

**A Selectionist Model of the Ego:  
Implications for Self-Control**

George Ainslie

Veterans Affairs Medical Center, Coatesville, PA,USA  
and Temple Medical College

Presented at *Disorders of Volition*,

a conference of

The Max Planck Institute for Psychological Research

Irsee, Germany, December 13, 2003

TO APPEAR IN ITS COPYRIGHTED PROCEEDINGS

## Abstract

The behavioral sciences increasingly regard “rational choice theory” (RCT) as a description of normal choice-making, even as they find more areas in which people violate it. In economics and behavioral psychology normal individuals are explicitly held to discount future outcomes in an exponential curve; in other fields exponential discounting is implied by RCT’s property of consistency, since all shapes other than exponential sometimes predict reversals of preference as a function of time. Given exponential discounting, the role of a self or ego is to obtain the individual’s greatest advantage by seeking as much information and freedom of action as possible. The ego is a hierarch that coordinates obedient subordinate processes; will in the sense of willpower is superfluous, and self-defeating choices must be explained by a separate motivational principle.

However, parametric experiments on discounting prospective events in animal and human subjects have repeatedly found that a hyperbolic shape (inverse proportionality of value to delay) describes spontaneous choice better than an exponential shape. Three implications of hyperbolic discounting—preference reversal toward smaller sooner (SS) rewards as a function of time (*impulsiveness*), early choice of committing devices to forestall impulsiveness, and decreased impulsiveness when choices are made in whole series rather than singly—have also been found experimentally. Such findings suggest an alternative to the hierarchical model of the self: Behavioral tendencies are selected and shaped by reward in a marketplace of all options that are substitutable for one another. Temporary preferences for SS options like substance abuse and other self-defeating behaviors create a state of limited warfare among successive motivational states. Thus a currently dominant option must include means of forestalling any incompatible options that are likely to become dominant in the future. Neither better information nor greater freedom of action necessarily serves the person’s longest range interest, which is the basic test of rationality. Consistency of choice is only partially achievable; and the most effective means of achieving it—perception of prisoners dilemma-like relationships among successive choices of a similar kind—leads to compulsive side-effects. In this view the ego is not a faculty but an emergent property of the internal marketplace, analogous to Adam Smith’s unseen hand; and the will is a bargaining situation analogous to the “will” of nations. This view also provides a rationale for the compulsive clinical disorders.

## **Text**

Behavioral science offers a smorgasbord of principles describing how people make choices (Mellers *et.al*, 1998), but where actual social planning is necessary, as in economics and law, these principles are winnowed down to the refinement of utility theory that was initiated by Samuelson (1937) and has come to be called expected utility theory, or, more generally, rational choice theory (RCT) (Boudon, 1996; Korobkin & Ulen, 2002, Sugden, 1991). In this theory a person with enough information and time to assimilate it will arrive at hierarchies of preference that are internally consistent (transitive, commensurable, etc.), maximize her probability of getting what she prefers, and do not shift as the perspective of time changes. Lawyers and economists are well aware of evidence from all the behavioral sciences of how people violate RCT. Jolls *et.al*(1998) summarized these violations as bounded willpower (a failure to follow your own plans), bounded rationality (failure to correctly interpret environmental contingencies) and bounded self-interest (a tendency to invest altruism where it will not bring returns), but the violations have seemed haphazard (Posner,1998), and RCT offers at least a uniquely coherent system.

However chaotic actual choices seem, RCT is the strange attractor that pulls possibilities back to a single consistent solution. It has “a unique attractiveness... [because] we need ask no more questions about it.” (Coleman, 1986) It is demonstrably the norm for competitions in marketplaces, which themselves become increasingly rationalized and interconnected, so it seems a small jump to say that people normally use it to make their decisions. Its tenets become “assumptions about how people respond to incentives “ (Korobkin & Ulen, 2002, p. 1055). Violations are abnormalities that require explanation. This way of thinking has spread from the policy-making disciplines to individual psychology, where it is the rationality that cognitive therapists teach their clients (Baumeister *et.al.*, 1994; Beck, 1976). A lower principle such as Plato’s passion or Freud’s id has historically been seen as a competing mechanism of choice, but the lower principle is now seen as mere noise that sometimes obscures the clear signal of RCT. I will argue, however, that the observed deviations from RCT are coherent, that they motivate coherent strategies for dealing with them, and that the competition of these strategies with their target deviations generates familiar complexities of choice that RCT does not begin to contemplate.

## **The Problem of Lower Mental Processes**

People have always divided mental life into lower and higher processes. Lower processes appear at an early age, are spontaneous and strongly motivated, tend to seek goals that are obviously useful to organisms in evolution, and are often thought of as the animal part of our nature. Higher processes develop later, often seem arbitrary, are less connected with biological need, and are often thought of as transcending our animal nature. They are not refined lower

processes, but respond to them and often conflict with them in asymmetrical combats, in which the weapon of the lower processes is superior force and the weapon of the higher processes is superior organization and foresight. Ancient thinkers often held that higher processes should simply replace lower ones, as in the Buddhist and stoic ideals of escaping from desire, the Zoroastrian end of light replacing darkness, and the Judeo-Christian practice of mortifying the flesh. However, it became evident that the relationship of these processes is not one of good versus evil. As Freud pointed out, "The substitution of the reality principle for the pleasure principle implies no deposing of the pleasure principle, but only a safeguarding of it." (1911, p. 223). Conversely, psychotherapies often attribute patients' miseries to overgrown higher processes—"cognitive maps (Gestalt)," "conditions of worth (client-centered)," "masturbation (rational-emotive)," and of course the punitive superego (summarized in Corsini, 1984). It is perhaps less clear than it ever was what makes lower processes lower and higher processes higher.

Most theories have had what has been called in this conference a top-down approach to the topic. An autonomous faculty—the Vedic *tapas*, St. Augustine's *temperance*, Plato's *reason*—imposes logical consistency and stability over time on the lower process. In top-down theories this faculty is not governed by the same determinants as the lower process, which is the slave of reward and—if this is something different—of passion. Perhaps attributing the same determinants would make us expect the same results; and in any case any dependence on lawful principles that make "the human person a closed system" is said to reduce people to "powerless victims of mechanism" (Miller, 2003, p. 63). The higher principle is the "you-noun (*ibid*)," the ego, that must be and perhaps should be impenetrable.

It is possible, of course, that our higher processes cannot be explained by mechanisms, that is, by a bottom-up approach. It is also possible that the right mechanisms simply have not been discerned. Certainly many authors have leapt from the discovery of a new atom of learning or motivation to an encompassing theory in which these atoms are merely multiplied or writ large, making the world into a procrustean Skinner box that fails to fit the subtleties of human experience. However, the science of motivation has finally become a cumulative one, in which the current generation stands on the shoulders of previous generations rather than rediscovering the same phenomena in different frames. I will argue that developments during the last four decades in behavioral research, bargaining theory, and even the philosophy of mind permit a model that comes significantly closer than previous models to fitting the subtleties of human character. In particular, I will show how it improves on a currently dominant atom-writ-large, RCT.

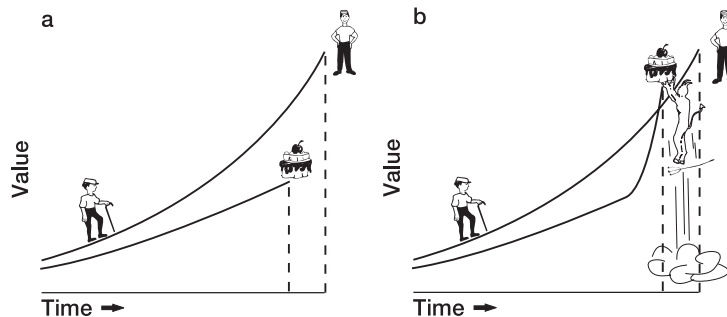
Of the three kinds of deviation from RCT catalogued by Jolls *et.al*, the most attention has been paid to bounded rationality and bounded self-interest. I will not discuss them here.<sup>1</sup> A far more serious problem is bounded willpower--the widespread violation of temporal consistency. People regularly express a preference for one course of action and then take the opposite course when they actually choose. This is sometimes a minor foible, mere fickleness, but often

immerses the person in substance abuse, pathological gambling, destructive rage-- indeed a large part of the psychiatric diagnostic manual (American Psychiatric Association, 1994). An even larger number of “bad habits” never reach the level of diagnosis: smoking, overeating, credit card abuse, rash attachments, impatience for pleasant things and procrastination of unpleasant ones—all the activities that you plan to avoid when you’re at a distance from them, and regret after you’ve done them.

RCT holds, against all intuition, that simple insight should prevent these lapses. Consistent choice implies an exponential discount curve of the value of delayed goals, such that they lose a constant proportion of their remaining value for every additional unit of delay. Financial transactions are universally conducted on the basis of the exponential discount curve, for any curve more bowed than this would lead a good to change its value relative to alternatives simply as it drew closer, an irrational instability. People regularly make their investment choices on the basis of exponential curves, so it makes sense to think that these curves are part of attainable insight. According to RCT, the choice between dessert now and fitness down the road should be reducible to a graph like figure 1a (given that an extended reward like that from fitness can be represented as a single event—Mazur, 1986; Ainslie, 1992, pp. 147-152, 375-385).

Confronted with the prevalence of temporary preferences, utility theorists have borrowed a mechanism from popular culture, a surge of preference for the less valued alternative when it looms close. Spirit possession was popular in more superstitious times, and you can still hear, “the Devil made me do it.” However, since the surge often follows a cue that signals the imminent possibility of the bad option, psychology has attributed it to classical conditioning: Appetite is assumed to be an unmotivated response transferred from a hardwired stimulus, and its sudden appearance makes the prospective reward<sup>ii</sup> from the bad option jump above that of the good option; hence the effect seen in figure 1b.

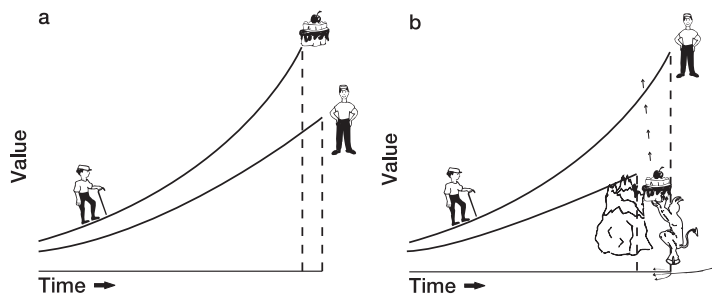
Figure 1



*In RCT, the values of LL fitness and SS dessert keep the same proportion at all times (a), unless a special mechanism (here, a demon) creates a surge in the dessert’s value (b).*

The problem with this model, aside from serious questions about whether classical conditioning represents a selective principle separate from reward (Ainslie, 1992, pp. 39-48, 2001, pp. 19-22), is that most if not all rewards are preceded by predictive cues. Almost all rewards must be “conditioned,” even the rewards that seem to be discounted rationally. Cues merely tell us the likelihood of occurrence and delay of the rewards they predict. A cue that regularly precedes a reward should become predictable in turn, and if it makes the bad reward more valuable it should soon raise the height of the discount curve all the way back, causing it to be revalued as a straightforwardly better reward (figure 2a). Thus the conditioning theory of impulses has to assume that you can’t learn the connection between cues and the kind of rewards that get temporarily preferred, or at least that you can’t learn the hedonic implications of this cue/reward pair. The “visceral rewards” that are frequent offenders in impulsive choice (Loewenstein, 1996) must thus stay surprising, and jump out at an unwary person however often she has previously lapsed and chosen the bad reward in the same circumstances (figure 2b).

Figure 2



*After several surges like that in 1a, the person should come to anticipate them and simply re-value dessert (a), unless the surge cannot be anticipated and thus stays surprising (b).*

This would be a somewhat anomalous occurrence, given that animals evaluate the prospect of the same visceral rewards with great accuracy (Herrnstein, 1969), and human addicts often anticipate lapses enough to take precautions against them. The experience of suddenly becoming conscious of an overwhelming appetite is common, and needs explanation in its own right; but it is not an adequate mechanism for temporary changes of preference in general.

## Theoretical Models of the Will

In RCT the person continually maximizes her future prospective reward. Higher processes involve only estimating what course will do this (Becker & Murphy, 1988). If we graft unpredictable conditioned appetites onto this model, we add the task of forestalling these preferences. Most people would say that the tool they use for this task is willpower or some synonym— resolve, intentionality, etc. However, this has not been a robust concept, rather a will-o'-the-wisp, which has eluded definition and study to the point where some authors deny its existence. Part of the problem has been that the term refers to at least three distinct processes—not only the maintenance of long range plans but also the simple initiation of any behavior—the sense in which Ryle found the concept unnecessary (1949/1984)—and the integration of specific plans with the whole self, the “ownership” process whose familiar lacunes seem to be what leads Wegner to call the will illusory (2002; see Ainslie, 2004). It is only in the first sense of maintaining long range plans that the concept of willpower is relevant; and there is no generally accepted mechanism for how this happens. Perhaps that is why the will is an exemplar of what are held to be impenetrable higher processes.

Although a mechanism has been lacking, there has been agreement about several properties of willpower. First of all, gimmicks are excluded. Seeking external means of control, like taking appetite-spoiling drugs, committing your funds to money managers, or joining social groups that will exert pressure, would not be called will. Positive properties were well defined by Victorian psychologists. Willpower was said to:

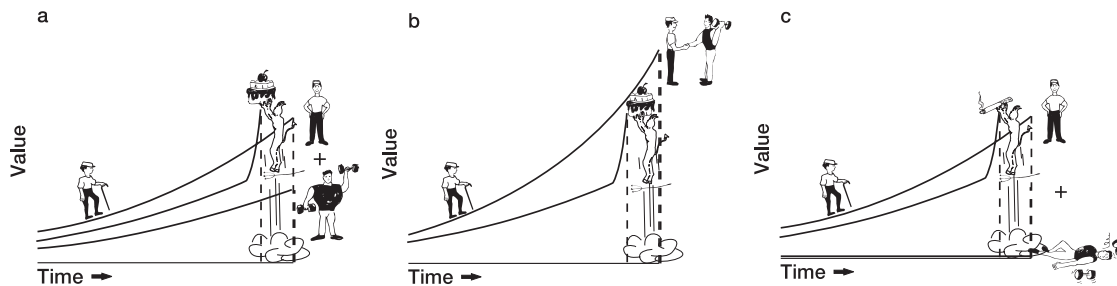
- come into play as "a new force distinct from the impulses primarily engaged (Sully, 1884, p. 669);"
- "throw in its strength on the weaker side... to neutralize the preponderance of certain agreeable sensations (*ibid*);"
- "unite... particular actions... under a common rule," so that "they are viewed as members of a class of actions subserving one comprehensive end (*ibid* p. 631);"
- be strengthened by repetition (*ibid* p. 633);
- be exquisitely vulnerable to nonrepetition, so that "every gain on the wrong side undoes the effect of many conquests on the right (Bain, 1886, p. 440);" and
- involve no repression or diversion of attention, so that "both alternatives are steadily held in view, and in the very act of murdering the vanquished possibility the chooser realizes how much in that instant he is making himself lose (James, 1890, vol. 2, p. 534)."

Three internal mechanisms have been proposed that are at least roughly compatible with these properties: building “strength,” making “resolute choices,” and deciding according to principle. However, we need to ask each of these hypotheses both whether its mechanism is complete or requires another will-like faculty to guide it, and whether it recruits adequate motivation to govern the

decision. Given a motivational structure made up of exponential (consistent) discount curves and conditioned cravings, these models all have problems.

*Strength.* Baumeister and others have proposed an organ of self-control, the main property of which is that, like a muscle, it gets stronger with use in the long run but can be exhausted in the short run (1994, pp. 17-20; 1996). Presumably it adds motivation to what is otherwise the weaker side (figure 3a), pushing it above the temporary surge of motivation (figure 3b). 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, overeating, when it has fallen flat in another such as smoking (figure 3c)? The strength concept merely elevates one of the familiar properties of will into a mechanism in its own right, without grounding it on any robust source of motivation.

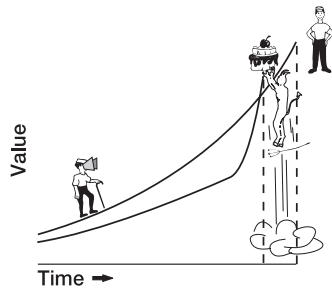
Figure 3



*In the strength model of will, there is an additional faculty that can add its own value to that of fitness (a), leading to a combined value that overcomes the attraction of dessert (b). It is not clear why this strength can be absent when there is a different kind of temptation (here, smoking-- c).*

*Resolute Choice.* Philosophers of mind favor the idea of “resolute choice” (e.g. McClennen, 1990; Bratman, 1999). When they venture to specify a mechanism it mostly involves not re-examining choices, at least while the person expects the bad choice to be dominant. There have been a number of experiments suggesting how children learn to do this: Mischel and his collaborators sometimes refer to a combination of controlling attention and avoiding emotionally “hot” thoughts as willpower (e.g. Metcalf & Mischel, 1999), essentially a use of mental blinders (figure 4).

Figure 4

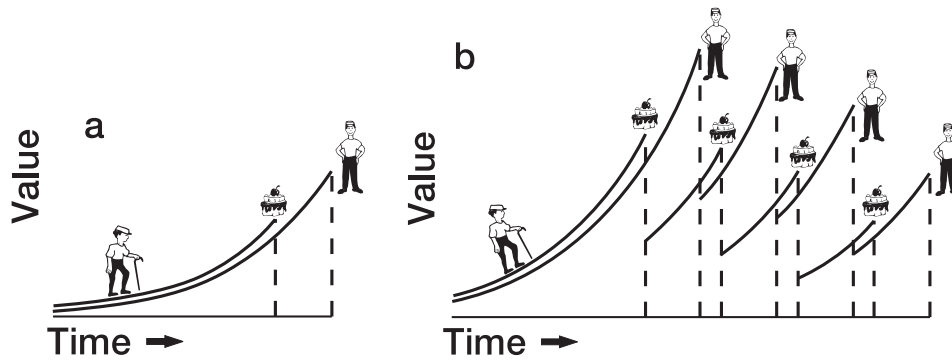


*In resolute choice, the person may avoid re-evaluating the options (blinders)—or there may be more to it.*

However, I have argued that diverting attention and nipping emotion in the bud are distinct and less powerful mechanisms of committing your behavior in advance (Ainslie, 1992, pp.133-142). The ability to control yourself in such a way that “both alternatives are steadily held in view” requires something more. Metcalfe and Mischel describe a growing interconnectedness of a child’s “cool” processes, which implies something more than just diversion of attention. Mere diversion after all is an act of holding your breath, useable, as hypnosis has demonstrated, against very short range urges like panic and the affective component of pain, but not against addictions, the urge for which forces a re-evaluation over the hours or days that the diversion must be maintained (McConkey, 1984). The philosophers, too, sense the need for a more complex mechanism: McClennen refers to “a sense of commitment” to previously made plans (1990, pp. 157-161), which sounds like more than diversion of attention, 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. This raises the issue of deciding according to principle, which we will now examine.

*Principle.* Since ancient times looking away from tempting options has been the main folk ingredient of self-control, but a subtler technique is just as venerable: deciding according to principle. Referring to dispositions to choose as “opinions” Aristotle said, “We may also look to the cause of incontinence [akrasia] scientifically in this way: One opinion is universal, the other concerns particulars...” (*Nichomachean Ethics* 1147a24-28). Deciding according to universals made you more continent. Many authors have repeated this advice (some listed in Ainslie, 2001, pp 79-81), but mostly without speculating as to how people can maintain their motivation to narrow their range of choice in this way. Simply summing series of exponentially discounted rewards together does nothing *per se* to change their relative values (figure 5).

Figure 5



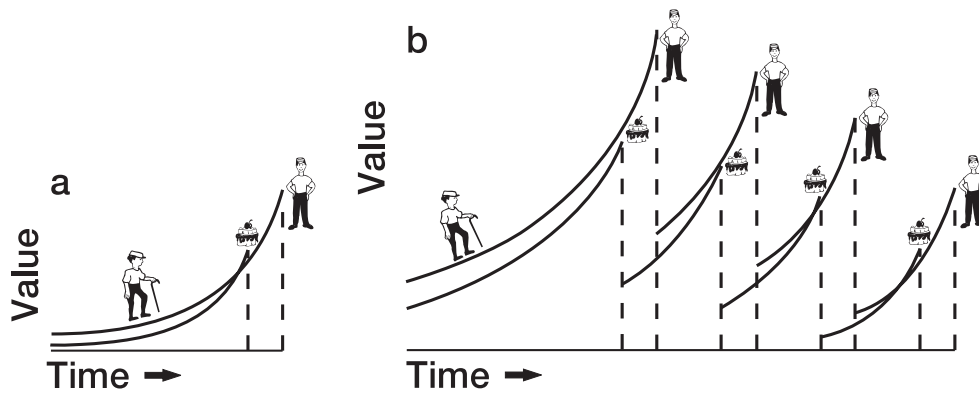
*Using principled choice, membership in a larger category of LL rewards must increase the relative value of individual LL rewards. This does not happen with exponential curves; a prospective series of reward (b) keep the same relative values as a single choice (a).*

However, Howard Rachlin has given considerable thought as to how people come to choose in “molar,” overall patterns rather than making “molecular” decisions, by which he means going case by case (2000). He believes that there comes to be an aesthetic factor in molar choice itself, just as, with learning, a whole symphony comes to be more rewarding than the sum of its parts. Thus a recovered addict might avoid lapses because of the aversiveness of spoiling her pattern of sobriety. In this model the strength or resolve that feels like the active ingredient in willpower is hypothesized to come from a specific mechanism, molar appreciation of an overall pattern, leading to distaste for options that break the pattern. This model does have the advantage of specifying the extra motivation to overcome temptations that choosing in categories seems to supply.

This aesthetic factor does not seem robust enough; distaste is not how most people would describe even the temptations that they manage to avoid. However, without it, there seems to be no way that bundling exponentially discounted options together could be expected to shift the direction of choice.

I have argued that no satisfactory theory of impulsiveness or impulse control can be based on exponential discount curves—that *a priori*, without data about the actual shape of the curves, there is a need to postulate curves more bowed than exponential ones (Ainslie, 1992, 2001, pp. 117-140). Highly bowed curves can account for both temporary preferences and the motivation to forestall them, as figure 6 demonstrates.

Figure 6



Principled choice *boosts LL reward values only when discount curves are hyperbolic or otherwise deeply bowed. Such curves from a series of paired SS and LL rewards may come never to cross (b), with the same amounts that cause curves from a single pair to cross (a, = last pair in b).*

A hyperbolic discounter who faces a choice between smaller-sooner (SS) and larger-later (LL) rewards will evaluate them roughly in proportion to their objective size—their values at zero delay—when both are distant, but value the SS reward disproportionately when it is close (figure 6a). Thus she will have an innate tendency to form temporary preferences for SS rewards, purely as a function of elapsing time. Furthermore, if she makes a whole series of choices at once—for instance a class of choices united by a principle—the curve describing her valuation of the LL rewards will be much higher (figure 6b).

Hyperbolic discount curves are a radical theoretical departure and lead to converse problems with how choice becomes stable, but they are not an outrageous leap. The degree of most psychophysical changes—from one intensity of warmth or brightness or heaviness to another—is experienced proportionately to the original intensity, a relationship expressed by a hyperbolic rather than an exponential curve (Gibbon, 1977). It does not strain our beliefs about nature that amounts of reward might be experienced proportionally to their immediacies.

### **Empirical Evidence about Temptation and Will**

Fortunately, the shape of the discount curve can be studied by controlled experiment, with at least four different methods and in both people and nonhuman animals. A large body of such research has occurred in the thirty years since I first proposed the hyperbolic shape (Ainslie, 1974, 1975); this research has found a robust and apparently universal tendency to discount delayed events in a curve more bowed than an exponential curve. Where the method has permitted

estimation of the exact shape, the shape that has best fit the data produced by that method has been a hyperbola. I will summarize the findings briefly:

1. Given choices between rewards of varying sizes at varying delays, both human and nonhuman subjects express preferences that by least squares tests fit curves of the form,

$$V = A / (1 + kD)$$

a hyperbola, better than the form,

$$V = A e^{kD}$$

an exponential curve (where  $V$  is motivational value,  $A$  is amount of reward,  $D$  is delay of reward from the moment of choice, and  $k$  is a constant expressing impatience; Grace, 1996; Green, Fry & Myerson, 1994; Kirby, 1997; Mazur 2001). It has also been observed that the incentive value of small series of rewards is the sum of hyperbolic discount curves from those rewards (Brunner & Gibbon, 1995; Mazur, 1986; Mitchell, 2003).

2. Given choices between SS rewards and LL ones available at a constant lag after the SS ones, subjects prefer the LL reward when the delay before both rewards is long, but switch to the SS reward as it becomes imminent, a pattern that would not be seen if the discount curves were exponential (Ainslie & Herrnstein, 1981; Ainslie & Haendel, 1983; Green *et.al*, 1981; Kirby & Herrnstein, 1995). Where anticipatory dread is not a factor (with nonhumans or with minor pains in humans), subjects switch from choosing SS aversive stimuli to LL ones as the SS ones draw near (Dinsmoor, 1998; Novarick, 1982; Solnick, 1980).

3. Given choices between SS rewards and LL 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 (Deluty *et.al*, 1983). This design has not been run with human subjects, but it has been argued that illiquid savings plans and other choice-reducing devices serve this purpose (Laibson, 1997). Such a pattern is predicted by hyperbolic discount curves, while conventional utility theory holds that a subject has no incentive to reduce her future range of choices (Becker & Murphy, 1988).

4. When a whole series of LL rewards and SS alternatives must be chosen all at once, both human and nonhuman subjects choose the LL rewards more than when each SS vs. LL choice can be made individually. 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 (2002). They reported an even greater effect for different amounts of pizza. Ainslie and Monterosso reported that rats made more LL choices when they chose for three trials all at once than they chose between the same contingencies separately on each trial (2003). The effect of such *bundling* of choices is predicted by hyperbolic but not exponential curves: As I described above, exponentially discounted prospects do not change their relative values

however many are summed together (figure 5); hyperbolically discounted SS rewards, although disproportionately valued as they draw near, lose much of this differential value when choices are bundled into series (figure 6).

Thus hyperbolic discounting seems to be an elementary property of the reward process. The resulting notion that our choices are intrinsically unstable is obviously disturbing, and requires a fair amount of theoretical re-tooling. Several counter-proposals have attempted to account for temporary preference phenomena as variants of exponential discounting. The simplest possibility is that different kinds of reward are discounted at different rates, so that the prospect of sobriety, say, might be discounted more slowly than that for intoxication. Such an explanation could account for temporary preferences, precommitment, and the effect of summing series of choices, as long as the SS rewards were of a different modality than the LL rewards. However, in all of the above experiments the SS rewards were of the same kind as the LL.

Other proposals have included:

- Noise in the valuation process, such that discount curves wobble randomly across one another (Strotz, 1956, Skog, 1999). However, since exponential curves draw further apart as delay decreases, this wobble should create fewer changes of preference, or at least no more, when the SS is near than when it is distant. The opposite is regularly observed.
- A step function in which immediate events are valued exceptionally and events at all delays are discounted exponentially (Simon, 1995); the most prominent example is Laibson's hyperboloid discount function (1997). This grossly accounts for the incentive for precommitment; but this function, not seen elsewhere in nature, is contradicted by the monotonic form of the available data.
- An exponential discount rate whose exponent itself varies as a function of amount (Green & Myerson, 1993). However, to explain changes of preference as a function of delay, the exponent would have to be determined only by the value at delay zero, so that a fifty dollar prize would be discounted more rapidly than a hundred dollar prize *even after the discounted value of the hundred dollar prize had fallen to fifty dollars*.
- The summation of separate exponential discount rates for association and valuation (Case, 1997). However, the association component that gives the necessary bowing to the overall curve should affect only new learning, not choice between the familiar alternatives that confronted subjects in most of the above research.

None of these proposals contradict hyperbolic discounting except in the precise fitting of the curve itself, and in this respect, the data for best least squares fit overwhelmingly support the hyperbola.

The finding of evidence for hyperbolic discounting in nonhumans as well as humans is crucial, because social psychology experiments are notoriously vulnerable to unprogrammed incentives, not the least of which is compliance with

perceived experimenter demand (Orne, 1973). Phrasing a choice one way or another can reverse the direction of the findings (Tversky & Kahnemann, 1981), and subjects are apt to express what they believe to be rational rather than what their spontaneous preference is; thus six-to-ten-year-olds are actually poorer at some kinds of reward-getting tasks than four-year-olds, because they rigidly hold to what they expect is the right strategy (Sonuga-Barke *et.al*, 1989). Furthermore, human subjects learn to compensate for their tendencies to form temporary preferences, and express valuations that have this compensation already factored in; I am still surprised that people reveal hyperbolic preferences for future money to the extent that they do, given its demonstrable irrationality. Of course nonhuman animals have their own behavioral foibles (Breland & Breland, 1961), but we can be sure that these do not include social demand or preconceived plans.

Hyperbolic curves suggest rationales for many phenomena that RCT fails to predict, even with the help of its designated villain, conditioned craving. This shape can obviously account for reversals of preference as SS rewards become imminently available. At first glance, it does not explain the stimulus-driven quality often reported for these reversals: A switch in preference is often experienced as happening not simply when a reward can be had soon, but when a stimulus induces a “conditioned” surge of appetite for it. However, I will argue presently that hyperbolic curves have a role in this surge by the same kind of mechanism that leads to the willpower phenomenon. These curves also repair many other defects of RCT—its inability to account for anomalies of investment (Thaler, 1991); its silence on the value of emotion; its confusion about the most important occasion for emotion, the vicarious experience of other people; and its total failure to notice the discomfort, even embarrassment, that many people feel about analyzing these questions directly. I will discuss willpower and stimulus induction here, and refer the reader to a longer work for the other topics (Ainslie, 2001, pp. 161-197).

### **Will as Intertemporal Bargaining**

The most basic consequence of hyperbolic discounting is that we are strangers to ourselves, at least more so than is commonly assumed. Neither cognitive theory nor popular imagination has revised the renaissance image of the person as an internal hierarchy, with an ego as king over obedient agents (muscles) and passive support organs (viscera; Tillyard, 1959). At best this image has been modernized to a corporation controlled by a CEO, or an army controlled by a general. By contrast, if our preferences tend to change as one reward and then another get close, we are more like a marketplace in which any plan we make at one moment must be sold to ourselves at future moments if it is to have any chance of succeeding. This, indeed, is even what corporations and armies look like when the motives of the individuals who “serve” in them are examined closely (Brunsson, 1985, chapters 1 and 2; Brennan & Tullock, 1982, p. 226). Memos and orders have to be supported by a great deal of tacit bargaining.

What bargaining within individuals can make a future self obey the plan of the present self? Of course there is sometimes external or physiological commitment, as when the present self takes an appetite-altering medication, makes a promise to a friend, limits the information that will come to future selves, or just starts a behavior that will affect motivation in the immediate future (Ainslie, 2001, pp.73-78). However, these methods are often unavailable, or too costly or restricting. A more adaptable method is suggested by hyperbolic curves' property of increasingly favoring LL rewards when they are drawn from whole series of rewards, as demonstrated in the fourth kind of experiment, above. This property may be the basis for what authors from Aristotle to Rachlin have suggested: that self-control increases when you decide according to principle—that is, when you choose whole series of similar options instead of just “particular,” “molecular” cases. However, long range advantage is not an adequate explanation for why people stick to a principle in the face of individual short range temptations. For this we need to invoke a process that would make no sense to the continual reward maximizers envisioned by RCT, *intertemporal bargaining*.

Future selves partially share the goals of the present self—the LL rewards that it values at a discount—and partially have different goals—the SS rewards that only the present self values highly. This defines a relationship of limited warfare, the incentives for which, in interpersonal bargaining, form repeated prisoners' dilemmas (RPDs). Such conflicts among individuals can be solved by finding clear, albeit often tacit, criteria for what constitutes cooperation or defection, as long as mutual cooperation will benefit each player more than mutual defection will. Classical RPDs cannot occur among successive selves within an individual because a later self can never literally retaliate against an earlier one. However, I have argued that the dependence of your expectation of a whole series of LL rewards on seeing yourself pick LL rewards in current choices effectively creates the outcome matrix of an RPD (Ainslie, 2001, pp. 90-104). If you see yourself violate your diet today you reduce your expectation that your diet will succeed; tomorrow's self will have that much less at stake in its effort; and tomorrow's self, by violating your diet in turn and reducing your expectation still further, will have in effect retaliated against today's defector.

The incentive structure of intertemporal bargaining can replace not only Rachlin's supplementary reward from love of principle but also faculties like a transcendent self or overriding ego that have long been assumed to be inborn. With interpersonal bargaining, small, stable markets come to regulate themselves by “self-enforcing contracts” (Klein & Leffler, 1981)—self-enforcing in that the incentive for cheating in a given transaction is continuously less than the expected gain from continuing mutual trust. By the same logic, an individual has incentives to develop self-enforcing cooperative arrangements with her future selves. Higher mental functions can develop by trial and error on the basis of the relatively small but stable rewards that attend foresight. A person's cognitive machinery need not be run by an autonomous part of the person herself, an ego that stands apart from its gears and power trains; the internal factory itself is

autonomous, the ultimate bottom-up mechanism that Dennett envisions (this volume).

The contingencies of the intertemporal RPD were illustrated by a demonstration at this conference: I asked the audience to imagine that I was running a game show. I announced that I would go along every row, starting at the front, and give each member a chance to say "cooperate" or "defect." Each time someone said "defect" I would award a euro only to her. Each time someone said "cooperate" I would award ten cents to her and to everyone else in the audience. And I asked that they play this game solely to maximize their individual total score, without worrying about friendship, politeness, the common good, etc. I said that I would stop at an unpredictable point after at least twenty players had played. Like successive motivational states within a person, each successive player had a direct interest in the behavior of each subsequent player; and had to guess her future choices somewhat by noticing the choices already made. If she realized that her move would be the most salient of these choices right after she made it, she had an incentive to forego a sure euro, but only if she thought that this choice would be both necessary and sufficient to make later players do likewise.

In this kind of game, knowing the other players' thoughts and characters--whether they are greedy, or devious, for instance—will not help you choose, as long as you believe them to be playing to maximize their monetary gains. This is so because the main determinant of their choices will be the pattern of previous members' play at the moment of these choices. Retaliation for a defection will not occur punitively-- a current player has no reason to reward or punish a player who will not play again<sup>iii</sup>-- but what amounts to retaliation will happen through the effect of this defection on subsequent players' estimations of their prospects and their consequent choices. These would seem to be the same considerations that bear on successive motivational states within a person, except that in this interpersonal game the reward for future cooperations is flat (ten cents per cooperation, discounted negligibly), rather than discounted in a hyperbolic curve depending on each reward's delay.

Perceiving each choice as a test case for the climate of cooperation turns the activity into a positive feedback system—cooperations make further cooperations more likely, and defections make defections more likely. The continuous curve of motivation is broken into dichotomies, resolutions that either succeed or fail. Proximity to temptation still influences the outcome of choices, but much less so than when they did not serve as test cases with whole series of expectations riding on them. The interpretation of cases as tests or not, that is, as members or not of this particular RPD, becomes more important in determining whether a temptation is worth resisting. If you violate your diet on a special day like Thanksgiving, or if a single conspicuous outsider like a child in the game show audience defects, the next choice-makers will be much less likely to see it as a precedent. The importance of interpretation creates incentive for what Freudians call rationalization, or Sayette calls motivated reasoning (this volume). Making resolutions more explicit forestalls impulsively motivated reasoning and

increases their chances of being carried out (Gollwitzer, this volume), but at the risk of compulsive side effects, as we shall see.

The similar incentive structures of interpersonal and intertemporal bargaining might make it seem like a good idea to use the former to study the properties of the latter. In full blown form, however, this turns out to be a ponderous undertaking. John Monterosso, Pamela Toppi Mullen and I have tried out the game show experiment with repeated trials for real money in a roomful of recovering addicts, but it was evident that social pressure was more of a factor than the announced rewards (unpublished data). Practical use of this method would require subjects sitting at thirty or forty separate terminals, enough trials to make them familiar with the logic of choice, and enough payoff to make it worth their time—obvious material for a well-funded internet study. Meanwhile it has been possible to model some of the logic of intertemporal cooperation in a two person RPD: Subjects at computer terminals given false feedback about their partners' responses have shown that damage done by defections is greater and more long lasting than damage repair following cooperations (Monterosso *et.al*, 2002)—the same asymmetry described for lapses of will (Bain, 1886, p. 440).

Experimental analogs are a noisy way to study intertemporal bargaining, but direct experimentation on this recursive, internal process is even less practical. There are suggestive data. 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 the separate-choice subjects that their current choice might be an indicator of what they would choose on subsequent occasions (2002). However, nothing short of imaging techniques would allow direct observation of the separate steps of recursive choices within individuals, and these techniques are in their infancy. Meanwhile, the most convincing evidence for the dependence of will upon self-observation comes from thought experiments of the kind that have been finely honed by the philosophy of mind (Kavka, 1983; Sorensen, 1992). An example tailored to self-control:

Consider a smoker who is trying to quit, but who craves a cigarette. Suppose that an angel whispers to her that, regardless of whether or not she smokes the desired cigarette, she is destined to smoke a pack a day from tomorrow on. Given this certainty, she would have no incentive to turn down the cigarette—the effort would seem pointless. What if the angel whispers instead that she is destined never to smoke again after today, regardless of her current choice? Here, too, there seems to be little incentive to turn down the cigarette—it would be harmless. Fixing future smoking choices in either direction (or anywhere in between) evidently makes smoking the dominant current choice. Only if future smoking is in doubt does a current abstention seem worth the effort. But the importance of her current choice cannot come from any physical consequences for future choices; hence the conclusion that it matters as a precedent. (Monterosso & Ainslie, 1999)

## **Recursive Self-Prediction in Will and “Conditioned Craving”**

Sometimes resolutions are deliberate, and people monitor cooperation systematically. However, less deliberate resolutions that still depend on recursive self-observations are apt to be more widespread. We intend to donate blood or dive into a cold lake and do not take formal notice of whether we do or not; but if we do not, it will be harder to intend similar acts the next time. Resolutions and intentions shade into the kind of self-predictions that merely forecast the immediate future, are made according to no principle, and may well occur in nonhuman animals. On one end of the scale, Russell's example of fending off seasickness involves effort:

I suspect that I may be getting seasick so I follow someone's advice to "keep your eyes on the horizon..." The effort to look at the horizon will fail if it amounts to a token made in a spirit of desperation... I must look at it in the way one would for reasons other than those of getting over nausea... not with the despair of "I must look at the horizon or else I shall be sick!" To become well I must pretend I am well (1978, pp. 27- 28).

But this example is continuous with the effortless James-Lange phenomena on the other end that were described in the nineteenth century, actually first by Darwin:

The free expression by outward signs of an emotion intensifies it. On the other hand, the repression, as far as this is possible, of all outward signs softens our emotions. He who gives way to violent gestures will increase his rage; he who does not control the signs of fear will experience fear in greater degree (1872/1979, p. 366).

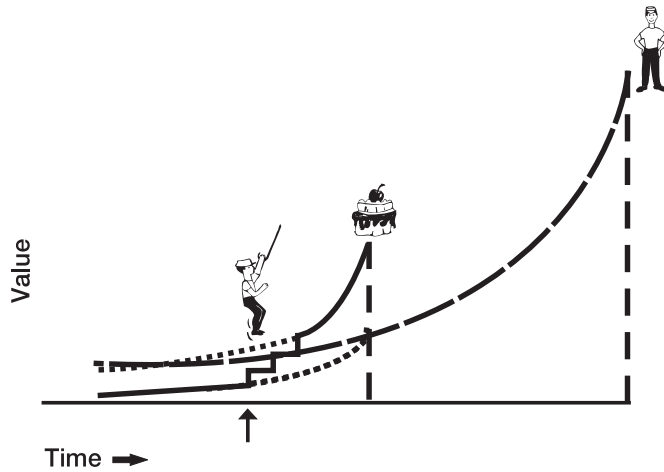
Anxiously hovering over your own performance is common in behaviors that you recognize to be only marginally under voluntary control: summoning the courage to perform in public (versus what comedians call "flopsweat") or face the enemy in battle, recall an elusive memory, sustain a penile erection, or, for men with enlarged prostates, void their bladders. To seem to be succeeding increases the actual likelihood of success. I suspect that it was not just to account for fate, but to describe the tenuous process of succeeding in just such behaviors, that polytheists discerned interactions with such gods as Mars, Venus, and Aesculapius. Will in the sense of willpower is a refinement of this recursive self-prediction to govern behaviors that are more reliable than the above examples in the short run, but become tenuous when they must be sustained over long periods. People pray to gods for success against temptations, too.

*Sudden craving* Recursive self-prediction is a likely mechanism for the apparent suddenness of "conditioned" craving—that part of the temptation experience which is not described simply by hyperbolic discount curves. In a reward-based view, craving is an example of appetite, a preparatory behavior that increases the effect of relevant rewards, given adequate biological need (or deprivation, or "drive"). Like many processes including the direction of thought itself, appetite occurs too rapidly to be inhibited by will, but it can be cultivated by will, as in daydreams. It occurs spontaneously only when there is sufficient

chance that it will be rewarded. It is a behavior that is inexpensive of resources and thus worth emitting even when the chance of reward is small, as a pet will beg even under circumstances when it is not usually rewarded. But consistent nonreward will cause appetite to extinguish, most rapidly in cases where appetite that is not followed by the relevant event turns aversive—For instance, if you prepare yourself for dinner and then do not get it, you bring on pangs of hunger; thus starving people often cease to get hungry (Carlson, 1916).

In the absence of certainty that appetite will *not* be rewarded, we try it out readily, in part to discover whether drive exists—we do not seem to get direct readouts of our drive states (Cameron, 2002). In situations where we can choose the relevant event or not, trying out appetite not only tells us whether it would be rewarding, but makes its prospect more rewarding by the second as we entertain the appetite. In these situations appetite does not just make the event more rewarding; the increase in rewardingness also makes the event more likely. This looks like a Darwin-James-Lange positive feedback cycle. If we never consume the reward in a particular circumstance we do not generate appetite there, just as orthodox Jews are said not to crave cigarettes on the Sabbath (Schachter *et.al*, 1977). At the opposite pole, if we accept that we usually consume the reward in this circumstance we will develop appetite in a monotonic, rising curve as the rewarding event gets closer, and the impact of the prospective reward will follow the same hyperbolic curve as it does in a nonhuman animal. But between these extremes, if we intend, without certainty, *not* to consume the reward, we will be prone to sudden increases in appetite that may or may not change the preference that was based on our previous anticipation (figure 7). The notorious dessert cart phenomenon occurs only in people who intend weakly not to have dessert. And if, starting in this middle, we add to our resolve and stop ever consuming the reward in this circumstance, it will still take many, many trials for the appetite to extinguish here.

Figure 7



*With hyperbolic curves, sudden craving may occur not only from proximity but also in recursive, James-Lange-Darwin fashion (steps), when appetite and the person's prediction of taking a nearby SS reward feed back positively to each other. Unless the person is sure of not indulging, a suggestive cue (at arrow) makes incentive move from the lower curve (value without appetite) to the upper curve (value with appetite).*

The question naturally arises whether this model of appetites as behaviors will work in the converse situation where there seem to be negative appetites. That is, there is a readiness to have anger, fear, and grief as well as the experiences that we seek to have; it is hard to imagine that the experiences that seemingly have to be imposed by conditioning are actually chosen for their rewardingness. Nevertheless, a consequence of hyperbolic discounting that I have described elsewhere is that reward can account for the selection of all kinds of behaviors, even those that have aversive but vivid consequences (Ainslie, 2001, pp. 48-70). Briefly, aversions may be rapid cycles of short, intense reward and relatively longer suppression of reward—the same pattern as itch and loss of concentration or, even slower, as binge and hangover, but condensed into so short a period that the rewarding and unrewarding components fuse in perception. This model, or any other that acknowledges the ability of aversive events to attract attention in a competitive internal marketplace, makes it possible to see unconditioned stimuli as selecting for the behaviors they follow in exactly the same way as acknowledged rewards do. A separate conditioning principle is no longer necessary to account for the apparent imposition of aversive or otherwise undesirable processes on unwilling subjects. Thus even when craving is unwelcome, it can be seen as arising only insofar as it is rewarded in the very short run.

What the theories of choice that have converged on RCT describe, then, is only a specialized part of choice, perhaps shaped by the conditions of competitive interpersonal markets. RCT represents a set of bargains that a person might make

with herself, some of her *personal rules*. Because of the inescapable ambiguities in these bargains she is fated to achieve “rationality” only imperfectly. Furthermore, insofar as personal rules are all that protect her from her own nature, red in tooth and claw, she is their prisoner. I have argued elsewhere that extensive or unskilled reliance on the perception of RPDs for self-control will motivate the development of four side effects (Ainslie, 2001, pp. 143-160):

- When an option is worth more as a test case than as an event in its own right you are less able to experience it in the here-and-now and your choice-making becomes rigid;
- a lapse that you see as a precedent reduces your hope for self-control in similar situations in the future, a reduction that recursively reduces your power of self-control;
- the incentive not to recognize a lapse may lead to gaps in your awareness of your own behavior;
- explicit criteria for defining lapses will tend to replace subtle ones, making your choice-making overly concrete.

Clinically, these side effects manifest themselves as compulsive symptoms, in the extreme as obsessive-compulsive personality disorder (Pfohl & Blum, 1991).<sup>iv</sup> When the test cases are focused on specific topics they may produce the picture of modality-specific syndromes like anorexia nervosa (Gillberg & Rastam, 1992), or narrow character traits like miserliness. Like the internal marketplace itself, these compulsive side effects have social analogs where society uses laws to control interpersonal bargaining (Sunstein, 1995).

Thus if rationality is maximizing experienced reward over time, solving intertemporal RPDs with rules for cooperation is not necessarily rational. On the contrary, people may rationally follow an incentive to seek other means of avoiding temporary preferences, like physically limiting their future options or information about their future options, tactics that make no sense in RCT. We seek the influence of other people, about which RCT is silent but which Kohlberg classed as a primitive basis for self-control (1963). To the consternation of RCT we gamble prodigiously, literally and figuratively, because we cannot otherwise repair the premature satiation of our emotional appetites that is driven by our urges for SS satisfactions (Ainslie, 2001, pp. 161-189, and 2003). Finally, we conceal the rationale for these activities from ourselves by setting up bogus goals and unnecessary detours, just because our “rational” rules for self-control might otherwise stamp them out (Ainslie, 2001, pp. 189-196).

## Conclusions

I have described a pattern of recursive self-prediction that extends an organism’s basic ability to use the stimuli from its own current behaviors as cues. This extended ability would not be important if people evaluated choices with the exponential discount curves that are intrinsic to RCT; it becomes crucial in the

limited warfare engendered by hyperbolic discount curves. Recursive self-prediction can account for both the recruitment of willpower when you see current choices as test cases and the sudden evaporation of willpower when you toy with generating appetite.

This approach provides a bottom-up rationale for the growth and selection of higher functions; higher literally means more farsighted, for they will be selected according to how well they can anticipate and influence future urges. They do not depend upon an independent organ of reason. Rather they are selected by long range reward itself, an invisible hand like that of Adam Smith's marketplace. However, higher does not necessarily mean wiser, since they are prone, like agents in interpersonal marketplaces, to fall into perverse patterns through the demands of the bargaining situation itself. Although these emergent higher functions are necessary for achieving the reward-seeking priorities that are defined by RCT, they can only approximate what, taking the long view, we would call rational.

## 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. (1992) *Picoeconomics: The Strategic Interaction of Successive Motivational States within the Person*. Cambridge U.
- Ainslie, G. (1995) A utility-maximizing mechanism for vicarious reward: *Rationality and Society* 7, 393-403.
- Ainslie, G. (2001) *Breakdown of Will*. Cambridge U.
- Ainslie, G. (2003) Uncertainty as wealth. *Behavioural Processes* 64, 369-385.
- Ainslie, G. (submitted) The self is virtual, the will is not illusory. *Behavioral and Brain Sciences*.
- 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. (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.
- Baumeister, R. F. and Heatherton, T. (1996) Self-regulation failure: An overview. *Psychological Inquiry* 7, 1-15.
- Baumeister, R. F., Heatherton, T. F., and Tice, D. M. (1994) *Losing Control: How and Why People Fail at Self-Regulation*. Academic.
- Beck, A. T. (1976) *Cognitive Therapy and the Emotional Disorders*. International Universities Press.
- Becker, G. and Murphy, K. (1988) A theory of rational addiction. *Journal of Political Economy* 96, 675-700.
- Berridge, K. C. and Robinson, T. E. (1998) What is the role of dopamine in reward: Hedonic impact, reward learning, or incentive salience? *Brain Research Reviews*. 28, 309-369.
- Boudon, R. (1996) The "rational choice model:" A particular case of the "cognitive model." *Rationality and Society* 8, 123-150.
- Bratman, Michael E. (1999) *Faces of Intention: Selected Essays on Intention and Agency*. Cambridge U.

- Breland, K. and Breland, M. (1961) The misbehavior of organisms. *American Psychologist* 16, 681-684.
- Brunner, D. and Gibbon, J. (1995) Value of food aggregates: parallel versus serial discounting. *Animal Behavior* 50, 1627-1634.
- Brunsson, Nils (1982) *The Irrational Organization*. Stockholm School of Economics.
- Cameron, O. G. (2002) *Visceral Sensory Neuroscience: Interoception*. Oxford U.
- Carlson, A.J. (1916) The relation of hunger to appetite. *The Control of Hunger in Health and Disease*, Chicago, Illinois: University of Chicago Press.
- Case, D. A. (1997) Why the delay-of-reinforcement gradient is hyperbolic. Paper presented at the 20th Annual Conference of the *Society for the Quantitative Analyses of Behavior*. Chicago, May 22.. [www.SQAB.psychology.org/abstracts-1997](http://www.SQAB.psychology.org/abstracts-1997).
- Coleman, J. (1986) *Individual Interests and Collective Action: Selected Essays*. Cambridge U.
- Corsini, R. J. (1984) *Current Psychotherapies*. Third Edition. Peacock.
- Darwin, C. (1872/1979) *The Expressions of Emotions in Man and Animals*.: Julian Friedman.
- 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.
- Dinsmoor, J. A. (1998) Punishment. In W. T. O'Donohue, Ed., *Learning and Behavior Therapy*. Allyn & Bacon.
- Freud, S. (1911/1956) Formulations on the Two Principles of Mental Functioning. In J. Strachey and A. Freud (Eds.), *The Standard Edition of the Complete Psychological Works of Sigmund Freud*. Hogarth, vol. 12.
- Gibbon, J. (1977) Scalar expectancy theory and Webers law in animal timing. *Psychological Review* 84, 279-325.
- Gillberg, C. and Rastam, M. (1992) Do some cases of anorexia nervosa reflect underlying autistic-like conditions? *Behavioural Neurology* 5, 27-32.
- Grace, R. (1996) Choice between fixed and variable delays to reinforcement in the adjusting-delay procedure and concurrent chains. *Journal of Experimental Psychology: Animal Processes*, 22:362-383.
- 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 Letters*, 43-51.
- Green, L., Fry, A., and Myerson, J. (1994) Discounting of delayed rewards: A life-span comparison. *Psychological Science* 5, 33-36.

- Green, L., and Myerson, J. (1993) Alternative frameworks for the analysis of self-control. *Behavior and Philosophy*, 21, 37-47.
- Harris, C. and Laibson, D. (2001) Dynamic choices of hyperbolic consumers. *Econometrica* 69, 535-597.
- 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. (1969) Method and theory in the study of avoidance. *Psychological Review* 76, 49-69.
- James, W. (1890) *Principles of Psychology*. Holt.
- Jolls, C., Sunstein, C. R., and Thaler, R. (1998) A Behavioral Approach to Law and Economics, *Stanford Law Review* 50, 1471-1550.
- Kahneman, D., and Tversky, A. (Eds) (2000) *Choices, values, and frames*. Cambridge U.
- Kavka, G. (1983) The toxin puzzle. *Analysis* 43, 33-36.
- Kirby, K. 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 Guastello, B. (2001) Making choices in anticipation of similar future choices can increase self-control. *Journal of Experimental Psychology: Applied* 7, 154-164.
- Kirby, K. N. and Herrnstein, R. J. (1995) Preference reversals due to myopic discounting of delayed reward. *Psychological Science* 6, 83-89.
- 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 childrens orientations toward a moral order: I. sequence in the development of moral thought. *Vita Humana* 6 11-33.
- Korobkin, R. and Ulen, T. S. (2000) Law and Behavioral Science: Removing the Rationality Assumption from Law and Economics, *California Law Review* 88, 1051-1144.
- Laibson, D. (1997) Golden eggs and hyperbolic discounting. *Quarterly Journal of Economics*, 62, 443-479.
- Loewenstein, George (1996) Out of control: Visceral influences on behavior. *Organizational Behavior and Human Decision Processes* 35, 272-292.
- McClennen, Edward F. (1990) *Rationality and Dynamic Choice*. Cambridge U.
- McConkey, K. M. (1984) Clinical hypnosis: Differential impact on volitional and nonvolitional disorders. *Canadian Psychology* 25, 79-83.

- Mazur, J.E. (1986) Choice between single and multiple delayed reinforcers. *Journal of the Experimental Analysis of Behavior* 46, 67-77.
- Mazur, J. E. (2001) Hyperbolic value addition and general models of animal choice. *Psychological Review* 108, 96-112.
- Metcalf, J. and Mischel, W. (1999) A hot/cool-system analysis of delay of gratification: Dynamics of willpower. *Psychological Review* 106, 3-19.
- Miller, W. R. (2003) Comments on Ainslie and Monterosso. In R. Vuchinich and N. Heather, Eds., *Choice, Behavioural Economics, and Addiction*. Pergamon, pp. 62-66.
- Mitchell, S. H. & Rosenthal, A. J. (2003) Effects of multiple delayed rewards on delay discounting in an adjusting amount procedure. *Behavioural Processes* 64, 273-286.
- Monterosso, J. and Ainslie, G. (1999) Beyond discounting: Possible experimental models of impulse control. *Psychopharmacology* 146, 339-347.
- Monterosso, J. R., Ainslie, G., Toppi Mullen, P., and Gault, B. (2002) The fragility of cooperation: A false feedback study of a sequential iterated prisoner's dilemma. *Journal of Economic Psychology* 23:4, 437-448.
- Navarick, D.J. (1982) Negative reinforcement and choice in humans. *Learning and Motivation* 13, 361-377.
- Orne, Martin T (1973) Communication by the total experimental situation: Why it is important, how it is evaluated, and its significance for the ecological validity of findings. In Pliner, Patricia; Krames, Lester; et. al, Eds *Communication and Affect: Language and Thought*. Academic.
- Pfohl, B. and Blum, N. S. (1991) Obsessive-compulsive personality disorder: A review of available data and recommendations for DSM-IV. *Journal of Personality Disorders* 5, 363-375.
- Posner, R. (1998) Rational Choice, Behavioral Economics, and the Law *Stanford Law Review* 50, 1555-1556.
- Rachlin, H. (2000) *The Science of Self-Control*. Harvard U.
- Russell, J.M. (1978) Saying, feeling, and self-deception. *Behaviorism* 6, 27-43.
- Ryle, G. (1949/1984) *The Concept of Mind*. U. Chicago.
- Samuelson, P.A. (1937) A note on measurement of utility. *Review of Economic Studies* 4, 155-161.
- Schachter, S., Silverstein, B. and Perlick, D. (1977) Psychological and pharmacological explanations of smoking under stress. *Journal of Experimental Psychology: General* 106, 31-40.
- Shizgal, P., and Conover, K. (1996) On the neural computation of utility *Current Directions in Psychological Science* 5, 37-43.

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

Skog, O.-J. (1999) Rationality, irrationality, and addiction. In J. Elster and O.-J. Skog (Eds.) *Getting Hooked: Rationality and Addiction*. Cambridge 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.

Sonuga-Barke, E. J.; Lea, S. E.; and Webley, P. (1989) The development of adaptive choice in a self-control paradigm. *Journal of the Experimental Analysis of Behavior* 51, 77-85.

Sorensen, R. A. (1992) *Thought Experiments*. Oxford.

Strotz, R.H. (1956) Myopia and inconsistency in dynamic utility maximization. *Review of Economic Studies* 23,166-180.

Sugden, R. (1991) Rational choice: a survey of contributions from economics and philosophy. *Economic Journal* 101, 751-785.

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

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

Thaler, R. H. (1988) The ultimatum game. *Journal of Economic Perspectives* 2, 195-206.

Thomsen, P. H., and Mikkelsen, H. U. (1994) Development of personality disorders in children and adolescents with obsessive-compulsive disorder: A 6 to 22 year follow-up study. *Acta Psychiatrica Scandinavica* 87, 456-462.

Tillyard, E. M. (1959) *The Elizabethan World Picture*.

Tversky, A. and Kahneman, D. (1981) Framing decisions and the psychology of choice. *Science* 211, 453-458.

---

<sup>i</sup> Much of bounded rationality seems to arise from pure cognitive error (Kahneman & Tversky, 2000). However, some reported examples probably arise from strategic motives, either serving self-control (as when people pay a premium to keep money in an illiquid account—Harris & Laibson, 2001) or evading it (for instance if the sunk cost fallacy evades a personal rule for recognizing loss—Ainslie, 1992, pp.291-293). The strategic approach presented here also provides a rationale for vicarious experience as a primary good, which can explain the apparent boundedness of self-interest (Ainslie, 1995, 2001, pp. 179-186).

<sup>ii</sup> By prospective reward I mean the affective salience of an anticipated reward, which can differ from any purely cognitive estimate of what the enjoyment of the reward will feel like—“wanting” as opposed to “liking” (Berridge & Robinson, 1998). It turns out that

the tendency to seek a reward can be dissected by pharmacological means from the intensity of its enjoyment, and is not necessarily proportional to it. This article is not based on neurophysiological data, but is compatible with the finding that the calculation of utility = affective salience can be tracked in the brain (Shizgal & Conover, 1996).

<sup>iii</sup> In actual play subjects often sacrifice their ostensible interests to punish others (Thaler, 1988), but in the intertemporal game being modeled the programmed contingencies encompass all incentives.

<sup>iv</sup> This disorder is not the same entity as obsessive-compulsive disorder (without the “personality”), which is an itch-like syndrome associated with low brain serotonin (Thomsen & Mikkelsen, 1994).