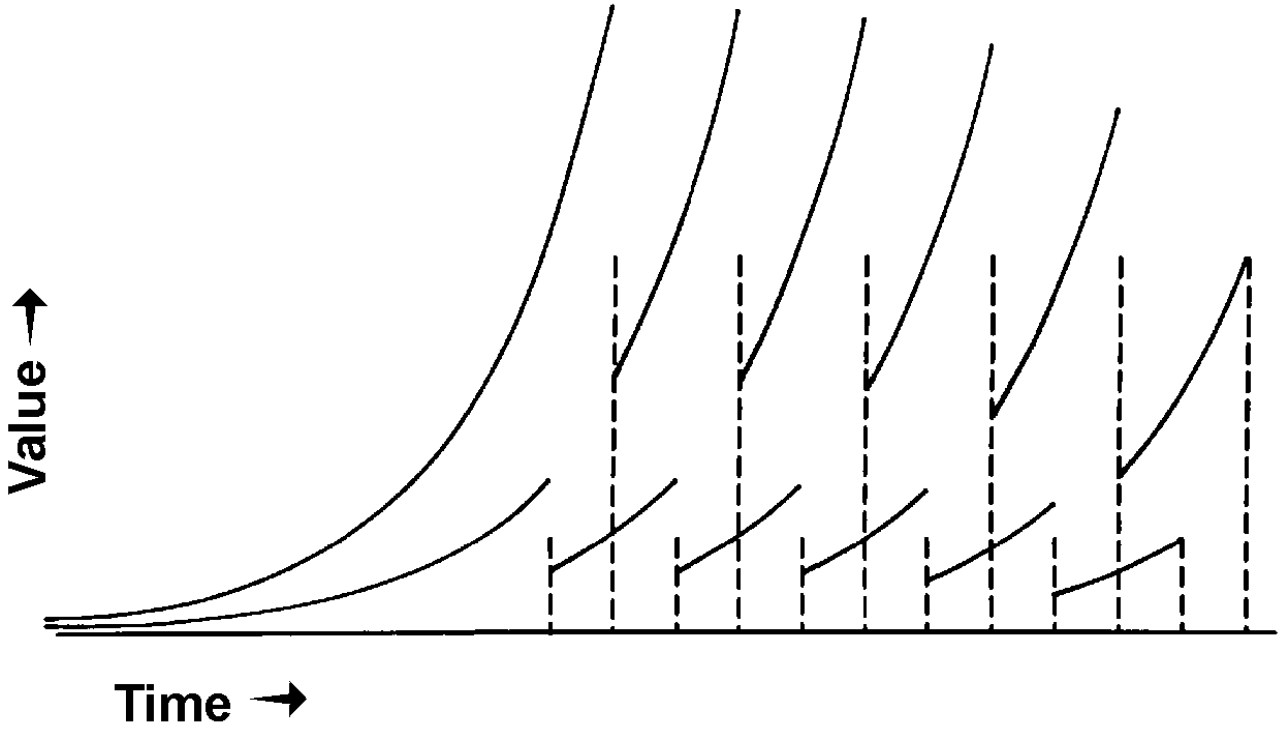
which is the case for most real-life categories, the sails may be added forward to a time horizon that stays a constant distance ahead, so that the height of the summed rewards stays roughly constant.

Figure 3A



Summed hyperbolic curves from a series of larger-later rewards and a series of smaller-earlier alternatives. As more pairs are added to the series, the periods of temporary preference for the series of smaller rewards shrink to zero. The curves from the final (rightmost) pair of rewards are the same as in figure 1B.

Figure 3B



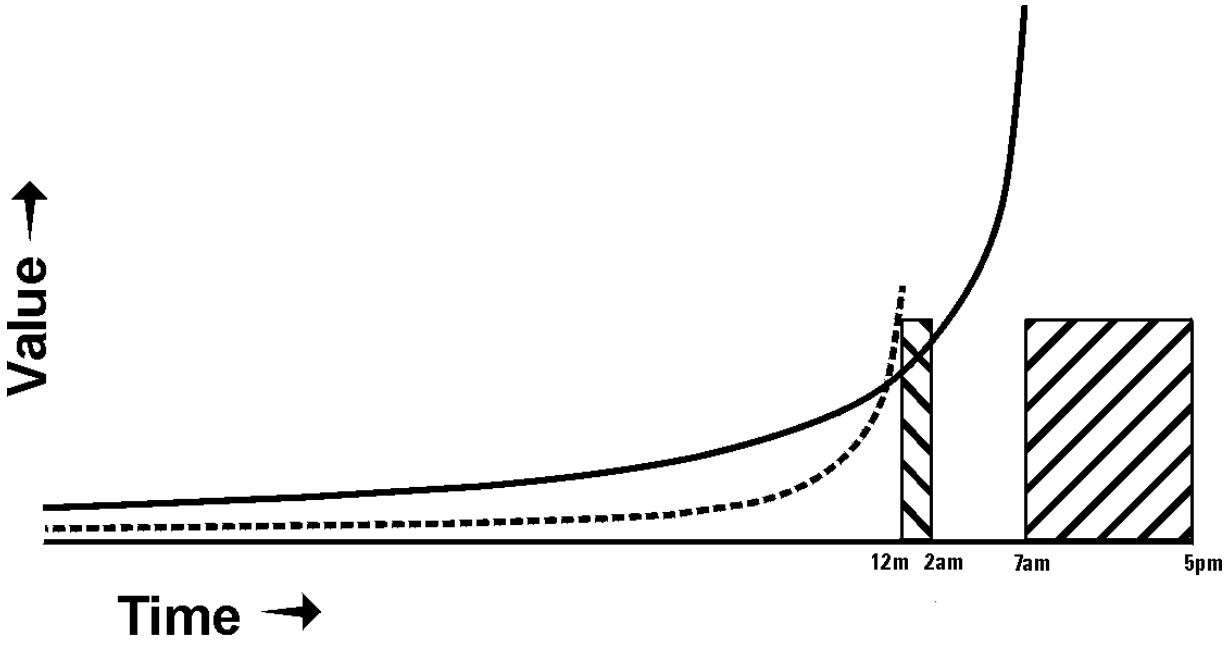
Summed exponential curves from the same series as in figure 3A. Summing doesn't change their relative heights. (This would also be true if the curves were so steep that the smaller, earlier rewards were preferred; but in that case summing would add little to their total height, anyway, because the tails of exponential curves are so low.)

But how does an individual arrange to bundle expected rewards together? This is where human perceptiveness is needed. Consider philosopher Michael Bratman's example of a pianist who throws his nightly performance off by drinking wine beforehand (1999, pp. 35-57). At a distance he prefers to abstain and perform well, but each night at dinnertime he changes his preference to drinking the wine. However, as Figure 3A suggests, even at dinnertime he may prefer abstaining all nights to drinking all nights for the foreseeable future. The incentives for choosing between these categories of reward will be the expected values of the series of rewards. The incentives for choosing just for one night will be just the curves from a lone pair, as in figure 1B. But if he perceives that his choice tonight is the best current predictor of what his future choices will be, he bundles his expectations together by that perception alone. Then if he has wine tonight, he sets a precedent, and sustains a greater expected loss than just tonight’s poor performance.

Most choices in real life aren't between momentary rewards, but between extended experiences-- the pleasure of a binge vs. feeling fit and having intact prospects Monday morning, or a good venting of rage vs. keeping a job and friends. Often the difference isn't between intensities of

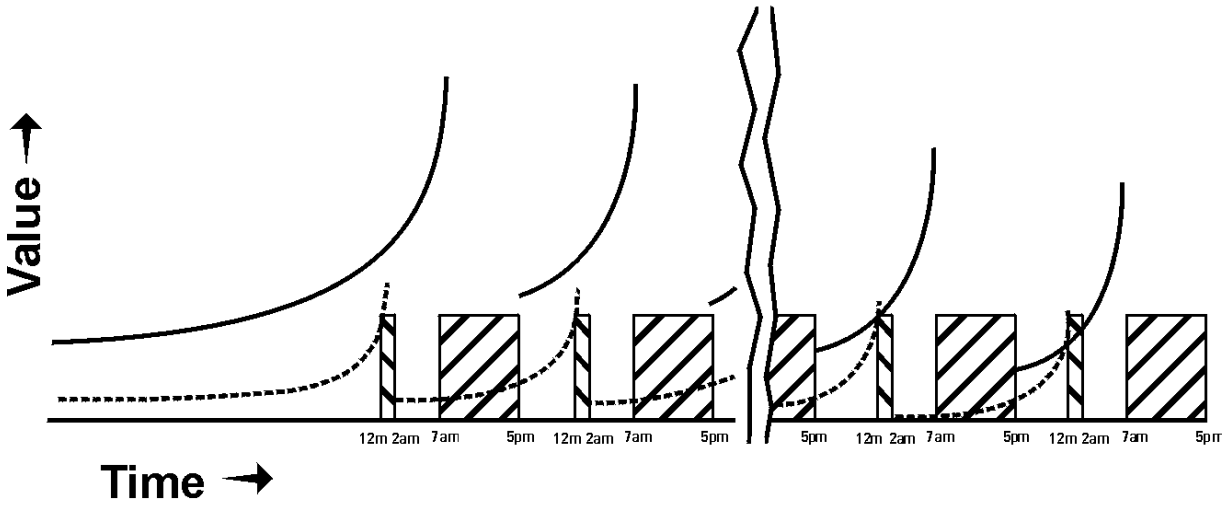
satisfaction-per-minute, but between different durations of comparable satisfactions. The pleasure of staying up for a couple more hours after midnight may be the same as the differential pleasure of feeling alert the next day, for instance, but the alertness lasts all day. However, if successive rewards are additive, it's easy to convert durations to total amounts (simple arithmetic derivations in Ainslie, 1992, pp. 155-162). If you value the fun of staying up at one unit per minute and expect to lose one unit per minute of comfort from when you get up at 7:00 the next morning until you leave work at 17:00, your discount curves from a day's aggregation of these rewards will look like those in Figure 4A. But if you see each night as a test case, your expectations will be bundled as in Figure 4B. As with more discrete moments of reward, bundling these experiences into series moves preferability toward the larger, later rewards.

Figure 4A



Curves that are the aggregate of hyperbolic discount curves from continuing rewards-- staying up from midnight to 2:00 vs. feeling rested from 7:00 to 17:00.

Figure 4B

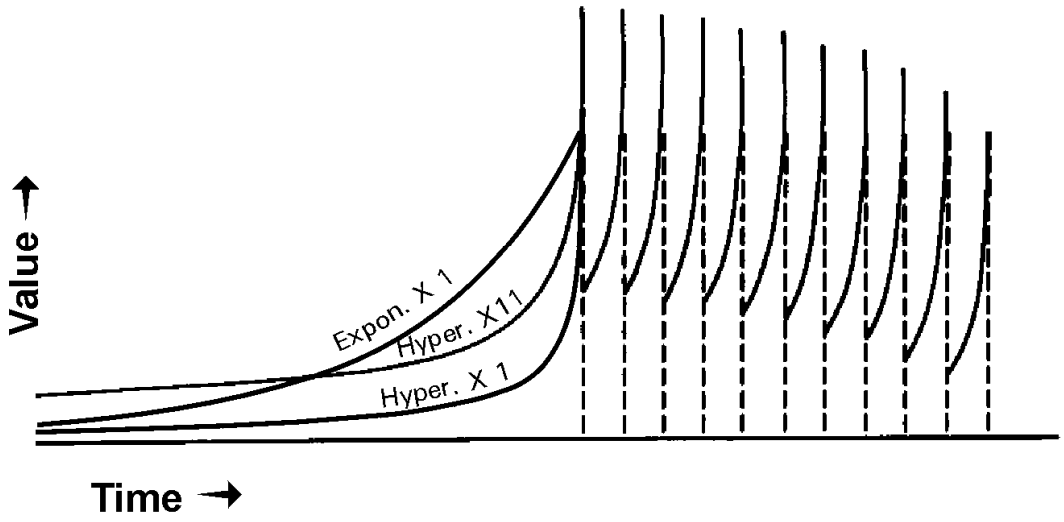


Summed curves from ten pairs of the rewards depicted in 4A. The effect of summation is the same as for the point rewards in Figure 3A.

6. Sophisticated Bargaining Among Internal Interests

The bundling phenomenon implies that you will serve your long range interest if you can obey a *personal rule* to behave alike toward all the members of a category. This is the equivalent of Kant's categorical imperative, and echoes the psychologist Lawrence Kohlberg's sixth and highest principle of moral reasoning, deciding according to principle (1963). It also explains how people with fundamentally hyperbolic discount curves may sometimes learn to choose as if their curves were exponential. Bundling whole series of choices together makes their summed discount curve look more exponential, as shown in Figure 5. Furthermore, if you adopt a personal rule to "discount all significant income at 6% per year," summed hyperbolic curves from all the expected amounts might be enough to motivate obedience to it, even though the shape of your summed curves did not approach this exponential curve closely. Summed hyperbolic curves from whatever goods accrued from the whole practice of exponential discounting might motivate rates of 3%, or any other rate including 0%; but the lower the rate to be enforced, the more vulnerable the rule would be to the lure of SS rewards.

Figure 5



Summed hyperbolic discount curves from eleven rewards, compared with a single exponential curve and a single hyperbolic curve such as were shown in figure 1. The summed curves come closer to exponential discounting than the lone hyperbolic curve does in the crucial sector of the curve where delay is low.

The problem with the bundling tactic is that there are many possible personal rules. The ice cream at hand may violate one diet but not another; and even if it's so outrageously rich as to violate all conceivable diets, there's apt to be a circumstance that makes the present moment an exception: It's Thanksgiving dinner or my birthday, or a host has taken special trouble to get it, or I have cause to celebrate or to console myself just today, etc. The molar principle that offers an exception just this once will be rewarded more than the one that doesn't, for it predicts the aggregation of LL rewards (as in figures #3A and #4B) for all but the first LL reward, *and* the first early spike of SS reward.

The possibility of seizing immediate rewards while protecting your expectation of later bundles-- by discerning exceptions-- makes the self-prediction upon which will depends potentially volatile, especially where self-control is tenuous. The hyperbolic discounter with an overeating problem can't simply estimate whether she's better off limiting her food intake or eating spontaneously, and then follow the best course, the way an exponential discounter could. Even if

she figures, from the perspective of distance, that dieting is better, her long range perspective will be useless to her unless she can avoid too many rationalizations. Her diet will succeed only insofar as she thinks that each act of compliance will be both necessary and effective-- that is, that she can't get away with cheating, and that her current compliance will give her enough reason not to cheat subsequently. The more she is doubtful of success, the more likely it will be that a single violation will make her lose this expectation and wreck her diet. Personal rules are a *recursive* mechanism; they continually take their own pulse, and if they feel it falter, that very fact will cause further faltering.

In this model deciding according to molar principles is not a matter of making dispassionate judgments, but of defending one way of counting your prospects against alternative ways that are also strongly motivated. Here the modified utility theory that I am proposing differs radically from a conventional top-down theory. In a top-down theory, the dieter, or pianist, does not need to predict her future choices because she (her ego, or other executive organ) can will them, and if her will is “strong” enough it will do just what she currently intends. But if, by contrast, choice is determined in a marketplace of competing interests, “she” is just the resultant of their activities, and stable choice has to be achieved as it is in the kind of markets that don’t have governors. The rules of this market are the internal equivalent of the "self-enforcing contracts" made by traders who will be dealing with each other repeatedly, contracts that let them do business on the strength of handshakes (Macaulay, 1963; Klein and Leffler, 1981). This recursive process of staking the credibility of a resolution on each occasion when it's tested gives your resolve momentum over successive times. The ongoing temptation to risk a damaging precedent-- and the ever-present anxiety that this may happen-- is probably what makes this strategy of self-control feel effortful. It separates intentions from expectations, and force of will from mere force of habit.

6.1. Intertemporal bargaining

Hyperbolic discount curves create a relationship of partial cooperation (*limited warfare*— Schelling, 1960, pp. 21-80) among your successive motivational states. Their individual interests in short range rewards, conflicting with their common interest in longer range rewards, create incentives much like those in the much studied bargaining game, repeated prisoner's dilemma. Choice of the better long range alternative at each point represents "cooperation," but this will look better than impulsive "defection" only as long as you see it as necessary and sufficient to maintain your expectation that future selves will go on cooperating. This is a useful way of modeling the will-- like the "will" of nations not to start a nuclear war-- rather than a cognitive hierarchy of some kind, but it needs to be modified for the intertemporal case. As Bratman has correctly argued (1999, pp. 35-57), a present "person-stage" can't retaliate against the defection of a prior one, a difference that disqualifies the prisoner's dilemma in its classical form as a rationale for consistency. However, insofar as a failure to cooperate will induce future failures, a current decision-maker contemplating defection faces a danger of the same kind as retaliation.

Intertemporal cooperation is most threatened by rationalizations that permit exceptions for the choice at hand, and is most stabilized by finding *bright lines* to serve as criteria for what constitutes cooperation. A personal rule never to drink alcohol, for instance, is more stable than

a rule to have only two drinks a day, because the line between some drinking and no drinking is unique (bright), while the two-drinks rule does not stand out from some other number, or define the size of the drinks, and is thus susceptible to reformulation. However, skill at intertemporal bargaining will let you attain more flexibility by using lines that are less bright. This skill is apt to be a key component of the control processes that get called ego functions.

This model proceeds from hyperbolic discounting with almost no extra assumptions-- only rough additiveness-- and predicts credible weapons for each side in the closely fought contests that occur as people decide about self-control: Long range interests define principles, and short range interests find exceptions.

7. The Subjective Experience Of Intertemporal Bargaining

Analyzing an activity that is second nature inevitably enlarges some features and slights others, so that the resulting picture seems foreign to familiar experience. Characterizing will as intertemporal bargaining may make it sound more deliberate, more effortful, and more momentous than casual introspection tells us our wills are:

1. Bargaining is usually thought of as requiring explicit consciousness of its contingencies; but the tacit bargaining that I have hypothesized as the basis of will may appear in a number of guises-- from prayers addressed to supernatural powers to beliefs in the factuality of propositions that are actually personal rules, guises that by chance or design conceal the active nature of our participation.
2. Bargaining might be thought to require continual re-evaluation; but bargaining may have its most important effect by establishing and only occasionally testing a dominance hierarchy of interests, just as social groups establish pecking orders that become habits.
3. The most conspicuous examples of bargaining stake huge incentives on all-or-none choices, such as when a recovering addict faces an urge to lapse; but resolutions like keeping your house neat can be mundane and largely based on intrinsic incentives while still having a recursive component. The only faculty you need in order to recruit the extra motivation that forms willpower is an awareness that your current decisions predict the pattern of your future decisions.

8. Getting Evidence About A Nonlinear Motivational System

If we conceive of the will broadly as whatever intentionality has some kind of force, it is possible to find five distinct models of it in the literature of motivational science. These models come from widely different intellectual traditions and often leave mechanisms unspecified, but they can be compared at least in their positions on whether or how extra motivation is recruited for impulse control:

- The *null* theory holds that there is no extra motivation, and that will is therefore a superfluous concept (e.g. Ryle, 1949/1984; Becker & Murphy, 1988). This theory seems to be based only on the absence of a rationale for will in an exponential system.
- The *organ* theory holds that the will is characterizable as strong or weak in general and directed rather like a muscle by an independent intelligence (e.g. Baumeister & Heatherton, 1996). The principal problem with this kind of model is it has to be guided by some evaluation process outside of motivation, since it has to act counter to the most strongly motivated choice at

the time. On what basis does this process choose? What keeps this strength from being co-opted by the bad option? Even granting a homunculus that governs from above, what lets a person's strength persist in one modality, say, smoking, when it has fallen flat in another such as overeating?

- The *resolute choice* theory holds that the will maximizes conventional utility by a rational avoidance of reconsidering plans (e.g.Bratman, 1999; McClennen, 1990). By this avoidance the proponents in the philosophy of mind may mean diversion of your attention, the second committing device I mentioned in chapter 5; but this would be effective only against brief urges like pain or panic, not against addictions (McConkey, 1984), the urge for which forces a re-evaluation over the hours or days that the diversion must be maintained. However, the philosophers may mean a more complex mechanism: McClennen refers to "a sense of commitment" to previously made plans (1990, pp. 157-161), and Bratman refers to "a planning agent's concern with how she will see her present decision at plan's end" (1999, pp. 50-56), which suggests that self-prediction is a factor. Resolute choice may turn out to be another name for intertemporal bargaining.
- The *pattern-seeking* theory holds that the will consists of an appreciation of pattern that is intrinsically motivating, like that which makes whole symphony more rewarding than the sum of its parts (Rachlin, 1995). Thus a recovered addict might avoid lapses because of the aversiveness of spoiling her pattern of sobriety. However, this aesthetic factor does not seem robust enough; distaste is not how most people would describe temptations, even the temptations that they avoid.
- The *intertemporal bargaining* model that I have described holds that the perception of precedents recruits motivation against impulses by bundling together classes of choices between hyperbolically discounted rewards. It is the only model that explains both temporary preference and adequate incentive to overcome it from the properties of the rewards involved. However, because the mechanism is recursive, it is hard to study directly by controlled experiment. There has been suggestive evidence. For instance, when Kirby and Guastello compared separate and bundled choices in their college subjects they found an intermediate degree of self-control if they suggested to some of the separate-choice subjects that their current choice might be an indicator of what they would choose on subsequent occasions (2001). However, I argue that better evidence comes from its ability to resolve paradoxes of intentionality that have been distilled into thought experiments by the philosophers of mind. One example:

8.1 Kavka's Problem

A person is offered a large sum of money just to intend to drink an overwhelmingly noxious but harmless toxin. Once she has sincerely intended it, as verified by a hypothetical brain scan, she's free to collect the money and not actually drink the toxin (Kavka, 1983). Philosophical discussion has revolved around whether the person has any motive to actually drink the toxin once she has the money, and whether, foreseeing a lack of such motive, she can sincerely intend to drink it in the first place, even though she would drink it if that were still necessary to get the money.

Kavka's problem poses the question: Are the properties of intention such that a person can move it about effortlessly from moment to moment, the way she raises and lowers an arm; and if not, what factors constrain changes of intention? Wholly unconstrained changes would make

intention seem no different from momentary preference. The problem makes it clear that intention must include a forecast of whether you'll carry it out; but this would seem to make it impossible to intend to drink the toxin, since mere forecasting leaves the intention powerless against a sudden change of incentive, even one that's entirely predictable. In that case, Ulysses couldn't intend to sail past the Sirens unaided, and Kavka's subject couldn't intend to drink the toxin, since they couldn't expect to fulfill their intentions.

However, if will is an intertemporal bargaining situation, an answer is at hand: Intending is the classification of an act as a precedent for a series of similar acts, so that the person stakes the prospective value of this series-- perhaps, in the extreme, the value of all the fruits of all intentions whatsoever-- on performing the intended action in the case at hand. Thus the person could meaningfully intend to drink the toxin, but only because she couldn't subsequently change her mind with impunity.

If I resolve to painfully donate bone marrow to a friend with leukemia, but then renege, I haven't gotten away with stealing altruistic pleasure during the period that my resolution was in force-- My failure to go through with it has reduced the credibility of my intending, and hence the size of the tasks I can subsequently intend. My willpower has suffered an injury, perhaps a costly one. Thus Kavka's subject does have an incentive to follow her original intention once she has the money: preservation of the credibility of her will; whether this incentive is adequate to overcome the approaching noxiousness of the toxin doesn't matter for purposes of the illustration. Will, in short, is a bargaining situation, where credibility is power. How a person perceives this bargaining situation is the very thing that determines how consistently she'll act over time.

Kavka's contribution has been to create a conceptual irritant that can't be removed until we supply a piece that is missing from conventional assumptions about intention. The piece I suggest is credibility, the stake that you add to a mere plan to keep yourself from reneging on it. To add a piece like this may be cheating; I imagine that Kavka envisioned philosophers working with only the elements he gave. But the theoretical problem may not have been a Chinese puzzle with a hidden solution, but rather a card game that we have been playing without a full deck. The fact that an intertemporal bargaining model can fill out the deck provides empirical support for its role in will. Drinking the toxin is irrational under the null theory and resolute choice theory. The organ and pattern-seeking theories seem to make no prediction about it. Only intertemporal bargaining makes it affirmatively rational.

Solutions to two other philosophical problems are discussed in the target book, but can only be mentioned here:

- In the problem of freedom of will, the determination of your choice by your own recursive prediction of your future choices makes choice neither indeterminate nor a straightforward estimation of external incentives.
- In Newcomb's problem (Nozick, 1993, p. 41) a choice that is defined as a diagnostic act is arguably made into a causal act by the postulation of an omniscient diagnostician; then it resembles the precedent-setting choice in intertemporal bargaining that is both diagnostic and causal.

THE ULTIMATE BREAKDOWN OF WILL: NOTHING FAILS LIKE SUCCESS

9. The Downside Of Willpower

Unfortunately, a person's perception of the prisoner's dilemma relationship among her successive selves-- and the willpower that results from this perception-- can't simply cure the problem of temporary preference. Willpower may be the best way we know to stabilize choice, but the intertemporal bargaining model predicts that it will also have serious side effects, side effects that have in fact been observed by clinicians. Such bargaining doesn't let us choose our best prospects from moment to moment as true exponential discounting would. Rather it formalizes internal conflict, making some self-control problems better, but some worse.

These side effects need to be discussed. Where they're noticed at all, they generally aren't recognized as the consequence of using willpower. In a dangerous split of awareness, we tend to see willpower as an unmixed blessing that bears no relation to such abnormal symptoms as loss of emotional immediacy, abandonment of control in particular areas of behavior, blindness toward one's own motives, or decreased responsiveness to subtle rewards. I will argue that just these four distortions are to be expected to a greater or lesser extent from a reliance on personal rules. They may even go so far as to make a given person's willpower a net liability to her.

9.1. Rules overshadow goods-in-themselves

The perception of a choice as a precedent often makes it more important for its effect on future expectations than for the rewards that literally depend on it. When this is true, your choices will become detached from their immediate outcomes and take on an aloof, legalistic quality. You will have an impaired ability to live in the here-and-now, the loss of authenticity that existential philosophers complain of in modern society generally.

It's often hard to guess how you'll interpret a current choice when looking back on it. Did eating that sandwich violate your diet or not? Where there's ambiguity, cooperation with your future selves will be both rigid and unstable. Under the influence of an imminent reward you may claim an exception to a rule, but later think you fooled yourself, that is, you may see yourself as having had a lapse. Conversely, you may be cautious beyond what your long-range interest requires, for fear that you'll later see your choice as a lapse. Every lapse reduces your ability to follow a personal rule, and every observance reduces your ability not to. Errors in either direction impose costs that would never result from the exponential curves of conventional rationality, since those curves wouldn't make choice depend on recursive self-prediction in the first place.

9.2. Rules magnify lapses

When you violate a personal rule, the cost is a fall in your prospect of getting the long range rewards on which it was based. But this prospect is what you've been using to stake against the relevant impulses; a lapse suggests that your will is weak, a diagnosis that may act recursively to weaken your will. To save your expectation of controlling yourself generally, you'll be strongly

motivated to find a boundary line that excludes from your larger rule the kind of choice where your will failed. This means attributing the lapse to a particular aspect of your present situation, even though it will make self-control much more difficult when that aspect is present in the future. You may decide that you can't resist the urge to panic when speaking in public, or to lose your temper at incompetent clerks, or to stop a doughnut binge once begun. Your discrimination of this special area has a perverse effect, since within it you see only failure predicting further failure. If you no longer have the prospect that your rule will hold here, these urges may seem to command obedience automatically, without an intervening moment of choice. Such an area, where a person doesn't dare attempt efforts of will, could be called a lapse district, by analogy to the vice districts in which Victorian cities encapsulated the vice they couldn't suppress. Where the encapsulated impulses are clinically significant, a lapse district gets called a symptom-- for instance, a phobia, a dyscontrol, or a substance dependence.

Thus the perception of repeated prisoner's dilemmas stabilizes not only long range plans but lapses as well (Discussed further in Ainslie, 1992, pp. 193-197). Alternative models of self-control failure based on exhaustion of "strength" (Baumeister and Heatherton, 1996) or an opponent process (Polivy, 1998), do not account for regular failure that is specific to a particular circumstance.

9.3. *Rules motivate misperception*

Personal rules depend heavily on perception-- noticing and remembering your choices, the circumstances in which you made them, and their similarity to the circumstances of other choices. And since personal rules organize great amounts of motivation, they naturally create temptations for you to suborn the perception process. When a lapse is occurring or has occurred, it will often be in both your long and short range interests not to recognize that fact: Your short range interest is to keep the lapse from being detected so as not to invite attempts to stop it. Your long range interest is also at least partially to keep the lapse from being detected, because acknowledging that a lapse has occurred would lower the expectation of self-control that you need to stake against future impulses.

After a lapse, the long range interest is in the awkward position of a country which has threatened to go to war in a particular circumstance that has then occurred. The country wants to avoid war without destroying the credibility of its threat, and may therefore look for ways to be seen as not having detected the circumstance. Your long range interest will suffer if you catch yourself ignoring a lapse, but perhaps not if you can arrange to ignore it without catching yourself. This arrangement, too, must go undetected, which means that a successful process of ignoring must be among the many mental expedients that arise by trial and error-- the ones you keep simply because they make you feel better without your realizing why. As a result, money disappears despite a strict budget, and people who "eat like a bird" mysteriously gain weight.

9.4. *Rules may serve compulsions.*

The fact that a decision comes to be worth more as a precedent than it is in its own right doesn't necessarily imply that it's the wrong decision. On the contrary, you'd think from the logic of summing discount curves that judging choices in whole categories rather than by themselves

would have to improve your overall rate of reward (figures 3A, 4B). Cooperation in a repetitive prisoner's dilemma would have to serve the players' long range interests, or else they'd abandon it. How, then, can self-enforcing rules for intertemporal cooperation ever become prisons? Why should anyone ever conclude that she was trapped by her rules, and even hire a psychotherapist to free her from a "punitive superego?"

The likeliest answer is that in everyday life a person can discern many possible prisoner's dilemmas in a given situation; and the way of grouping choices that finally inspires intertemporal cooperation need not be the most productive: Personal rules operate most effectively on distinct, countable goals. Thus the ease of comparing all financial transactions lets the value of a sum of money fluctuate much less over time than, say, the value of an angry outburst, or of a night's sleep. The motivational impact of a series of moods has to be much less than that of an equally long series of cash purchases. When some personal rules are based on well-marked criteria, and criteria for richer alternative rules are harder to specify, the well-marked criteria may win out simply because they offer more stability to the corresponding personal rules. The personal rules of anorectics or misers are too strict to promise the greatest satisfaction in the long run, but their exactness makes them more enforceable than subtler rules that depend on judgment calls. Here is a mechanism for the disorders of overcontrol, which impair a person’s capacity for satisfaction but seem to be enforced by an insistent will. The exemplar is obsessive-compulsive personality disorder, “control freak” disease, which differs from the more itch-like obsessive-compulsive disorder (without the “personality”) particularly in that people who have it endorse its strictures and seek to sustain them rather than seeking to be cured of them (American Psychiatric Association, 1994, pp. 417-423, 669-673).

So cooperation among successive motivational states doesn't necessarily bring the most reward in the long run. The mechanics of policing this cooperation may produce the intrapsychic equivalent of regimentation, which will increase your efficiency at reward-getting in the categories you've defined, but reduce your sensitivity to less well-marked kinds of reward.[2]

9.5. Rationality is elusive

Both hyperbolic discounting and the personal rules that compensate for it have distorting effects. Therefore, there can be no hard and fast principle that people should follow to maximize their prospective reward. Thus "rationality" becomes an elusive concept. Insofar as it depends on personal rules demanding consistent valuation, rationality means being systematic, though only up to the point where the system seems to go too far and we look compulsive. Even short of frank compulsiveness, the systemization that lets rules recruit motivation most effectively may undermine our longest range interests:

The attempt to optimize our prospects with personal rules confronts us with the paradox of definition-- that to define a concept is to alter it, in this case toward something more formalized. If you conclude that you should maximize money you become a miser; if you rule that you should minimize your vulnerability to emotional influence, you develop the numbing insensitivity that clinicians have named alexithymia (Nemiah, 1977); if you conclude that you should minimize risk, you become obsessively careful; and so forth. The logic of rules may come to so overshadow your responsiveness to experience that your behavior becomes legalistic

and inefficient. A miser's strict rules for thrift make her too rigid to optimize her chances in a competitive market; even the minor confinement of a rule to maximize profit on a yearly basis undermines a financier's effectiveness (Malekzadeh & Nahavandi, 1987). Similarly, strict autonomy means shielding yourself against exploitation by others' ability to invoke your passions; but alexithymics can't use the richest strategy available for maximizing emotional reward, the cultivation of human relationships (Ainslie, 1995).

In this way people who depend on willpower for impulse control are in danger of being coerced by logic that doesn't serve what they themselves regard as their best interests. Concrete rules dominate subtle intuitions; and even though you have a sense that you'll regret having sold out to them, you face the immediate danger of succumbing to short range urges like addictions if you don't.

10. An Efficient Will Undermines Appetite

The value of willpower seems to be limited not only by these four side-effects but also by two ways in which rewards seduce attention when they are too imminent to be offset by even bundled long range rewards: in the generation of appetites (including emotion and pain) and in premature satiation. Appetites can sometimes by avoided by other forms of trained foresight, as I described in chapter 4; but they can't be willed away, which has probably contributed to the common impression that they don't depend on reward. I also argued in chapter 4 that the seduction of attention is how negative emotions impose themselves on people who don't want them. I'll now discuss the converse problem of what constrains *ad lib* self reward with positive emotions. The key concept is premature satiation, the other process that can't be controlled by will. The limitation of reward by premature satiation is key in turn to three other puzzles that have only begun to be addressed by utility theory, which I will discuss in chapter 11 under the headings of construction of fact, vicarious reward, and indirection.

10.1. The limitation of positive emotion puzzle

Emotional rewards of one kind or another seem to be a large part of most people's incentives. We may decide to climb mountains, or become an object of envy, or achieve moral purity, or perform any number of other feats that aren't necessary for our physical comfort. We could ignore these tasks without any obvious penalty; but we somehow become committed to them, occasionally to the point of dying for them.

However, emotional reward is physically independent of any particular turnkey in the environment, an inconsistency with conventional utility theory. To function as a reward according to that theory, a good has to be limited in supply or accessibility; if it's available unconditionally, as emotion is, it should never induce significant motivation to obtain it. As Adam Smith originally observed, this is just the reasoning that makes air have less market value than diamonds, although air is more necessary. To let rewarding emotions be seen as economic goods, utility theory has had to assume that they are unmotivated reflexes that must be released by conditioned stimuli. But we saw in chapter 4 that conditioning is a superfluous mechanism, that supposedly conditioned responses can be accounted for by the brief predominance of

hyperbolically discounted rewards—except for the deferred question of how nature prevents the liberal coining of self-reward. It is to that question that I now return.

10.1.1. Avoiding premature satiation. The strongest emotions do seem to require a sense of necessity, so that we experience them not as choices but responses to an external provocation. Although emotions are physically available, something makes them less intense in proportion as the occasion for them is arbitrary. To the extent that someone learns to access them at will, doing so makes them pale, mere daydreams. Even an actor needs to focus on appropriate occasions to bring them out with force. But what properties must an event have in order to serve as an occasion for emotion? The fact that there's no physical barrier opposing free access to emotions raises the question of how emotional experiences come to behave like economic goods that are in limited supply. That is, how do you come to feel as if you have them passively, as implied by their synonym, "passions?"

The basic question is, how does your own behavior become scarce? I'll divide it into two parts: Why would you want a behavior of yours to become scarce, that is, to limit your free access to it? And given that this is your wish, how can you make it scarce without making it physically unavailable?

All kinds of reward depend on a readiness for it that's used up as reward occurs and that can't be deliberately renewed. This readiness is the potential for appetite, sometimes called appetite itself, although “appetite” then does not differentiate between an actual arousal for consuming a reward (as in "stimulating your appetite" or “becoming emotional”) and the adequately deprived or rested state that makes this arousal possible. The distinction is not important here, since exhausting the aroused appetite also exhausts the potential for it, so I will speak merely of appetite.

The properties of appetites are often such that rapid consumption brings an earlier peak of reward but reduces the total amount of reward that the appetite makes possible, so that we have an amount-vs.-delay problem of the kind that was described in figure 1B. Where people-- or, presumably, any reward-governed organisms-- have free access to a reward that's more intense the faster it's consumed, they'll tend to consume it faster than they should if they were going to get the most reward over time from that appetite. In a conflict of consumption patterns between the long and pleasant versus the brief but even slightly more intense, an organism that discounts the future hyperbolically is primed to choose brief but intense.

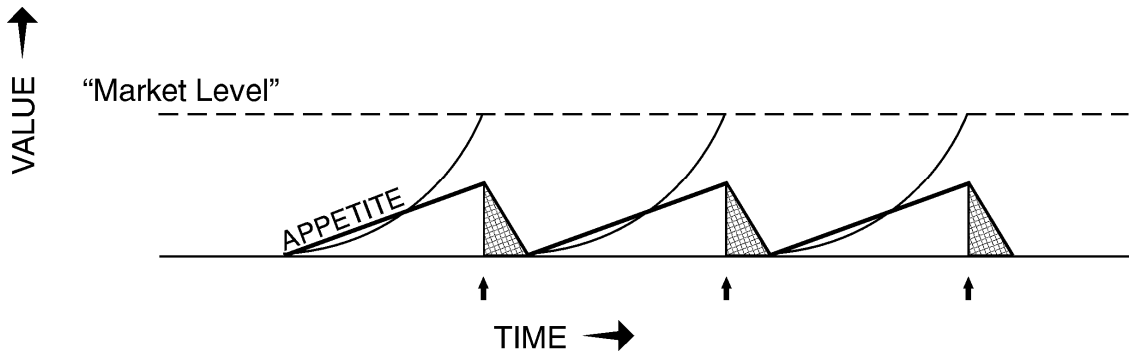
This problem makes no sense in a world of exponential discounting. In an exponential world, an adept consumer should simply gauge what the most productive way to exploit an appetite will be, and pace her consumption accordingly. People could sit in armchairs and entertain themselves optimally by waiting for just enough emotional appetite and then satisfying it. By contrast, common experience teaches that emotional reward, indulged in *ad lib*, becomes unsatisfactory for that reason itself. To get the most out of any kind of reward, we have to have-- or develop-- limited access to it.

Limiting access should be easiest for physical rewards: You can make a personal rule to consume them only in the presence of adequately rare criteria; but with emotional rewards, the

only way to stop your mind from rushing ahead is to avoid approaches that can be too well learned. Thus the most valuable occasions will be those that are either 1. uncertain to occur or 2. mysterious-- too complex or subtle to be fully anticipated, arguably the rationale of art. To get the most out of emotional reward, you have to either gamble on uncertainty or find routes that are certain but that won't become too efficient. In short, your occasions have to stay surprising-- a property that has also been reported as necessary for activity in brain reward centers (e.g. Hollerman et.al., 1998; Berns et.al., 2001).

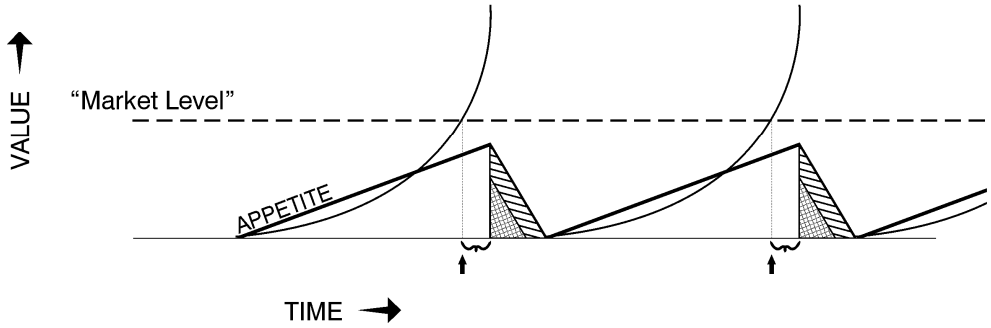
To restate this pivotal hypothesis: In the realm of emotional reward-- the great preponderance of the reward that even modestly well-off people pursue-- possible behaviors must compete on the basis of how well they can maintain your appetite. The processes that are rewarded by emotion compete for adoption on the basis of the extent to which their occasions defy willful control. Direct paths to reward become progressively less productive, because insofar as they become efficient they waste your readiness for reward. Conversely, if there's a factor that delays consumption from the moment at which the consumption could, if immediate, compete with available alternatives-- the moment it reaches what could be called the market level of reward-- that factor may substantially increase the product of [value x duration] before the appetite satiates. Figure 6 shows this using the simplest assumption that the build-up of potential appetite and the falling level of reward during consumption are linear over time; any concavity in the build-up or convexity in the consumption curve would accentuate the effect.

Figure 6A



Repeated cycles (not summed) of growing reward potential ("appetite," depicted schematically by the straight lines) and actual consumption to the point of satiety (gray areas). Consumption begins at the points (arrows) when discounted value of expected consumption reaches the competitive market level set by alternative sources of reward (which are not shown). Hyperbolic discount curves of the total value (the sum of the gray areas) of each act of consumption decline with delay from its anticipated onset (right to left as delay increases).

Figure 6B



Increased reward (striped areas) resulting from increased appetite when there is an obligatory delay from the moment of choice to the moment of starting consumption ("{" brackets); the choice to consume occurs at the points (arrows) when the discounted value of the delayed consumption reaches the market level. *Note that this figure was inaccurately drawn in the target book.*[3]

To repeat satisfactions that were once intense, you have at least to structure them as fantasies involving obstacles in order to achieve a modicum of suspense; but as a fantasy becomes familiar and your mind jumps ahead to the high points, the fantasy collapses further into being just a cursory thought-- an irritant if it retains any attractiveness at all, and a disregarded, empty option if it doesn't. Durable occasions for emotion have to be surprises, so that you don't have to restrain your attention from jumping ahead. Thus it's usually more rewarding to read a well-paced story than to improvise a fantasy, although even in fantasy some randomization is possible. Accordingly, surprise is sometimes said to be the basis of aesthetic value (Berlyne, 1974; Scitovsky, 1976). In modalities where you can mentally reward yourself, surprise is the only commodity that can be scarce.

Although there are wide variations in the equilibria people find between gratification at will and strict dependence on external occasions-- the fantasy-prone seem to have emotions that are more robust than other people's despite equally free access (Rhue & Lynn, 1987)-- everyone learns limits to her self-induction of emotions. Most people probably develop intuitions about how to foster sources of surprise, e.g. a rule not to read ahead, without ever making an explicit theory. People-- and presumably nonhuman animals-- wind up experiencing as emotion only those patterns that have escaped the habituation of voluntary access, by a selective process analogous to that described by Robert Frank for the social recognition of "authentic" emotions (1988): Expressions that are known to be intentionally controllable are disregarded, as with the false smile of the hypocrite. By this process of selection positive emotion is left with its familiar guise as passion, something that has to come over you. (The negative emotions habituate less, and need not be limited except by avoidance.)

It's undoubtedly adaptive for vivid rewards to fade away into habit as you get efficient at obtaining them; this process may keep you motivated to explore your environment, both when you're young and inept and when you've become a master problem-solver. If internal reward were strictly proportional to how much of some external stimulus you could get, then a reward rate that was sufficient to shape your behavior when you were a beginner would lead you to rest on your laurels once you'd become adept at getting it. But instead, as you become increasingly skilled in an activity, the reward it generates increases only at first, and then decreases again because your appetite doesn't last as long.

> The paradox is that it is just those achievements which are most solid, which work best, and which continue to work that excite and reward us least. The price of skill is the loss of the experience of value-- and of the zest for living (Tomkins, 1978, p. 212).

11. The Need To Maintain Appetite Eclipses The Will

I'll argue that the other three puzzles also hinge upon premature satiation, the impulse to harvest emotional reward before it’s ripe. Will not only cannot control this impulse, it may make you more vulnerable to it because of its demand for regular, distinct criteria for choice, the fourth side effect listed in chapter 9. The greatest limitation of the will comes from the same process as its greatest strength: its relentless systemization of experience through attention to precedent, which braces it against temporary preferences but also makes it unable to follow subtle strategies to overcome the premature satiation of emotional appetite.

11.1. *The construction of fact puzzle*

It's now common knowledge that people's beliefs are heavily influenced by their own tacit choices. Decisions about attending to or ignoring information shape perception so much that some "social constructivists" have put fact and fiction on a par, under the name "text" (e.g. Gergen, 1985; see Harland, 1987). To a great extent belief does seem to be a goal-seeking activity. However, it can't be based simply on rewardingness and still be experienced as belief. Belief differs from make-believe in depending on the ruling of some external arbiter, some test that's beyond your direct influence, rather than simply being chosen.

Instrumental beliefs, those shaped by external rewards, leave little room for construction. Construction can occur readily where the consequences of beliefs are emotional rather than externally determined, but the constraints on this process haven’t been explored. However, the pervasive urge for premature satiation discussed in the foregoing chapter is a likely limiting factor. This urge can be expected to create a selective process favoring emotions that are occasioned by adequately inaccessible texts, thereby promoting those texts to a status more significant than fantasy. That is, the premature satiation hypothesis predicts an incentive to cue emotions by something as inflexible as facts in order to optimize available appetite. Emotions tied to beliefs that can shift as convenience dictates will become daydreams, just like emotions that aren't tied to beliefs at all. The texts that get selected as beliefs for noninstrumental motives will be those interpretations of reality that serve as effective occasions for emotions. If they are adequately unique—the history everyone agrees upon, the answer that seems too hard to have

alternatives, the assumption you've held since childhood-- those texts have the feel of facts, and the recognition of their importance has the feel of belief.

According to this hypothesis, the very point of noninstrumental beliefs is to constrain the occasions for emotion. As with any mental process, the ultimate selective factor must be reward, but here long range rewardingness will depend on a balance between the production and restriction of reward; and since the production of emotions is not intrinsically limited, we learn to produce them when and only when there are adequate restrictions. Cues that have been selected on this basis as occasions for emotions become experienced as the facts that stimulate these emotions. For this purpose, accuracy *per se* will be only one selective factor for belief in a fact, and not an indispensable one at that.

11.2 The vicarious reward puzzle.

Other people are especially valuable as sources of emotional experience. Conventional utility theory calls this a simple putting-yourself-in-the-other's-place, and regards it as natural whenever "social distance" is short. This idea, first elaborated by Adam Smith (1759/1976), has been put into terms of utility by Julian Simon (1995). But the movingness of social experiences doesn't precisely depend on distance, or even on the existence of a real other person as opposed to a fictional character; and in many cases the experience that one person gets is obviously different from that of her vicarious object-- at the extreme, for sadist and victim. How do other people move us, and what are the constraints on that process?

There has been a lively debate between authors who believe that altruism is a primary motive (e.g. Batson & Shaw, 1991) and those who think it reduces to selfish pleasure (Piliavin et.al., 1982; Sen 1977). Economic man is supposed to maximize his own prospects, and help others only insofar as doing so will elicit reciprocity. However, you find counterexamples all the time, from transients who leave tips for waiters they'll never see again to heroes who give their lives to save strangers in fires and accidents. People also have the potential to derive satisfaction from others' pain-- even, in the extreme, from their death throes (e.g. Davies, 1981, pp. 78-82). Instrumentality again aside, what makes this range of perceived experiences in other people valuable to us?

The premature satiation hypothesis predicts that vicarious experience ought to be a good criterion for occasioning emotional reward, but should become less valuable to the extent that you can bring it under your control, because your control will inevitably undermine your appetite. Thus the greatest rewards from other people will come through gambles on their responses. But gambles that are rigged--interactions that are predictable, people you can boss around, relationships you're poised to leave if they turn disappointing-- push your emotional experiences in the direction of daydreams. These hedges are tantamount to exchanging a mutual game of cards for a game of solitaire, and perhaps even to cheating at solitaire; such an impulse is punished by a loss of suspense, and hence of all but fairly short range reward.

Given adequate appetite, the emotional payoff comes when the other person gives you a good occasion for emotion. Predicting other people becomes a highly rewarded activity for its emotion-occasioning value, quite aside from how it may help you influence them. However, this

is only part of the story. So far, there's no reason to think that gambling on other people's behavior would be any more rewarding than gambling on a horse race, or on your ability to solve a puzzle. The fact that this puzzle responds strategically to your choices might make it more challenging, but wouldn't qualitatively change the experience of succeeding or failing. But because this kind of puzzle is built like the person solving it-- that is, because it's another person-- it may foster what is likely to be a much richer strategy for occasioning emotions:

First, this similarity supplies a different way of solving the puzzle. Since other people's choices depend more on their interaction with you than on anything you know about them in advance, you soon learn that the best way to predict them is to use your own experience to model theirs. If the model isn't arbitrary-- if it's disciplined by observation-- it's apt to behave much more like the actual other person than a non-empathic model would, for instance one made like the model of an economy from statistical data. The best way to predict people is to put yourself in their shoes.

However, this empathic modeling process should yield more than just prediction. Putting yourself in the other person's shoes means adopting the criteria that you think she's using to occasion emotion. For the time being you entertain her emotions. But of course, they are hers only in the sense that you're having them according to a theory about her. *You* are the person through whose brain they're percolating. This means that you can use such a model to occasion emotions just as you use your own prospects.

Since emotions don't need a turnkey, just appetite and adequately rare occasions to preserve this appetite, you should be able to sometimes experience the emotions you're modeling in the other person as substantially as the ones you have as yourself. To model the other people is to have their expected feelings; and nothing makes these "vicarious" feelings differ in kind from "real" ones. The target book suggests a related rationale for the vicarious enjoyment of negative emotions (Ainslie, 2001, pp. 183-186). However, the emotional impact of these phenomena will be limited by the uniqueness of your relationship with the other person, just as the impact of texts is limited by their factuality; vicarious experiences from strangers picked for the purpose will be little more than daydreams.

To the extent that we've gambled on another person's discernable feelings, these feelings should become a good that we'll work for. Information about our gambles on other people will be the limited commodity that constrains the otherwise too-available resource of emotion. This, I argue, is how other people come to compete for our interest on the same footing as the goods of commerce.

11.3. The indirection puzzle.

Some goal-directed activities can't effectively approach their goals by direct routes. Trying to have fun usually spoils the fun, and trying to laugh inhibits laughter (Elster, 1981; Wegner, 1994). At first glance, this problem seems to strike at the heart of any motivational model, not just one that assumes exponential discounting. How can any goal-directed activity be undermined by striving toward its goal? How can a reward-dependent activity not be strengthened by reward?

I've described how the will can’t stop the premature satiation of suspense. I'll now argue that will may actually make premature satiation worse. The will needs conspicuous, discrete criteria of success or failure to maintain the incentive to cooperate with future selves at each choice-point. But systematically following well-defined criteria is exactly what makes your behavior predictable, by other people as well as yourself. It's a great way to achieve a goal as efficiently as possible, so that you can go on and do something else. It's a terrible way to enjoy an activity for its own sake, because it kills appetite. You inevitably learn to anticipate every step of the activity, so that it eventually becomes "second nature," making it so uninteresting that people used to think that ingrained habits were run by the spinal cord. You can't use will to prevent this anticipation, because clear criteria for rules directing attention aren't available, and even if they were, the attention required to test the choice would be the very behavior involved in the choice. So a too-powerful will tends to undermine its own motivational basis, creating a growing incentive to find evasions. The awkwardness of getting reward in a well-off society is that the creation of appetite often requires undoing the work of satisfying appetite.

Refreshing your emotional appetite without having to contradict what you've willed often requires believing in some seemingly rational, or arguably necessary, activity that is incompatible with the direct routes to a reward. That is, you need to find *indirect* routes to success: dummy activities that aren't actually worthwhile for their ostensible purpose, but stay desirable insofar as they maintain appetite by creating good gambles. In general you will need to believe in some larger quest that requires you to put your satisfaction at risk. To climb mountains or jump out of airplanes as a test of fortitude, to stay with an abusive lover to prove your loyalty, to join a religion that demands self-abasement, to play the stock market or the horses as a way to get rich, even to bet your dignity on staying in the forefront of fashion, leads to repeated losses or at least the credible threat of losses. You get your appetite back while struggling not to.

Activities that are spoiled by counting them, or counting on them, have to be undertaken through indirection if they are to stay valuable. For instance, romance undertaken for sex or even "to be loved" is thought of as crass, as are some of the most lucrative professions if undertaken for money, or performance art if done for effect. Too great an awareness of the motivational contingencies for sex, affection, money, or applause spoils the effort, and not only because it undeceives the other people involved. Beliefs about the intrinsic worth of these activities are valued beyond whatever accuracy these beliefs might have, because they promote the needed indirection.

12. Conclusions

Robust evidence has indicated that the basic function by which all vertebrates devalue delayed events is hyperbolic. Hyperbolic discounting has confronted conventional utility theory with the likelihood that it doesn’t describe elementary principles of choice, but represents a higher-order cultural invention that doesn't necessarily operate in all people or in all situations. Preferences that are temporary aren't aberrations any more, but the starting place for a strategic understanding of functions that used to be thought of as organs: the ego, the will, even the self.

Processes that pay off quickly tend to be temporarily preferred to richer but slower-paying processes, a phenomenon that can't be changed by insight per se. However, where people come to look at their current choices as predictors of what they will choose in the future, a logic much like that in the familiar bargaining game, repeated prisoner's dilemma, should recruit additional incentive to choose the richer processes. This mechanism predicts all the major properties that have been ascribed to both the power and freedom of the will. Further examination of this mechanism reveals how the will is apt to create its own distortion of objective valuation. Four predictions fit commonly observed motivational patterns: A choice may become more valuable as a precedent than as an event in itself, making people legalistic; signs that predict lapses tend to become self-confirming, leading to failures of will so intractable that they seem like symptoms of disease; there will be motivation not to recognize lapses, which creates an underworld much like the Freudian unconscious; and distinct boundaries will marshal motivation better than subtle boundaries, which impairs the ability of will-based strategies to exploit emotional rewards.

Furthermore, hyperbolic discounting suggests a distinction between short-lived reward and more durable pleasure that allows us to account for the often-observed seductiveness of pain and "negative" emotions. Conversely, the likelihood that this discounting pattern hastens our consumption of a reward where slower consumption would be richer explains why we seek external occasions for rewards that are otherwise at our disposal. The existence of both strong lures to entertain aversive mental processes and intrinsic constraints on freely available, pleasurable processes makes it possible to do without the hoary theory of classical conditioning. Instead: Emotions and hungers (together: "appetites") recur to the extent that they are rewarded. This means that the "conditioned stimuli" for appetites are not automatic triggers, but signs that emitting these appetites will be more rewarding, at least in the very short run, than not emitting them. These cues don't *release* appetites, they *occasion* them.

The urge to prematurely satisfy appetite teaches efficiency of reward-getting but brings about the decline of pleasures once they've become familiar. This problem provides a primary motive for the separation of belief from fantasy. Instrumental needs aside, beliefs determined by relatively rare events that are outside of your control are better occasions for feeling than your own arbitrary constructions, and hence come to be experienced as more meaningful. However, uniquely well-established social constructions may function about as well as objective facts in this regard. Similar logic explains the value of empathic interaction with other people, apart from any motives for practical cooperation: To gamble, in effect, on the experiences of others keeps your occasions for emotion surprising, and thus counteracts learned habituation.

Finally, there is an inevitable clash between two kinds of reward-getting strategies: Belief in the importance of appetite-satisfying tasks-- amassing wealth, controlling people, discovering knowledge itself-- leads to behaviors that rush to completion; but a tacit realization of the vulnerability of appetite motivates a search for obstacles to solutions, or for gambles that will intermittently undo them. Consciousness of the second task spoils the very belief in the first task that makes the first task strict enough to be an optimal pacer of reward. Thus the task of restoring appetite tends to be learned indirectly, and to be culturally transmitted via beliefs that seem superstitious or otherwise irrational to conventional utility analysis.

References

Ainslie, G. (1974) Impulse control in pigeons. *Journal of the Experimental Analysis of Behavior* 21, 485-489.

Ainslie, G. (1975) Specious reward: A behavioral theory of impulsiveness and impulse control. *Psychological Bulletin* 82, 463-496.

Ainslie, G. (1986) Beyond microeconomics: conflict among interest in a multiple self as a determinant of value. In J. Elster (Ed.), *The Multiple Self*. Cambridge University Press, pp. 133-175.

Ainslie, G. (1992) *Picoeconomics: The Strategic Interaction of Successive Motivational States within the Person*. Cambridge U.

Ainslie, G. (1995) A utility-maximizing mechanism for vicarious reward: Comments on Julian Simons "Interpersonal allocation continuous with intertemporal allocation" *Rationality and Society* 7, 393-403.

Ainslie, G. (2001) *Breakdown of Will*. Cambridge U.

Ainslie, G. and Haendel, V. (1983) The motives of the will In E. Gottheil, K. Druley, T. Skodola, H. Waxman (Eds.), *Etiology Aspects of Alcohol and Drug Abuse*. Charles C. Thomas, pp. 119-140.

Ainslie, G. and Herrnstein, R. J. (1981) Preference reversal and delayed reinforcement. *Animal Learning and Behavior* 9,476-482.

Ainslie, G. and Monterosso, J. (2003) Building blocks of self-control: Increased tolerance for delay with bundled rewards. *Journal of the Experimental Analysis of Behavior* 79, 83-94.

American Psychiatric Association (1994) *Diagnostic and Statistical Manual of Mental Disorders*, Fourth Edition. APA Press.

Baier, A. (1991) *A Progress of Sentiments: Reflections on Hume's Treatise.* Harvard U.

Batson, C. D. and Shaw, L. L. (1991) Evidence for altruism: Toward a pluralism or prosocial motives. *Psychological Inquiry* 2, 159-168.

Baumeister, R. F. and Heatherton, T. (1996) Self-regulation failure: An overview. *Psychological Inquiry* 7, 1-15.

Becker, G. and Murphy, K. (1988) A theory of rational addiction. *Journal of Political Economy* 96, 675-700.

Beecher, H. (1959) *Measurement of Subjective Responses*. Oxford.

Berlyne, D.E. (1974) *Studies in the New Experimental Aesthetics*. Hemisphere.

Berns, G. S., McClure, S. M., Pagnoni, G., and Montague, P.R. (2001) Predictability Modulates Human Brain Response to Reward, *Journal Of Neuroscience* 21, 2793-2798

Berridge, K. C. and Robinson, T. (1998) What is the role of dopamine in reward: Hedonic impact, reward learning, or incentive salience: *Brain Research Reviews* 28, 309-369.

Bratman, M. E. (1999) Faces of Intention: Selected Essays on Intention and Agency. Cambridge U.

Chung, S. and Herrnstein, R. J. (1967) Choice and delay of reinforcement. *Journal of the Experimental Analysis of Behavior* 10, 67-74.

Clum, G. A., Clum, G. A., & Surls, R. (1993) A meta-analysis of treatments for panic disorder. *Journal of Consulting and Clinical Psychology* 61, 317-326.

Davies, N. (1981) *Human Sacrifice in History and Today*. Morrow.

Deluty, M.Z., Whitehouse, W.G., Mellitz, M., and Hineline, P.N.(1983) Self-control and commitment involving aversive events. *Behavior Analysis Letters* 3, 213-219.

Donahoe, J. W., Burgos, J. E., and Palmer, D. C. (1993) A selectionist approach to reinforcement. *Journal of the Experimental Analysis of Behavior* 60, 17-40.

Donahoe, J. W., Palmer, D. C., and Burgos, J. E. (1997) The S-R issue: Its status in behavior analysis and in Donahoe and Palmer's Learning and Complex Behavior. *Journal of the Experimental Analysis of Behavior* 67, 193-211.

Downs, D. (1995) *The Actor's Eye: Seeing and Being Seen.* Applause Theatre Books.

Elster, J. (1981) States that are essentially by-products. *Social Science Information* 20, 431-73. Reprinted in Elster, J. (1983) *Sour Grapes: Studies in the Subversion of Rationality*. Cambridge U., pp. 43-108.

Forzano, L. B. and Logue, A. L. (1992) Predictors of adult humans' self-control and impulsiveness for food reinforcers. *Appetite* 19, 33-47.

Frank, R. H. (1988) *Passions Within Reason*. Norton.

Gerall, A.A. and Obrist, P.A. (1962) Classical Conditioning of the Pupillary Dilation Response of Normal and Curarized Cats. *Journal of Experimental Psychology* 50, 261-263.

Gergen, K. J. (1985) The social constructionist movement in modern psychology. *American Psychologist* 40, 266-275.

Gosselin, P., Kirouac, G. and Dore, F. Y. (1998) Components and recognition of facial expression in the communication of emotion by actors. *Journal of Personality and Social Psychology* 68, 83-96.

Grace, R. C. (1994) A contextual model of concurrent chains choice. *Journal of the Experimental Analysis of Behavior* 61, 113-129.

Granda, A.M. and Hammack, J.T. (1961) Operant behavior duringsleep. Science 133, 1485-1486.

Green, L., Fisher, E.B., Jr., Perlow, S., and Sherman, L. (1981) Preference reversal and self-control: choice as a function of reward amount and delay. *Behaviour Analysis Letter* 1, 43-51.

Green, L., Fristoe, N., and Myerson, J. (1994) Temporal discounting and preference reversals in choice between delayed outcomes. *Psychonomic Bulletin & Review* 1, 386.

Green, L., Fry, A., and Myerson, J. (1994) Discounting of delayed rewards: A life-span comparison. *Psychological Science* 5, 33-36.

Harland, R. (1987) *Superstructuralism: The Philosophy of Structuralism and Post-Structuralism*. Methuen.

Harvey, C. M. (1994) The reasonableness of non-constant discounting *Journal of Public Economics* 53, 31-51.

Hayes, S.C., Kapust, J., Leonard, S.R., and Rosenfarb, I. (1981) Escape from freedom: Choosing not to choose in pigeons. *Journal of the Experimental Analysis of Behavior* 36, 1-7.

Herrnstein, R. J. & Prelec, D. (1992) A theory of addiction. In eds. G. F. Loewenstein & J.

Elster, *Choice over time*. Sage, 331-360.

Heyman, Gene M. (1996) Resolving the contradictions of addiction. *Behavioral and Brain Sciences* 19, 561-610.

Hilgard, E.R. and Hilgard, J.R. (1994) *Hypnosis in the Relief of Pain, Revised Edition*. New York, Burnner/Mazel.

Ho, M.-Y., Al-Zahrani, S.S.A., Al-Ruwaitea, A.S.A., Bradshaw,C.M., and Szabadi, E. (1998) 5-hydroxytryptamine and impulse control: Prospects for a behavioural analysis. .*Journal of Psychopharmacology* 12, 68-78.

Hollerman, J. R., Tremblay, L., and Schultz, W. (1998) Influence of reward expectation on behavior-related neuronal activity in primate striatum. *Journal of Neurophysiology* 80, 947-963.

Kant, I. (1793/1960) *Religion Within the Limits of Reason Alone* (T. Green and H. Hucken, Trans.). Harper and Row, pp. 15-49.

Kavka, G. (1983) The toxin puzzle. *Analysis* 43, 33-36.

Kilic, C., Noshirvani, H., Basoglu, M., & Marks, I. (1997) Agoraphobia and panic disorder: 3.5 years after alprazolam and/or exposure treatment. *Psychotherapy and Psychosomatics* 66, 175-178.

Kirby, K. N., and Guastello, B. (2001) Making choices in anticipation of similar future choices can increase self-control. *Journal of Experimental Psychology: Applied* 7, 154-164.

Kirby, Kris N. (1997) Bidding on the future: Evidence against normative discounting of delayed rewards. Journal of Experimental Psychology: General 126, 54-70.

Kirby, K. N. and Marakovic, N. N. (1995) Modeling myopic decisions: Evidence for hyuperbolic delay-discounting within subjects and amounts. *Organizational Behavior and Human Decision Processes* 64, 22-30.

Klein, B. and Leffler, K.B. (1981) The role of market forces in assuring contractual performance. *Journal of Political Economy* 89, 615-640.

Kohlberg, L. (1963) The development of children's orientations toward a moral order: I. sequence in the development of moral thought. *Vita Humana* 6, 11-33.

Laibson, D. (1997) Golden eggs and hyperbolic discounting. *Quarterly Journal of Economics*, 62, 443-479.

Licklider, J.C.R. (1959) On psychophysiological models. In W.A. Rosenbluth (Ed.), *Sensory Communication*, M.I.T..

Loewenstein, G. (1996) Out of control: Visceral influences on behavior. *Organizational Behavior and Human Decision Processes* 35, 272-292.

Macaulay, S. (1963) Non-contractual relations in business: A preliminary study. *American Sociological Review* 28, 55-67.

McClennen, E. F. (1990) *Rationality and Dynamic Choice*. Cambridge.

McConkey, K. M. (1984) Clinical hypnosis: Differential impact on volitional and nonvolitional disorders. *Canadian Psychology* 25, 79-83.

Malekzadeh, A., R. and Nahavandi, A. (1987) Merger mania: Who wins? Who loses? *Journal of Business Strategy* 8, 76-79.

Marks, I. & Tobena, A. (1990) Learning and unlearning fear: A clinical and evolutionary perspective. *Neuroscience and Biobehavioral Reviews* 14, 365-384.

Mazur, J.E. (1987) An adjusting procedure for studying delayed reinforcement.. In M.L. Commons, J.E. Mazur, J.A. Nevin, and H. Rachlin, (Eds.), *Quantitative Analyses of Behavior V: TheEffect of Delay and of Intervening Events on Reinforcement Value,* Erlbaum.

Mazur, J. E. (1997) Choice, delay, probability, and conditioned reinforcement. *Animal Learning and Behavior* 25, 131-147.

Melzack, R. and Casey, K.L. (1970) The affective dimension of pain. In M.B. Arnold (Ed.), *Feelings and Emotions*. Academic, pp. 55-68.

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

Millar, A. and Navarick, D.J. (1984) Self-control and choice in humans: effects of video game playing as a positive reinforcer. *Learning and Motivation* 15, 203-218.

Miller, Neal (1969) Learning of visceral and glandular responses. *Science* 163, 434-445.

Mischel, H.N. and Mischel, W. (1983) The development of childrens knowledge of self-control strategies. *Child Development* 54, 603-619.

Myerson, J. and Green, L. (1995) Discounting of delayed rewards: Models of individual choice. *Journal of the Experimental Analysis of Behavior* 64, 263-276.

Navarick, D.J. (1982) Negative reinforcement and choice in humans. *Learning and Motivation* 13, 361-377.

Nemiah, J.. C. (1977) Alexithymia: Theoretical considerations. *Psychotherapy and Psychosomatics*, 28, 199-206.

Nozick, R. (1993) *The Nature of Rationality*. Princeton U.

Parrott, W. G. (1991) Mood induction and instructions to sustain moods: A test of the subject compliance hypothesis of mood congruent memory. *Cognition and Emotion* 3, 41-52.

Parrott, W. G. (1993) Beyond hedonism: Motives for inhibiting good moods and for maintaining bad moods. In D. M. Wegner & F. W. Pennebaker, (Eds.) *Handbook of Mental Control* Prentice Hall.

Piliavin, J. A., Callero, P. L., and Evans, D. E. (1982) Addiction to altruism? Opponent-process theory and habitual blood donation. *Journal of Personality and Social Psychology* 43, 1200-1213.

Polivy, J. (1998) The effects of behavioral inhibition: Integrating internal cues, cognition, behavior, and affect. *Psychological Inquiry* 9, 181-204.

Rachlin, H. (1995) Self-control: Beyond commitment. *Behavioral and Brain Sciences* 18, 109-159.

Rader, N., Bausano, M., and Richards, J. E. (1980) On the nature of the visual-cliff-avoidance response in human infants. *Child Development* 51, 61-68.

Ramsay, R. W. (1997) Behavioural approaches to bereavement. In S. Rachman & H. J. Eysenck, (Eds.) *The Best of Behavior Research and Therapy*. Pergamon.

Rescorla, R. A. (1988) Pavlovian conditioning: It's not what you think it is. *American Psychologist* 43, 151-160.

Rhue, J. W. and Lynn, S. J. (1987) Fantasy proneness: The ability to hallucinate "as real as real." *British Journal of Experimental and Clinical Hypnosis* 4, 173-180.

Ryle, G. (1949/1984) *The Concept of Mind*. U. Chicago.

Sartre, J.-P. (1939/1948) *The Emotions: Sketch of a Theory* (B. Frechtman, Trans.) Philosophical Library.

Schelling, T. C. (1960) *The Strategy of Conflict*. Harvard University.

Scitovsky, T. (1976) *The Joyless Economy: An Inquiry into Human Satisfaction and Consumer Dissatisfaction*. Oxford.

Sen, A. K. (1977) Rational fools: A critique of the behavioral foundations of economic theory *Philosophy and Public Affairs* 6, 317-344.

Shizgal, P. and Conover, K. (1996) On the neural computation of utility *Current Directions in Psychological Science* 5, 37-43.

Siegal, S. (1983) Classical conditioning, drug tolerance, and drug dependence. In R. Smart, F. Glaser, Y. Israel, H.Kalant, R. Popham, and W. Schmidt (Eds.), *Research Advances in Alcohol and Drug Problems*, vol. 1,.Plenum.

Simon, J. L. (1995) Interpersonal allocation continuous with intertemporal allocation: Binding commitments, pledges, and bequests. *Rationality and Society* 7, 367-430.

Smith, A. (1759/1976) *The Theory of Moral Sentiments*. Oxford U.

Solnick, J., Kannenberg, C., Eckerman, D. and Waller, M. (1980) An experimental analysis of impulsivity and impulse control in humans. *Learning and Motivation* 2, 61-77. Review, 217-225.

Strasberg, L. (1988) *A Dream of Passion: The Development of the Method*. Dutton.

Sully, J. (1884) *Outlines of psychology*. Appleton.

Sunstein, C. R. (1995) Problems with rules. *California Law Review* 83, 953-1030.

Tomkins, S. S. (1978) Script theory: Differential magnification of affects. *Nebraska Symposium on Motivation* 26, 201-236.

Wegner, D. M. (1994) Ironic processes of mental control. *Psychological Review* 101, 34-52.

Vuchinich, R. E. and Simpson, C. A. (1998) Hyperbolic temporal discounting in social drinkers and problem drinkers. *Experimental and Clinical Psychopharmacology* 6, 292-305.

Zimmerman, J. and Ferster, C.B. (1964) Some notes on time-out from reinforcement. *Journal of the Experimental Analysis of Behavior* 7, 13-19.

Acknowledgements

Thanks to Lynne Debiak for corrected art work, and John Monterosso, David Spurrett, and Andries Gouws for comments on this précis.

NOTES

[1] Even if these elements are governed by different brain centers, neurophysiologists Shizgal and Conover have pointed out that there has to be an "evaluative circuitry" that reduces them to a common currency: "For orderly choice to be possible, the utility of all competing resources must be represented on a single, common dimension (1996)."

[2] For an analogous problem in social organization, see Sunstein, 1995.

[3] Figure 6B is figure 10B in the target book. In figure 10B the slope of increasing appetite is steeper than it is in figure 10A, whereas to illustrate my point it has to be the same as in figure 10A. Figure 6B has thus been corrected here.