BEHAVIORAL AND BRAIN SCIENCES (2005) **28,** 635–673
*Printed in the United States of America*

# Précis of *Breakdown of Will*



**George Ainslie**
*Department of Veterans Affairs Medical Center, 116A, Coatesville, PA 19320 and Temple Medical College, Philadelphia, PA 19140*
**George.Ainslie@va.gov** **http://www.Picoeconomics.com**

**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 periods from minutes to 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: over-concern 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.

**Keywords**: altruism; appetite; behavioral economics; classical conditioning; compulsions; 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 do so. 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 main parts and twelve chapters, but have necessarily been selective in what I describe in detail. In the first part, "Breakdowns of Will" (sects. 2 to 4 here), I criticize the two main conventional approaches to impulsiveness and self-control (Ch. 2, sect. 2.1 here), then present experimental evidence that vertebrates' evaluation of future options is basically hyperbolic, rather than exponential as conventionally assumed (Ch. 3, sect. 3 here), and argue that the hyperbolic form offers an alternative to classical conditioning as a mechanism for involuntary behaviors (Ch. 4, sect. 4 here). In the second part, "A Breakdown of the Will" (sects. 5 to 8 here), I argue that hyperbolically based uncertainty about our own future choices leads us to see current choices as test cases (Ch. 5, sect. 5.1 here), that this perception establishes willpower through an intertemporal version of the repeated prisoner's dilemma (Ch. 6, sect. 6 here), that this model fits common experiences of will (Ch. 7, sect. 7 here), and that substantial evidence fa-

While a student at Harvard Medical School, GEORGE AINSLIE proposed that a recently discovered function describing choice among unpredictable, recurring rewards, Herrnstein's matching law, could be applied to discrete, predictable rewards in the form of a hyperbolic discount curve. Empirical demonstration of this "irrational" curve and its implications – choice that changes as a function of proximity, durable conflicting interests within the individual, and intertemporal bargaining among those interests – has been the work of nearly forty years. His behavioral and bargaining experiments and theoretical deductions have been published in journals of psychology, philosophy, economics, and law, in many book chapters, and in a previous book, *Picoeconomics: The Strategic Interaction of Successive Motivational States Within the Person* (Cambridge, 1992).

 0140-525X/05 $12.50

vors the bargaining model over other models of willpower (Ch. 8, sect. 8 here). In the last part "The Ultimate Breakdown of Will" (sects. 9 to 11 here), I describe how intertemporal bargaining leads to compulsive side effects (Ch. 9, sect. 9 here) and how a hyperbolically based impulse toward premature satiation of appetite gives emotions their quasi-voluntary quality (Ch. 10, sect. 10 here), and motivates the social construction of facts, the quest for vicarious experience, and indirect approaches to goals (Ch. 11, sect. 11 here). I summarize the conclusions of these arguments in section 12 (Ch. 12 of the book).

## 2. Breakdowns of will: The puzzle of akrasia (Part I, Ch. 1 of book)

### *2.1. The dichotomy at the root of decision science: Do we make choices by judgments or by desires? (Ch. 2)*

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 does not 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 (Fig. 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 [1992] "primrose path" to addiction), short time horizons (Becker & Murphy 1988), conditioned cravings (Loewenstein 1996), and recent discoveries about the neurophysiological process of reward (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, and short time horizons do not predict people's plans to avoid temptations when they face them from a distance. There is no reason to think that conditioned cravings should operate differently from other appetites, all of which have conditioned elements; and although 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 (Ch. 3)

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 et al. 1994b; Kirby 1997; Mazur 1997). The resulting hyperbolic discount curve is seen over all time ranges, from seconds to decades (Harvey 1994). This curve is a variant of Herrnstein's matching law as applied to delay (Chung & Herrnstein 1967), and 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 upon which most utility theories rely. 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 et al. 1994b; Kirby 1997) money. Investigators sometimes report that their data fit even 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 (Fig. 1B).

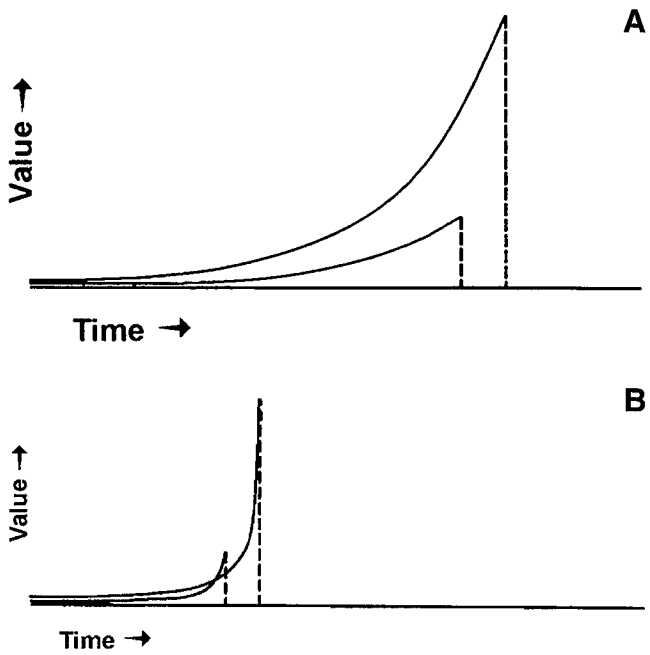


Figure 1.A. Conventional (exponential) discount curves from a smaller-sooner (SS) and a larger-later (LL) reward. At every point their heights stay proportional to their values at the time that the SS reward is due. B. Hyperbolic discount curves from an SS and an LL reward. The smaller reward is temporarily 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.

In contrast to exponential curves, hyperbolic discount curves depict a strong but temporary tendency to prefer smaller and sooner (SS) rewards to larger and 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), as well as 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 et al. 1994a; 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.

### *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 are 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 unchosen 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). Because 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 are 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 simply a preference, any more than the determinant of whether a piece of legislation becomes law is simply 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-micro-economics (hence, "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 (Ch. 4)

Because 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 deni-

able 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 minutes to days – we experience them as decisions. Even here, however, we may not feel fully responsible. An alcoholic learns that he 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 section 9 (Ch. 9 of the book). At the other end of the scale of durations, there are processes that usually feel involuntary but nevertheless have incentive value, positive or negative. These include 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. Because 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 here (see Ainslie 1992, pp. 19–22; Ainslie 2001, pp. 100–14). Because 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.

### 4.1. The problem of pain

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, choices 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 (1998) have described as "wanted" but not "liked," 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 these behaviors 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 (see 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 came to avoid being pulled into this undesirable pattern of responses by short-term rewards.

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 wide-

spread 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. 157–90), 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., McGaw 1966; Strasberg 1988), 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 (Fig. 2A). Each reward is dominant so briefly that it can command only attention, not a motor response (Fig. 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.

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 overall valence will be.

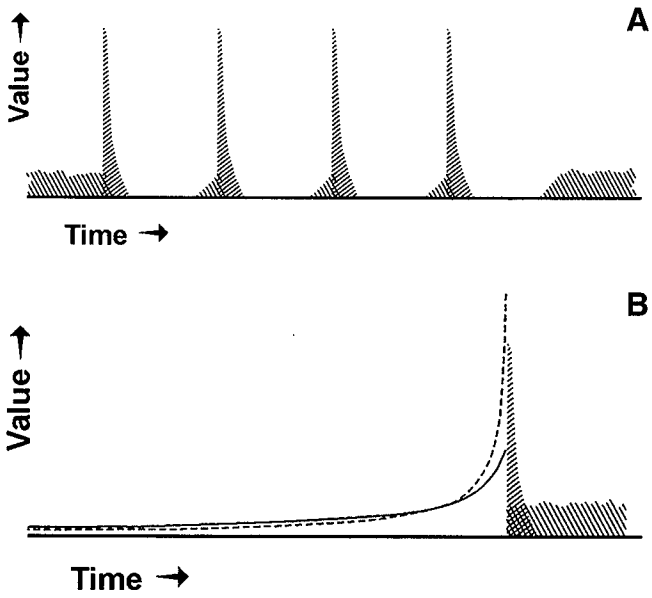


Figure 2.A. Aversion as a cycle of brief, intense reward (rightward hatching) that interrupts an on-going baseline reward (leftward hatching) for a relatively longer time. B. Hyperbolic discount curves drawn from a single spike in an aversion sequence such as that in Figure 2A. (Each curve is the sum of the curves from each moment of reward, see Fig. 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.

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 section 10 (Ch. 10 of the book).

## 5. A breakdown of will: The components of intertemporal bargaining (Part II of book)

### *5.1. The elementary interaction of interests (Ch. 5)*

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 0.5 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 0.5 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* 1147a; Aristotle 1984, pp. 24–28). Kant said that the highest kind of decision-making involves making all choices as if they defined universal rules (the "categorical imperative"; Kant 1793/1960, 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" (Sully 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 (Fig. 3A).

Experiments in both humans and rats have verified the predicted anti-impulsive effect of bundling choices together. Kirby and Guastello (2001) 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. The authors reported an even greater effect for SS versus LL amounts of pizza. Ainslie and Monterosso (2003a) reported that rats made more LL choices of sugar water when they chose for a bundle of three trials all at once than when they chose between the same SS versus LL contingencies on each separate trial. 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 (Fig. 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 to-

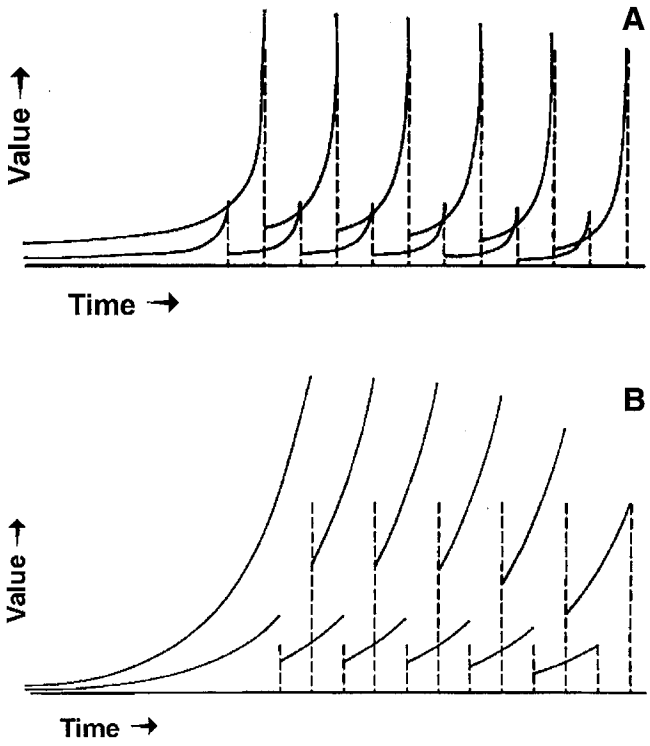


Figure 3.A. Summed hyperbolic curves from a series of larger-later rewards and a series of smaller-earlier alternatives. As more pairs are added to the series, the periods of temporary preference for the series of smaller rewards shrink to zero. The curves from the final (rightmost) pair of rewards are the same as in Figure 1B. B. 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.)

gether. 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.

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 (Bratman 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 versus feeling fit and having intact prospects Monday morning, or a good venting of rage versus 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. 155–62). 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.

## 6. Sophisticated bargaining among internal interests (Ch. 6)

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 (Kohlberg 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.

The problem with the bundling tactic is that there are

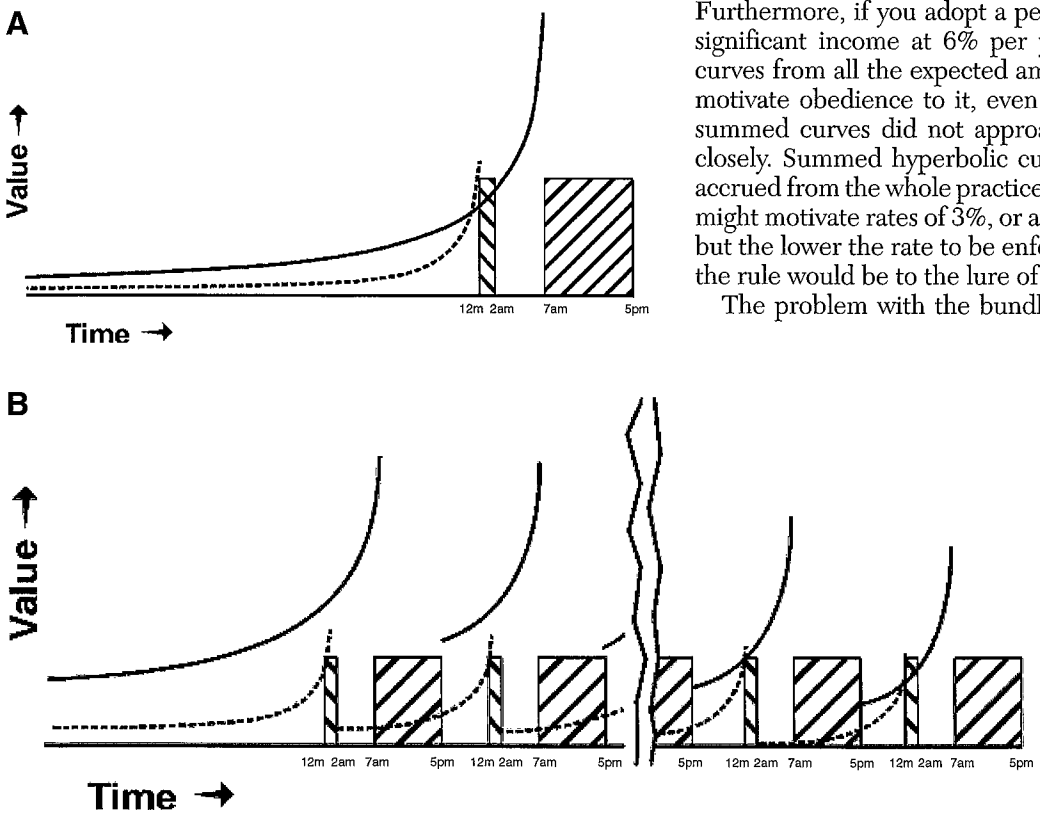


Figure 4.A. Curves that are the aggregate of hyperbolic discount curves from continuing rewards – staying up from midnight to 2:00 versus feeling rested from 7:00 to 17:00. B. Summed curves from ten pairs of the rewards depicted in Figure 4A. The effect of summation is the same as for the point rewards in Figure 3A.

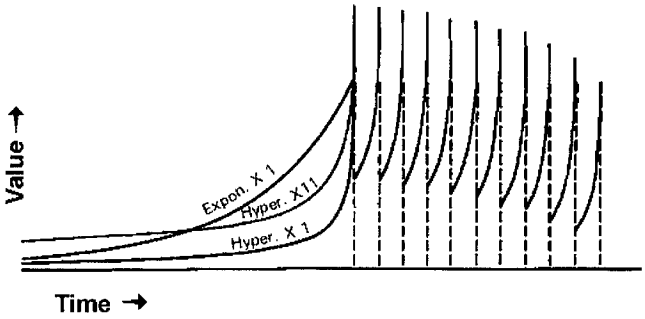


Figure 5. Summed hyperbolic discount curves from eleven rewards, compared with a single exponential curve and a single hyperbolic curve such as shown in Figure 1. The summed curves come closer to exponential discounting than the lone hyperbolic curve does in the crucial sector of the curve where delay is low.

many possible personal rules. The ice cream at hand may violate one diet but not another; and even if it is so outrageously rich as to violate all conceivable diets, there is apt to be a circumstance that makes the present moment an exception: It is Thanksgiving dinner or my birthday, or a host has taken special trouble to get the ice cream, or I have cause to celebrate or to console myself just today, and so forth. The molar principle that offers an exception just this once will be rewarded more than the one that does not, for it predicts the aggregation of LL rewards (as in Figs. 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 who has an overeating problem can't simply estimate whether she is 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 making 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 (Klein & Leffler 1981; Macaulay 1963). This recursive process of staking the credibility of a resolution on each occasion when it is tested gives your resolve momentum over successive times. The on-going 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 mere 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 risk 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 (Bratman 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 (Ch. 7)

Analyzing an activity that is second nature inevitably enlarges some features and slights others, so that the resulting picture seems foreign to familiar experience. Charac-

terizing 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 (Ch. 8)

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 and 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., Becker & Murphy 1988; Ryle 1949/1984). 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 that 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 section 5.1; 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 (McClennen 1990, pp. 157–61), and Bratman refers to "a planning agent's concern with how she will see her present decision at plan's end" (Bratman 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 a 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 example, when Kirby and Guastello (sect. 5.1) 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 (Kirby & Guastello 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 is Kavka's problem.

### 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 is 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 one will 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 is 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, because 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 will act over time.

Kavka's contribution has been to create a conceptual irritant that cannot 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: (1) 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. (2) 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.

## 9. The ultimate breakdown of will: Nothing fails like success (Part III of book)

### *9.1. The downside of willpower (Ch. 9)*

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 are 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.

### *9.2. 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, perhaps the loss of authenticity that existential philosophers complain of in modern society generally.

It is 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.

### *9.3. 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 have 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, because 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. 193–97). Alternative models of self-control failure based on exhaustion of "strength" (Baumeister & Heatherton 1996) or an opponent process (Polivy 1998), do not account for regular failure that is specific to a particular circumstance.

### 9.4. 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 that 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.

### 9.5. 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 is the wrong decision. On the contrary, you would 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 (Figs. 3A and 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. 417–23, 669–73).

So, cooperation among successive motivational states does not 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 have defined, but reduce your sensitivity to less well-marked kinds of reward.[2]

### 9.6. 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 (Ch. 10)

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 (sects. 4 here); 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 (sect. 4 here) that the seduction of attention is how negative emotions impose themselves on people who don't want them. I shall 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 section 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 is available unconditionally, as emotion is, it should never induce significant motivation to obtain it. As Adam Smith originally observed (Smith 1776/1976, pp. 44–45), 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 (sect. 4 here) 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 as 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 is 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 is used up as reward occurs and that cannot 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 versus 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 is more intense the faster it is consumed, they will 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 the 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 uncertain to occur or 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., Berns et al. 2001; Hollerman et al. 1998).

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 is 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.

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 does not. 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 (Frank 1988): Expressions that are known to be intentionally controllable are disregarded, as with the false smile of

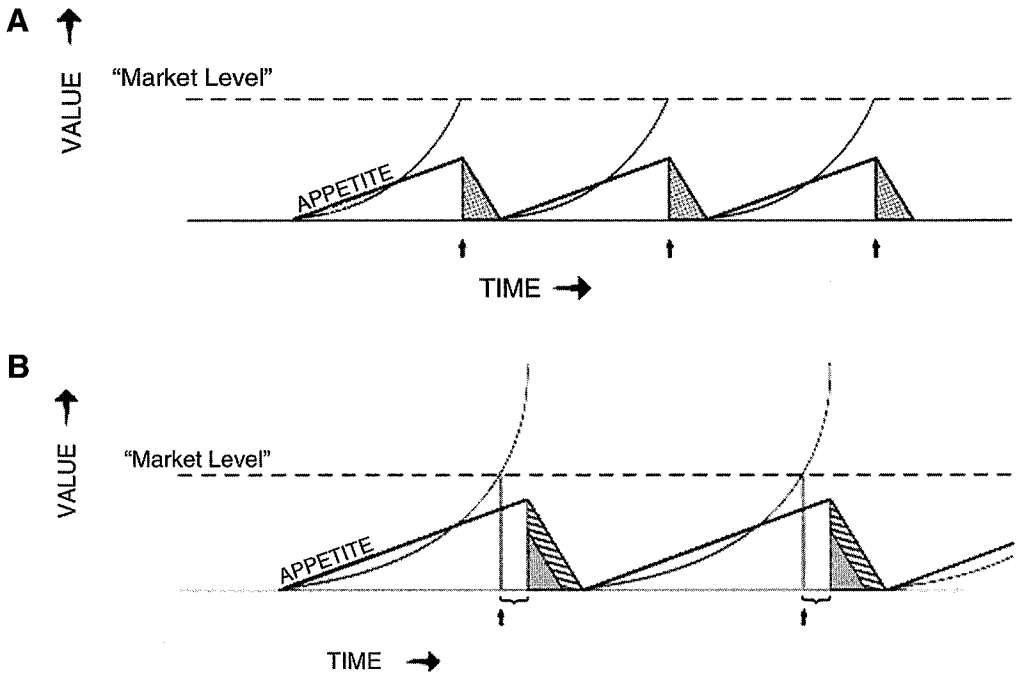


Figure 6.A. Repeated cycles (not summed) of growing reward potential ("appetite," depicted schematically by the straight lines) and actual consumption to the point of satiety (gray areas). Consumption begins at the points (arrows) when discounted value of expected consumption reaches the competitive 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). B. 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]

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 is 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 are young and inept and when you have 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 had 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 (Ch. 11)

I'll argue that the other three puzzles also hinge upon premature satiation, the impulse to harvest emotional reward before it is 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 section 9 (Ch. 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 is 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 cannot 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 is 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 previous section, 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 because 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. Making predictions about 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 is 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 would not qualitatively change the experience of succeeding or failing. But when the puzzle is built like the person solving it – that is, when it is 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. Because 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 is disciplined by observation – it is apt to behave much more like the actual other person than a nonempathic 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 as your own the criteria that you think she is using to occasion emotion. For the time being you entertain her emotions. But of course, they are hers only in the sense that you are having them according to a theory about her. *You* are the person through whose brain they are percolating. This means that you can use such a model to occasion emotions just as you use your own prospects.

Because 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 are modeling in the other person as substantially as the ones you have as yourself. To model 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. 183–86). 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 have gambled on another person's discernable feelings, these feelings should become a good that we will 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 cannot 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 have described how the will cannot 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 to 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 have 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 do so.

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 (Ch. 12)

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 the 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 have 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.

ACKNOWLEDGMENTS

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

NOTES

The author of this Précis is employed by a government agency and, as such, this Précis is considered a work of the U.S. government and not subject to copyright within the United States.

**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" (Shizgal & Conover 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.

# Open Peer Commentary

## Models of preference reversals and personal rules: Do they require maximizing a utility function with a specific structure?

Horacio Arló-Costa

*Philosophy Department, Carnegie Mellon University, Pittsburgh, PA 15213.*
**hcosta@andrew.cmu.edu**
**http://www.phil.cmu.edu/faculty/arlocosta/**

**Abstract:** One of the reasons for adopting hyperbolic discounting is to explain preference reversals. Another is that this value structure suggests an elegant theory of the will. I examine the capacity of the theory to solve Newcomb's problem. In addition, I compare Ainslie's account with other procedural theories of choice that seem at least equally capable of accommodating reversals of preference.

One of the ambitions of Ainslie's book is to develop a *descriptive* theory of choice and intentionality (Ainslie 2001). The theory abandons some of the main tenets of Rational Choice Theory (RCT) although it is still compatible with explanations in terms of maximization of time-indexed utility.

The main departure from RCT is based on Ainslie's conviction that discount curves are not only nonexponential, but also "specifically hyperbolic" (p. 31). Although this hypothesis has some cur-

rency among economists, the idea is not completely consensual. The phenomenon of preference reversals is certainly an empirically robust fact. But as Rubinstein (2003) and other economists have recently remarked, the adoption of hyperbolic discounting does far more than just modeling the psychological phenomenon that the present has a special status. "It assumes the maximization of a utility function with a specific structure and as such it misses the core of the psychological decision-making process. Thus, I find it to be no more that a minor modification to the standard discounting approach" (Rubinstein 2003, p. 1215).

The objection raised by Rubinstein is that the same sort of evidence provided by preference reversals can also reject hyperbolic discounting as well; although most of these phenomena can be explained in terms of a decision procedure based on similarity relations, Rubinstein's own proposal constitutes an even more severe departure from RCT (e.g., transitivity is not satisfied). It has nevertheless the merit of delivering an account of other important choice anomalies that inspired the main theories of nonexpected utility.

Of course, a sophisticated theory of rationality cannot be based on an unmodified version of RCT. This is so in virtue of *normative* reasons alone. For example, some phenomena, like Ellsberg's, seem to require a normative relaxation of RCT of the sort advocated by Amartya Sen (2002) or Isaac Levi (1986). An adequate descriptive theory of choice should also be able to explain problematic patterns of choice of this kind. But it seems doubtful that the adoption of a particular functional form for utility can accomplish that. The so-called theories of nonexpected utility, which also relied on the use of special utility curves, have been relatively successful in explaining only a limited range of recalcitrant phenomena (e.g., Ellsberg's puzzle cannot be explained by appealing to the special utility and probability curves of prospect theory; especial and questionable ad hoc hypotheses have been used instead by Fox and Tversky [1991] – see Arló-Costa and Helzner [2005] for a critical appraisal of this account).

In Chapter 4 of Ainslie's book, hyperbolic discounting is applied via ingenious arguments to develop accounts of emotions, pain, and other aversive behavior. The extension requires the neat separation of the notions of reward and pleasure. Therefore, the theory is quite different from the standard hedonic articulations of rational choice. In addition, the model presents an account of emotions as "pulled" by reward, which seems, prima facie, controversial.

But the bulk of the book is devoted to showing that one can develop a model of the will as emerging from a process of intertemporal bargaining among units called *interests*. This is one of the most imaginative and interesting parts of the book. Perhaps the most charitable account of interests is as time slices of the self.

Stroz concluded in his seminal paper (Stroz 1955) that "the individual always decides what to do on the assumption that he has no authority over his future self" (p. 180). Since then, many philosophers have questioned this assumption by postulating that agents can bind themselves through the operation of their wills. This strategy *postulates* rather than explains intentionality. Ainslie suggests instead that when someone seems to be choosing according to principle (or by using a personal rule), what literally happens is that his successive selves form a repeated prisoner's dilemma relationship, which is solved in the manner of interpersonal bargainers. So the idea here is to explain the will away, rather than invoke it to explain commitment.

Ainslie concedes that the existence of the resulting internal feedback process is probably impossible to study via controlled experiments. Nevertheless, he claims that postulating hyperbolic discounting (and therefore recursive decision-making) solves some well-known philosophical conundra on intentionality and choice. I only have space here to focus on one of them: the so-called Newcomb's problem.

As originally formulated, Newcomb is a single-play situation and the puzzle is to determine what is *normatively* required in this case. Ainslie seems to think (citing Nozick 1969) that in this case RCT requires choosing both boxes (p. 134). But Nozick's argument insists that if one maximizes "evidential" expected utility (EU), one should choose one box. Dominance is invoked to rationalize the two-box solution. Later on, deviant versions of RCT (causal versions) have been developed in order to articulate the latter solution (Joyce 1999). Presystematically, Ainslie seems to be a two-boxer (p. 137). But this conclusion is deeply controversial (see, e.g., Levi 1975; Meek & Glymour 1994), and it does not seem to be based on using hyperbolic discounting or recursive decision-making.

In addition, the author argues that the temptation to hedge on a personal rule could be modeled as a problem with the same arithmetic structure as a repeated version of Newcomb. In his model, he argues, diagnostic acts are also causal acts and one can then explain the sequential analogue of the "evidential" solution in the single-play case (cooperating). But this does not seem to explain why Ainslie defends the solution analogous to mutual defect in the one-shot case of Newcomb. And this is the gist of Newcomb's problem.

Ainslie's very ambitious idea is to open the impenetrable black box of intentionality, which is then modeled as a sort of brokerage process. But this process is hard to dissect on account of its recursive nature. I do not think that the appeal to philosophical puzzles like Newcomb's adds credibility to it. Many of these puzzles share with Ellsberg's example the feature of being thought experiments designed to uncover *normative* inadequacies of RCT. And the theory presented in the book, like many other descriptive theories of choice, seems unable to deal with this type of problems. Ultimately, the plausibility of Ainslie's theory seems to rest on how well it fares in comparison to other attempts to "open the black box of decision."

The final chapters of the book (on the "dyscontrol" symptoms that can be induced by bargaining strategies) are full of fascinating insights. Probably, Ainslie offers in this book one of the most complete and theoretically unified theories of the will available today. The resulting overall picture is certainly quite impressive.

## Three other motivational factors

Kent Bach

*Department of Philosophy, San Francisco State University, San Francisco, CA 94132.* **kbach@sfsu.edu** **http://online.sfsu.edu/~kbach**

**Abstract:** Ainslie uses his hyperbolic discount model to explain a dazzling array of puzzling motivational phenomena. In so doing, he assumes that the motivational force of a given option at a given time is directly proportional to its discount-adjusted reward as assessed at that time. He overlooks three other factors which, independently of the perceived reward, can affect motivational force.

Ainslie (2001) assumes that the motivational force of a given option at a given time is directly proportional to its discount-adjusted reward as assessed at that time. Evidently, he rejects any independent role for cognition in mediating or arbitrating competing rewards or in deliberating and deciding what to do when. Rather, he conceives of these interests as a population of quasi-independent agents engaged in tacit bargaining, each aimed at its own temporally discounted reward. He argues that the curves representing this discounting are "highly bowed," hyperbolic rather than exponential in form, thereby allowing for temporary reversals of preferences. He briefly mentions four other discounting patterns (p. 208), but, mathematically speaking, there are countless others consistent with temporal reversal. Indeed, perhaps the discounting takes different shapes for different rewards, and maybe the discounting is sometimes non-monotonic, as with highly unstable desires. But let's assume that Ainslie's contention that they are hyperbolic is not hype and that he is not taking us for pigeons.

The key idea is that hyperbolic discounting, by devaluing future rewards and punishments (negative rewards) proportionately to their delay, lets "utility theory move beyond its stalemate with cognitivism" (p. 38). So-called "dynamic inconsistency" is really a side effect of the fact that the discount curves for different rewards can cross: "people will naturally go for smaller, earlier over later, larger rewards. . . . Akrasia is just maximizing expected reward, discounted in highly bowed curves" (p. 39). In explaining how preferences can temporarily reverse, this simple model eliminates the apparent mystery of how we can act against our better judgment and against our "true" interests. Ainslie contends that hyperbolic discounting can explain a host of phenomena, including impulsiveness, addiction, compulsion, ambivalence, procrastination, and back-sliding. It can explain "the irony of smart people doing stupid things or having to outsmart themselves in order not to" (p. 27), by adopting "personal rules," cultivating good work habits, and making commitments that increase the cost of yielding to temptation. Consideration of future rewards doesn't take us outside the realm of reward and require higher-level judgments. There is just the ongoing competition among the rewards themselves.

It may seem an exaggeration to treat a person's different values as autonomous agents engaged in intertemporal bargaining with one another. Ainslie himself recognizes that his seemingly schizophrenic model of "the self as a population" (p. 39) makes it puzzling how, in a dog-eat-dog world of competing bargaining agents, "a marketplace of hyperbolically discounted choices [could] ever come to look like a single individual" (p. 40) rather than a kennel. What most worries me, though, is something else: At a given time, the motivational force of a desire (drive, urge, goal, value, or whatever you want to call the members of this population) is not a function merely of the reward associated with it, even as adjusted for the odds of success and the cost of failure and as temporally discounted.

Here are three other factors that can contribute to the motivational force of a particular desire: (1) the frequency and persistence of the desire's coming to mind, (2) the desire's degree of insatiability, and (3) its resistibility to the second-order desire to get rid of it. Each of these factors can vary even as the perceived reward of what is desired stays the same. For example, (1) something you want but deem of little importance can be more zealously pursued simply because the thought of it keeps occurring to you and capturing your attention. Playing another video game when you are trying to finish writing a commentary does not seem all that important, but its urgency is enhanced just because the thought of doing it keeps occurring to you. This doesn't make it seem like a better thing to do. Rather, you think you had better do it. What is rewarding is not playing the game but eliminating the clamoring desire to play it. Even worse, sometimes (2) the thought of playing the game does not go away after you play it. You did what you wanted to do, but now you want to do it again, just as if you hadn't done it in the first place. It's not that you want it more but that you want it again. In such a predicament, you may desire to make this desire go away, but (3) try as you may, it keeps rearing its head, keeping you from concentrating on that commentary. Now you're back to square (1), and vulnerable to (2) and (3) all over again.

You could have the opposite problem, say with a long-term project that requires irregular but frequent attention over time. As much as you value the ultimate reward of cultivating your garden, for example, (1) the thought of doing even a little puttering does not occur as frequently as it should. Not only that, (2) when it does occur and you act accordingly, the mere satisfaction of doing a little something makes you feel as though you have made significant progress even though you haven't. And, to make matters worse, (3) the thought of cultivating your garden resists staying in mind even when you want it to.

It might seem that these three factors are reducible to the perceived size of the desired reward. However, to suppose that would confuse the assessed size of the desired reward with these three independent dimensions of strength of the desire itself.

I have three further worries about the explanatory depth and breadth of the hyperbolic discount model. First, it ignores the distinction between something's being desired because it is rewarding and something's being rewarding because it is desired. Second, this model does not explain the magnitudes assigned to particular rewards in the first place. And third, the fact, if it is a fact, that discounting is hyperbolic itself cries out for an explanation.

## Hyperbolas and hyperbole: The free will problem remains

Bruce Bridgeman

*Department of Psychology, University of California–Santa Cruz, Social Sciences 2, Santa Cruz, CA 95064.* **bruceb@ucsc.edu**
**http://psych.ucsc.edu/faculty/bruceb/**

**Abstract:** Hyperbolic theories have the fatal flaw that because of their vertical asymptote they predict irresistible choice of immediate rewards, regardless of future contingencies. They work only for simple situations. Theories incorporating intermediate unconscious choices are more flexible, but are neither exponential nor hyperbolic in their predictions. They don't solve the free will paradox, which may be just a consistent illusion.

Will a hyperbolic theory of reward discounting solve the persistent problems of the role of will in governing behavior? Ainslie (2001) makes a case for that idea, but hyperbolic theories have problems of their own. The hyperbola is defined by two perpendicular asymptotes, in this case one at the baseline of zero reward value and the other at the time of reward. Inevitably, a hyperbolic theory must predict reward value approaching infinity as the function moves toward the moment of reward. Rewards should become irresistible, no matter what other conditions apply, when the moment of reward gets very close. The exponential curve, in contrast, always has a finite value at a given time. True, the hyperbola has a predictive advantage over most of its range, where it predicts lower reward value (more discounting of future rewards) than exponential models, but the choice of either conic section is more a matter of mathematical convenience than theory, because no reasonably developed theory of reward motivates either model.

Another problem with hyperbolic theories became clear to me when I applied the only hyperbolic model that I have developed during my career, an ideal-observer model that predicts reading rate at any rate of display flicker, as on a CRT monitor. It was a harrowing experience, because once the positions of the two asymptotes have been established (in this case at zero display frequency and reading speed at infinite frequency) there is only one free parameter left, a scaling parameter. It's not much to go on. The model worked well enough, though, enabling us to predict reading speed at one frequency with great precision, given only the reading speed at another frequency (Montegut et al. 1997).

The hyperbolic model of reading rate worked because it dealt with a low-level stimulus sampling issue, not with deep psychological issues of reward and choice. The hyperbola simply doesn't leave enough room to take complexities into account. Recognizing this, Ainslie proposes that reward-delay decisions involve a whole cascade of hidden intermediate decisions, each with its own hyperbolic function that is replaced by another hyperbolic function at the time of an intermediate decision. The resulting predictions of future reward value are neither exponential nor hyperbolic, but depend on the timings of the intermediate decisions. Each intermediate decision brings with it a new free parameter, making the model more predictive but less theoretically useful, because the number of free parameters expands faster than the number of predicted actions.

Ironically, the decision cascade idea highlights an essential paradox of linear decision theories – decisions are themselves nonlinearities, places where everything that has gone before is applied to a single binary choice: you either accept the reward or you

don't. After each binary choice, represented by the vertical lines in Ainslie's diagrams, the data that previously went into the choice are discarded, and the process starts again. This is the sense in which decisions are nonlinear. And it's difficult to make a linear theory of nonlinear behavior.

Even this expanded theory, however, fails to resolve the deepest paradox of reward theory, the issue of what seems like free will. Even if the intermediate decisions are driven by utility theory or by some other deterministic algorithm, we still do not know how that last decision gets made. Substituting intertemporal bargaining within an individual adds flexibility, but there still must be rules governing each bargaining agent. The paradox of free will remains, the two alternatives of choice determined either by some unknown influence or by uncaused action. In the first alternative, free will is an illusion, a feeling of free choice in an environment where the decision is in fact determined by unconscious information-processing in the brain. The second flies in the face of everything we know about the physical world. Indeterminacy from random or chaotic processes doesn't solve the problem, for it only adds a bit of noise to the reasoning process, whether conscious or unconscious. And noise is not the same as free will; even in introspection, we don't confuse lack of control with freedom. The only alternative left is the uncomfortable first alternative, that is, that free will is indeed an illusion, but since it is a consistent illusion, it is accepted as reality.

The idea that consistent illusions are perceived as reality has precedents from more prosaic, but better-defined, domains in perceptual research. Illusions can be defined as situations that change upon closer inspection, whereas reality remains the same upon closer inspection. Length illusions, such as the Müller-Lyer arrows-in versus arrows-out figure, for example, can be tested easily by measuring the two lines in question, or by superimposing them for direct comparison. The illusion becomes obvious, but even after years of experience the figures still appear to be of distorted lengths. Other illusions are more difficult to expose as illusions. The slopes of hills, for instance, are grossly overestimated by most people, who will go through their lives believing that the steepest streets in San Francisco are perhaps 45 degrees, when they are actually about 10 degrees. If no one corrects them on this, the 45 degrees is reality for them, with none of the conflict that pertains to illusions when they are exposed. Analogously, if the unconscious information-processing that goes into decision-making is never exposed, people can go through their lives believing that their thought processes are guided by free will, and never be confronted by the paradoxes of uncaused action or hidden determinism. Still, a theory that can help us to predict behavior in serious situations, such as addictions, is a pragmatic step forward.

ACKNOWLEDGMENT

Preparation of this commentary was supported by a Faculty Research Grant from the Academic Senate of the University of California, Santa Cruz.

## Regret and the control of temporary preferences

Terry Connolly[a] and Jochen Reb[b]

[a]*Department of Management and Policy, The Eller College, University of Arizona, Tucson, AZ 85721;* [b]*Singapore Management University, Lee Kong Chian School of Business, Singapore 178899.* **connolly@email.arizona.edu jochenreb@smu.edu.sg**

**Abstract:** Regret is often symptomatic of the defective decisions associated with "temporary preference" problems. It may also help overcome these defects. Outcome regret can modify the relative utilities of different payoffs. Process regret can motivate search for better decision processes or trap-evading strategies. Heightened regret may thus be functional for control of these self-defeating choices.

In Ainslie's taxonomy of "temporary preference" problems, the defining feature of addictions is that "the imminent prospect of such activities is strongly rewarding but they're avoided if foreseen from a distance *and regretted afterward*" (Ainslie 2001, p. 49; italics added). In compulsions, similarly, "regret may still occur . . . and the person may even expect the regret while indulging in the behavior" (p. 50), but the behavior still persists. Regret, then, may accompany, or even define, the problematic behaviors. What Ainslie appears to overlook is the possibility that regret may help to control them.

The "temporary preference" problem involves, minimally, the integration of two payoffs, one of which arrives earlier than the other, and is thus prone to the immediacy or hyperbolic discount effect. Viewed from a suitable distance before or after the choice, the later option is superior. Close to the decision point the earlier option is (temporarily) more attractive, a phenomenon that, as Ainslie notes (p. 198), "can't be changed by insight per se." The addict surrenders to the overwhelming desire for a fix, and the compulsive for yet another hand-wash, despite a clear intellectual understanding that she would, on balance, prefer not to and that the temporary urge will pass. Intellection is overpowered by emotion, System 2 by System 1 (Kahneman 2003).

Although emotions commonly involve both cognition and feeling (Frijda 1988), regret does so to an unusually large extent (Landman 1993). Asked if we feel regret over our choice of job, spouse, or vacation, most of us would reply "Let me *think*." We comfort a friend torn by feelings of regret by offering consoling *thoughts*: "How could you have known?" "You made a careful choice, there's nothing to blame yourself for." This intimate interweaving of thought and feeling has made regret the variable of choice for decision theorists interested in emotions. Perhaps it has a role to play in the complete understanding of temporary preference problems.

It is useful to distinguish two targets of decision-related regret: (1) regret associated with the outcome of a decision, and (2) regret associated with the choice process itself (Connolly & Zeelenberg 2002). The former seems to be essentially a reference-point phenomenon, in which the value of an outcome is reduced (or, if positive, enhanced: e.g., rejoicing) by comparison with some reference point, commonly the outcome of a foregone alternative (see Bell 1982; Loomes & Sugden 1982; Mellers et al. 1999). The second seems to involve a stronger component of self-blame or remorse, and is tied to the feeling that the decision made or the process used in making it was, in retrospect, insufficiently justified. The two regrets may be compounded, as for a mother who feels both outcome regret at the sickness of her small child, and self-blame regret at not having thought more carefully about his medical care (Reb & Connolly 2005). The failed dieter regrets both the additional weight gained and the poor decision about the chocolate cake.

There is abundant evidence that anticipated regret can influence decisions in a variety of domains, including medical care (Connolly & Reb 2003), consumer decisions (Simonson 1992), and negotiations (Larrick & Boles 1995). Richard et al. (1996) report some success in one temporary preference context, curbing unsafe sexual behavior, by shifting time-frames and making regret salient. They asked their respondents about either their feelings *about* unsafe sex or the feelings they would anticipate *after having had* unsafe sex. Participants in the second condition reported "safer" behavioral expectations immediately, and less actual risky sexual behavior in the six months following the experiment. How, exactly, might such a manipulation of regret salience achieve this promising result?

Two mechanisms might be suggested, paralleling the two sorts of regret described above. One possibility is that the manipulation enhanced outcome regret associated with the smaller, sooner behavior (unsafe sex), lowering its payoff value, and/or increased the larger, later payoff value by adding a component of rejoicing. It is clear from the conventional portrayal of discounted payoff values (e.g., Ainslie 2001, Fig. 4, p. 63) that modestly lowering the ear-

lier payoff or raising the later one could resolve the preference reversal in favor of the later behavior.

A second, perhaps complementary, mechanism might rely on the self-blame, process-oriented component of regret. We have shown in recent work (Reb 2005) that making regret salient to experimental participants can lead them to use more careful decision processes, acquire more decision-relevant information, and deliberate longer before deciding. Perhaps the participants in Richard et al.'s study responded to the regret-salience manipulation by searching more diligently for alternative choices, weighing the costs and benefits of the unsafe behavior more carefully, or considering one of the familiar self-control strategies discussed by Ainslie (p. 73ff).

The hypothesis, then, is that regret can be more than a mere symptom of failed decision making. Regret may, in some circumstances, play a role in improving decisions: the experience of regret can drive learning in repeated decisions; its anticipation can shape single decisions. Outcome regret affects decisions by modifying the relative attractiveness of different payoffs. Process-related regret does so by motivating the search for trap-evading strategies such as decision bundling, precommitment, and the like. In both cases, the interweaving of thought and feeling that characterize regret provide the bridge between System 1 and System 2 processes, between the thoughtful appraisal of the distant goal and the visceral appeal of the immediate indulgence. Without venturing into evolutionary speculation that regret may have developed to serve such a system-bridging purpose, it is not difficult to see that some level of regret can be highly functional for control of the self-defeating processes that temporary preference problems represent. The hypothesis seems to us worthy of serious consideration.

## The will: Interpersonal bargaining versus intrapersonal prediction

Luca Ferrero

*Department of Philosophy, University of Wisconsin–Milwaukee, Milwaukee, WI 53201-0413.* **ferrero@uwm.edu** **http://www.uwm.edu/~ferrero**

**Abstract:** Ainslie is correct in arguing that the force of commitments partly depends on the predictive role of present action, but this claim can be supported independently of the analogy with interpersonal bargaining. No matter whether we conceive of the parties involved in the bargaining as interests or transient selves, the picture of the will as a competitive interaction among these parties is unconvincing.

I am unpersuaded by Ainslie's central claim that the will is the product of transtemporal bargaining among successively dominant, transient interests analogous to the emergence of cooperation in a repeated Prisoner's Dilemma (Ainslie 2001, pp. 90–93). It is questionable that we could make sense of the parties involved in this bargaining as truly separate sources of agency. And even if we could, it is hard to see how cooperation could emerge out of the interactions between these parties. It is doubtful, therefore, that the will can be understood as a genuine *inter*personal phenomenon.

Consider the alleged competition among separate interests. Ainslie presents the interests as independent agencies that strive for selection (Ainslie 2001, pp. 39–41, 61, 73) as if they were replicators in a process of natural selection. But he gives no reason to believe that there are heritability and differential fitness in the competition among interests. What the selection amounts to is just that the strongest interest is satisfied at the expense of the conflicting, weaker interests. This satisfaction does not alter the chances of any interest to reappear with equal strength in the future. Nor does it promote the development of any adaptive strategy by the interests themselves. Understanding of the effects of hyperbolic discounting does not depend on the unwarranted reification of the agent's preferences into independent sources of agency that compete strategically in a genuinely selective process. Talk of selection among the preferences and the development of strategies to deal with the conflicts of preferences is more appropriate at the level of agents, even if the agents operate in response to the varying strengths of their preferences. In any event, interaction among interests could not explain the emergence of commitments. A short-range interest has no incentive to submit to a commitment, because commitments preclude the interest's present and future satisfaction. Interests seek nothing other than their satisfaction, hence nothing can be offered to them in exchange for their frustration.

What if the parties are not interests, but successive temporal selves? Ainslie occasionally shifts from talk of transient interests to talk of temporal selves (Ainslie 2001, pp. 40, 93, 161). The two notions are not identical, however. Contrary to transient interests, temporal selves are sources of agency and can have multiple interests. It seems that temporal selves might agree to be under commitments that frustrate their dominant short-range preferences in exchange for the satisfaction of other preferences. However, temporal selves are transient, hence they have no incentive to settle for less than the satisfaction of their short-range dominant interest.

The problem would not arise for parties that are transient in the sense that they act just once, but have stakes in the long-term outcomes of their actions (see Ainslie 2001, p. 93). These parties have no problem seeing the long-term benefits of a commitment. Nevertheless, they are tempted to make an *exception* now, thereby satisfying their dominant short-range interest while still reaping the long-term benefits of future compliance. However, if the present action counts as a precedent, a single exception to the commitment is self-defeating, given that present defections invite future ones. For Ainslie, transient interests/selves happen to be related so that their actions count as precedent for future ones, whence the stability of commitment. However, the fact that transient parties with long-term stakes can strategically agree to cooperate does not explain the will. First, there is no need to look at interpersonal scenarios to appreciate that actions can work as *intra*personal precedents. Given that the *same* agent is going to face the *same* choice at the future time with the *same* set of preferences, it is not surprising that her present action is a precedent for her future ones, thereby defeating temptations to make exceptions to her commitments. This is not really *strategic* thinking, but just reflection on the import of one's present action in the context of one's continued existence as one and the same agent who is going to face exactly similar choices in the future. Appeal to transient selves adds nothing to this straightforward *intra*personal explanation. Moreover, in order to make the repeated Prisoner's Dilemma scenario envisaged by Ainslie (2001, p. 93) truly explanatory, special interpersonal conditions must be assumed: The parties must face exactly the same choice over time and share the same long-term preferences. But these conditions are not special from the intrapersonal point of view. They are just distinctive features of the agent's temporal identity.

Second, the fact that we are subject to hyperbolic discounting and thus prone to inconsistent shifts in short-term preferences is no reason to think that we are made up of competing transient selves with long-term stakes. What makes this false picture attractive is the misleading focus on scenarios like "Ulysses and the sirens" as if they were paradigmatic of diachronic agency. Ulysses' situation, however, is unusual. When Ulysses listens to the sirens, he does not just reverse his short-range preferences, rather, he is also insensitive to long-term considerations. Hence, he does not care that his action could be a precedent. But this makes him impervious to commitments. He can be controlled only by physical restraints or short-range disincentives. If hyperbolic discounting were to make us always like Ulysses, our lives would indeed be best described in interpersonal terms. But then there would be no will, just crude transtemporal manipulation. On the other hand, if temporal selves are depicted as having not just shifting short-term

preferences, but also stable long-term ones, then they have no explanatory force. They are just like ordinary extended agents with stable long-term preferences, except that they act just once. Speaking of an agent as made up of a succession of these selves is just a convoluted way of saying that, as a result of hyperbolic discounting, we are going to have dynamically inconsistent short-term preferences.

Ainslie is right in claiming that we are subject to hyperbolic discounting and that the will is partly a matter of predicting future conduct from one's present action. But this conclusion is independent of the misleading picture of a bargaining among competing interests/selves. The will is not an organ, but it is not a bargaining situation either. It is rather the product of the agent's reflection on the long-term effects of her actions, including their role as precedents, under the assumption of her own identity over time.

## Hyperbola-like discounting, impulsivity, and the analysis of will

Leonard Green and Joel Myerson

*Department of Psychology, Washington University, St. Louis, MO 63130.*
**LGreen@wustl.edu JMyerson@wustl.edu**

**Abstract:** Ainslie's insightful treatment of dynamically inconsistent choice stands in contrast to traditional views in psychology, economics, and philosophy. We comment on the form of the discounting function and on new findings regarding choice between delayed rewards. Finally, we argue that the positive correlation between temporal and probability discounting is inconsistent with the view that impulsivity represents a unitary trait.

In *Breakdown of Will*, Ainslie (2001) proposes a radical departure from the traditional approach to the concept of will. He argues against the notion of will as a faculty that exerts top-down control (an executive function, as it were). A major problem with this traditional, top-down approach, as Ainslie points out, is that it cannot easily account for dynamically inconsistent choice. For example, an individual may agree to give a talk at an upcoming meeting, yet regret having made that commitment as the date gets closer. From the traditional perspective, it seems odd that the same person who is endowed with a sufficient degree of will at an earlier point in time would not possess it at a later point. Even odder from the traditional perspective is that the same person often will agree to give yet another talk if asked well in advance of a subsequent meeting.

Rather than blaming such apparently inconsistent behavior on inadequate willpower or selective stupidity, Ainslie suggests that such dynamic inconsistencies, as well as the behavioral phenomena traditionally explained in terms of will, can be better accounted for by conceptualizing will as a strategy that involves thinking of individual choices as precedents. Thinking of choice in terms of precedents that establish choice policies has the effect of "bundling" the sequence of rewards that would be obtained by following a specific policy. Giving a talk, for example, may be thought of as part of a larger, ongoing plan for career advancement associated with larger rewards than might result from a single talk (see also Rachlin 1995; 2000). According to Ainslie, the utility of this bundling strategy derives from the hyperbolic nature of the discounting of future consequences.

Ainslie was among the first to argue that hyperbolic discounting predicts that preferences may reverse with the passage of time, whereas the standard economic model, which assumes exponential discounting, predicts stable preferences. That is, although a larger, later reward may be chosen over a smaller, sooner reward when both are sufficiently delayed, as time passes and the delay until both rewards is reduced equally, preference may reverse and the smaller, sooner reward becomes the more attractive alternative. In the absence of some form of overt commitment, well-intended resolutions frequently are broken as preference reverses. Such inconsistency emerges naturally from hyperbolic discounting and, according to Ainslie and others, provides compelling evidence against exponential discounting.

Preference reversals, however, do not, in and of themselves, rule out exponential discounting. If larger delayed rewards are discounted less steeply than smaller delayed rewards, then exponential and hyperbolic discounting both lead to preference reversals (see Fig. 1 in Green & Myerson 1993). This is important because a number of studies have shown that discounting rate is, in fact, inversely related to amount of reward (Chapman & Elstein 1995; Green et al. 1997; Kirby 1997; Raineri & Rachlin 1993), and thus an alternative method for distinguishing between exponential and hyperbolic discounting models is needed. Curve fitting provides such a method, and when exponential and hyperbolic functions are fit to individual discounting data, the hyperbolic consistently accounts for a larger proportion of the variance (Kirby & Marakovic 1995; Myerson & Green 1995; Simpson & Vuchinich 2000).

As it turns out, curve fitting reveals that a hyperbola-like discounting function in which the denominator is raised to a power provides an even better fit to individual discounting data than either an exponential or a simple hyperbola (Myerson & Green 1995; Simpson & Vuchinich 2000). More specifically, the power to which the denominator is raised is often significantly less than 1.0, and never significantly greater than 1.0. Moreover, often when the data from an individual cannot be described by a simple hyperbola, it can be described by a hyperbola-like discounting function. It is important to note that the hyperbola-like function preserves the essential characteristic required by Ainslie's framework – steeper discounting at short delays and less steep discounting at longer delays, rather than the constant (stationary) discounting rate predicted by the exponential. Thus, although we disagree with Ainslie's argument against exponential discounting based on preference reversals, nevertheless we agree with Ainslie's bottom-line conclusion that the discounting function is hyperbolic (or at least hyperbola-like).

Studies in which curve fitting has been used to evaluate the form of the discounting function almost always have examined choice between an immediate and a delayed reward. An important issue arises when choice involves two delayed rewards. In such situations, people do not simply compare the present (hyperbolically discounted) values of the two rewards, as most behavioral-economic analyses of preference reversals, including Ainslie's, would imply. Neither does behavior conform to the predictions of an *elimination-by-aspects* decision process (Tversky 1972). For example, in a situation where one has to choose between a smaller reward available in one year and another, larger reward available in two years, both alternatives share the common aspect of a one-year wait, yet people do not ignore the common one-year wait. Instead, recent work in our laboratory (Green et al., in press) suggests that discounting functions are still hyperbola-like, but people give less-than-full weight to the common aspect of the delays. By not taking the actual delays fully into account, individuals increase the likelihood that they will choose the sooner, smaller reward, thereby exacerbating their problems with self-control and furthering the "breakdown of will."

Which brings us to the question of what is the best way to talk about the important issues under consideration in Ainslie's insightful book. It should be clear to his readers that when Ainslie refers to *self-control*, he means control *of* the self (in the sense of control of one's own thoughts and behavior), rather than control *by* the self, and we strongly endorse his use of the term. A related issue concerns the use of the term *impulsivity*, a term that does not occur in the index to *Breakdown of Will* but which is increasingly used by researchers to refer to the tendency to choose smaller, sooner over larger, later rewards. Ainslie's arguments against a unitary faculty of willpower also apply to the notion of a unitary trait of impulsivity, which appears to mean simply a lack of willpower. According to the *Diagnostic and Statistical Manual of*

*Mental Disorders* (*DSM-IV*, American Psychiatric Association 1994), for example, "the essential feature of Impulse-Control Disorders is the failure to resist an impulse, drive, or temptation" (p. 609) when the consequences are harmful. Thus, impulsivity and lack of impulse control (an inhibition deficit, as it were) represent the modern version of a lack of willpower.

To Ainslie's arguments against such constructs we would add the fact that the degree to which individuals discount delayed rewards is positively (but weakly) correlated with the degree to which they discount probabilistic rewards (Myerson et al. 2003). A unitary faculty of willpower and an impulsivity trait both imply that a negative correlation between temporal and probability discounting should be observed. For instance, the term "impulsivity" often is used to refer to both an unwillingness to wait for delayed rewards and a willingness to gamble in the hope of getting larger rewards. If the tendencies to discount steeply the value of delayed rewards and a failure to discount the value of probabilistic rewards were both reflections of a single, impulsive trait, then a negative correlation would exist between temporal and probability discounting. Our data are clearly inconsistent with this view.

In *Breakdown of Will*, Ainslie builds on his early insights into the implications of hyperbolic discounting of delayed rewards. Using these insights as a foundation, he constructs a theoretical framework that is able to account for violations of preference consistency, violations that often are mistakenly viewed as representing failures of willpower or as the result of misjudgments about oneself and the world. These mistaken views are represented in many disciplines – economics, philosophy, and experimental, social, and clinical psychology – and it is to be hoped that Ainslie's book will provide a needed corrective. In addition, the book may be of help to individuals seeking better control of their own behavior (not to mention providing fodder for the self-help industry). In closing, we point out that we share with Ainslie the belief that a better understanding of the fundamental properties of discounting will shed light on the mechanisms that underlie both successes and failures in controlling one's own behavior. Our comments on specific aspects of these fundamental properties (e.g., on the form of the discounting function and the relationship between temporal and probability discounting) are offered in the hope of furthering such understanding.

ACKNOWLEDGMENT

Support for the preparation of this commentary was provided by grant MH55308 from the National Institute of Mental Health.

## Comparing apples to oranges: Who does the framing?

Richard Griffin and Daniel Dennett

*Center for Cognitive Studies, Tufts University, Medford, MA 02155.*
**Richard.Griffin@ tufts.edu Daniel.Dennett@tufts.edu**
**http://www.ase.tufts.edu/cogstud/**

**Abstract:** The idea of "bundling" lesser later rewards so they outweigh smaller sooner rewards is compelling, but the sophisticated cognitive activity involved in this bundling is not yet modeled; in particular the role of language is hard to assess.

Contrary to the old adage, you *can* compare apples to oranges, but if you want the results to be worthwhile, you have to do some clever framing. Just "bundling" them together and weighing them on some scale is not apt to give you a meaningful (or effective) result. Ainslie's mathematical model is enticing, but he hasn't yet told us how the backstage carpenters do all the framing.

Ainslie (2001) wants to show how it is that we sometimes come to maximize larger, later (LL) rewards in the face of the default hyperbolic discounting so readily seen in nonhuman animals. He suggests that choosing "according to a principle" or "bundling" can approximate exponential discounting, overwhelming the smaller, sooner (SS) rewards with the summed force of an *imagined* or at least *anticipated* series of LL rewards. He also suggests that this ability requires a self-awareness that is probably unavailable to nonhumans. It probably requires language. In Ainslie's model, however, the bundles of LL rewards are simply summed and if the value is high enough, this will trump the SS impulse. This is mathematically neat, but is it psychologically accurate? Does each bit of the bundle have the same value? One voice, one vote? This doesn't seem likely. A pianist who likes to have several glasses of wine at night might refrain from doing so because of a simple albeit value-laden reason: he has an important performance tonight. What will he do in the future? It may depend on the importance of the gig. A couple of glasses of Pinot Noir may be just the ticket for a chamber rehearsal. The mechanism may be a summing function, but "I" get to prepare the bundles – indeed *have to* prepare the bundles – before they are put on the scales. (What do you get when you cross B. F. Skinner with J. P. Sartre? George Ainslie!)

Ainslie follows a tradition dating back to Plato when he proposes disassembling this "I" into a group of lesser agents or agencies, and composing the competence (and incompetence) of the whole person out of the interactions of these subpersonal homunculi, which he calls "interests." There is nothing wrong with this tactic so long as none of the homunculi are *too smart*. Here is where Ainslie tempts fate somewhat. His *interests* are not like the myopic and obedient clerks that populate most cognitive models or even the eager-beaver competitive homunculi of pandemonium models. Interests are more like different personalities than subpersonal modules or cognitive organs. What is particularly striking about them is that the rules or principles they bargain with are explicitly expressed, not implicit somehow in the hardware. We do not yet know how to devise mechanistic models that can discharge such worldly sub-agencies, with their long time horizons, world knowledge, and access to language.

Language is a tool for thinking, and thinking is a tool for doing a lot of things, or *not* doing a lot of things. Children are often told to "think before they act" and we often remind ourselves to do the same. Indeed, research with children (and chimpanzees) shows that learning apparently simple associations can be quite difficult when prepotent responses are involved (e.g., pointing to a larger food reward but receiving the smaller). Although perseverative errors abound in these cases, they can be alleviated when symbols are introduced as an intermediary. For instance, in tasks indexing young children's ability to deceive, 3-year-olds have great difficulty actually pointing somebody in the wrong direction (it's that-a-way), though their performance increases dramatically if they use an arrow symbol rather than their finger (Carlson et al. 1998). There are many similar findings, and the dictum among developmentalists might as well be "reduce/enhance the salience, reveal the competence." Language is just such an intermediary, and can both reduce and enhance the salience of various phenomena. But the conceptual/emotional structures of which language is part are neither born of a cookie cutter nor are they immutable to change over time, whether isolated or bundled.

What Ainslie provides, then, is a proposal about the architecture of a model of choosing, rather than a model whose mode of operation is already even sketched. This does not diminish the importance of what he proposes. Inverting the modeler's imagination, getting us all to see that our self-control (and its many problems, so revealingly analyzed by Ainslie) must somehow be constructed out of the hyperbolically discounted urges that we are born with, is a contribution that opens up new vistas. It is hyperbolic except when it is not, and that's when things get interesting.

## Is the evidence for hyperbolic discounting in humans just an experimental artefact?

Glenn W. Harrison[a] and Morten Igel Lau[b]

[a]*Department of Economics, College of Business Administration, University of Central Florida, Orlando, Florida 32816-1400;* [b]*Centre for Economic and Business Research, Copenhagen Business School, DK-2000 Frederiksberg, Denmark.* **Glenn.Harrison@bus.ucf.edu** **mol@cebr.dk**
**http://www.bus.ucf.edu/gharrison** **http://www.cebr.dk/mol**

**Abstract:** We question the behavioral premise underlying Ainslie's claims about hyperbolic discounting theory. The alleged evidence for humans can be easily explained as an artefact of experimental procedures that do not control for the credibility of payment over different time horizons. In appropriately controlled and financially motivated settings, human behavior is consistent with conventional exponential preferences.

Ainslie's (2001) book, *Breakdown of Will*, is based on hyperbolic discounting theory. This theory predicts that the individual could behave in a dynamically inconsistent manner, by holding and acting on preferences at one point in time that contradict the preferences of the same individual at a later date. However, before worrying about ways that the individual could address possible dynamic inconsistencies, we need to be sure that the behavioral premise is valid.

A critical design feature in the empirical literature on hyperbolic discounting is the use of a time delay to the *early* payment option in order to control for any confounding effects from fixed premia due to transactions costs. The use of this front end delay (FED) means that one cannot differentiate between "quasi-hyperbolic preferences" and "exponential preferences," and we do not believe that any credible design can do so.

Ainslie concludes his discussion of the empirical evidence on hyperbolic discounting with the following passage:

> There is extensive evidence that both people and lower animals spontaneously value future vents in inverse proportion to their expected delays. The resulting hyperbolic discount curve is seen over all time ranges, from seconds to decades. Because a hyperbolic curve is more bowed than the exponential curve that most utility theories go by, it describes a preference pattern that these theories would call irrational: It predicts temporary preferences for the poorer but earlier of the two alternative goals during the time right before the poorer alternative becomes available. (Ainslie 2001, p. 47)

This passage confounds three things. The first is whether the discount rate varies with the length of the time horizon over which it is being elicited, such as it does with continuously hyperbolic preferences. The second is whether the discount rate for a given horizon and elicited with a FED is different than the discount rate for the same horizon and elicited with no FED. For an experimenter, and for subjects evaluating the credibility of being paid, these are very different questions. The potential importance of this distinction seems to have been first noticed by Benzion et al. (1989).[1] It was also highlighted by Roberts (1991, p. 344), in the context of comments on Ainslie and Haendel (1983) and Winston and Woodbury (1991). The third issue is whether nonexponential preferences imply dynamic inconsistency when one relaxes the restrictive assumption of temporally separable preferences (Machina 1989; McClennan 1990).

The FED design was introduced into discount rate experiments to address concerns about differential credibility. Although it may not completely solve the potential credibility problem, it arguably mitigates it. The FED also serves to equalize any other unspecified differences subjects may perceive between the two payment options. For example, if subjects have a "passion for the present," they demand a premium in order to accept a delay of any length. In a choice between immediate payment and delayed payment, this premium is attached only to the delayed payment. Thus, the subject is being asked to compare "good apples today" with "bad apples tomorrow," confounding the discount rate with the credibility of receiving the commodity. However, if both payments are delayed, the premium applies to both choices and thus becomes irrelevant to a choice between them. Harrison et al. (2002) used a FED in a major field experiment in Denmark, and found that elicited discount rates are proximately invariant with respect to horizon.

There are, however, many field settings in which the relevant issue is what the discount rate is for "money today" versus "money in the future."[2] Even if the experimenter faces the inferential problem of having to then tease apart the effects of time horizon from credibility, transactions, or other subjective costs, it is entirely appropriate that experiments with no FED be considered. If there is a finding that discount rates are not constant when there is no FED, then it is a matter for interpretation as to whether this is a subjective differential cost effect or a time-inconsistency effect (or both).

Evidence for the behavioral importance of a 30-day FED was provided by Coller and Williams (1999). In one of their experimental treatments they had no such delay, and the results from those experiments can be directly compared to their other experiments. After some minor modifications to their statistical analysis, Coller and Williams's results provide evidence that the use of a FED decreases elicited rates by a large amount. The average effect of having no FED is to increase elicited rates by 28 percentage points, with a 95% confidence interval between 52 percentage points and 3 percentage points. Coller et al. (2003) provide additional laboratory evidence on the role of the FED, and show that a 7-day FED is sufficient to overcome the effects of subjective transactions costs.

Finally, there have been no direct tests of the implication of dynamically inconsistent choice behavior using real rewards. Such longitudinal tests require that one allow for possible changes in the states of nature that the subject faces, since they may confound any in-sample comparisons of discount rate functions at different points in time. Harrison et al. (2005) have reported the results of a large-scale panel experiment undertaken in the field to examine this issue and found evidence strikingly consistent with dynamic consistency.

ACKNOWLEDGMENT

We thank the Danish Social Science Research Council for research support under project No. 24-02-0124.

NOTES

**1.** Holcomb and Nelson (1992) reexamined the role of a FED with monetary payoffs, motivated by a concern that Benzion et al. (1989) only studied hypothetical choices. Their FED was only one day long, so it is not obvious that the subjects viewed this as substantially different from there being no FED. They observed no apparent effect of the one-day FED on behavior.

**2.** Such settings might include individual decisions of whether to consume now or save for future consumption, or to purchase a more expensive but energy efficient appliance. We believe that individual decisions involving more significant sums of money or public policy decisions are better characterized as having a FED.

## Shaping your past selves

Jeanne Peijnenburg

*Faculty of Philosophy, University of Groningen, 9712 GL Groningen, The Netherlands.* **Jeanne.Peijnenburg@rug.nl**

**Abstract:** I propose to complement Ainslie's idea of "bargaining with your future selves" with that of "shaping your past selves." The result of such a complementation is that an action can work in two ways: (1) as a predecent for future behavior and (2) as a shaper of past behavior. I argue that this diminishes the unwanted effects of hyperbolic discounting even further.

Weakness of will, or akrasia, comes in two different forms. Broad, apparent, or diachronic akrasia covers cases where an agent fails

to stand by a previous decision about what he will do; strict, clear-eyed, or synchronic akrasia comprises actions that go against an overall judgement that the agent still considers the best at the time of the action. Broad akrasia constitutes the easy problem: How does one explain that an agent changes his mind? Strict akrasia presents the hard problem: How does one understand that an agent believes at time $t_1$ that action $A$ is the best, all things considered, and yet performs not-$A$ at $t_1$? Some have given short shrift to the hard problem by declaring that strict akrasia is an illusion (Socrates, notably); others have tried hard to solve it (e.g., Donald Davidson, whose seminal work [Davidson 1969] spawned dozens of papers and books on the subject).

Be that as it may, Ainslie's (2001) book deals with the easy problem (where "easy" should of course be read tongue-in-cheek). For Ainslie locates akrasia in reversals of preference that occur whenever an agent comes close to a tempting, lesser reward. Ainslie's explanation of this phenomenon is very original in that it is based on the idea that broad akrasia is the rule, whereas its opposite – enkrateia or strength of will – is the exception. Hence, the problem is not why people do not stick to their guns, but rather, why they often do. Ainslie's solution to *that* problem lies in the view of an agent, $P$, as being a collection of agents $P_1$, $P_2$, and so on, at different times, $t_1$, $t_2$, and so on. These $P_1$, $P_2$, . . . have different and often competing interests, but there are also interests that they all have in common. By cleverly bargaining together in the intrapersonal version of a repeated Prisoner's Dilemma, they might succeed in letting the common interests prevail, thereby accounting for $P$'s strong will.

Ainslie's explanation of enkrateia is ingenious, and if couched in less technical jargon, it might well become a useful therapeutic instrument. However, it has a questionable implication as well. For it presupposes that an earlier $P_j$ must be interested in a later $P_k$, and if $P$ is a pitiable alcoholic, this presupposition is doubtful. The hallmark of an addict suffering from weakness of will is that he does not care how he will feel tomorrow or next year. (Ainslie denies this on pp. 17–18, where he argues that a "rational addict" *does* care about the future, because she "wouldn't even try to kick her habit." What Ainslie means, of course, is that she would not even try to kick her habit *now*. But this illustrates, contrary to what Ainslie suggests, precisely her care*less*ness about the future. In the usual sense of "caring about the future" the agent is able to see and reason further than the present moment – something that an addict qua addict is unable to do.)

If an akrates were really to care about the future (in the usual sense), the first step towards his recovery would have been taken. Ainslie's happy thought is to model this first step as a decision that functions as a precedent for future decisions, and hence keeps pace with a "personal rule" that, if followed, will generate a greater reward in the end. Nonetheless, each $P_i$ can still fall prey to hyperbolic discounting by choosing the earlier, sooner reward over the larger, later one; and if he does, he can always logically claim that this was a special case and not a violation of the general rule. However, I think we can make this problem less pressing.

Imagine that I am a happily married mother of five. One day I go to a party, where I dance and drink exuberantly, only to wake up the following morning in a hotel bed next to an attractive man whom I cannot recall having seen before. Although it seems all too clear what happened, I still have some latitude in determining what I have already done. In particular, I can make it the case, through my future actions, that this adventure becomes either a mere incident or the beginning of a long and secret affair.

This example shows that sometimes I can, to a certain extent, determine my past actions. Moreover, my knowledge of the fact that I have this possibility, and hence my understanding that I am at a bifurcation point, might motivate me to pursue the one rather than the other course. Thus, we have here another way of evading the effect of hyperbolic discounting. For if choosing the larger, later reward (continue a happy family life) simultaneously means determining a past action (make my adventure a mere incident), then the smaller, sooner reward (date the attractive stranger again) loses much of its temptation. The reason for this is, of course, that the shaping in retrospect of a past action is already very rewarding in itself. Similarly, when an alcoholic realizes that, through his future actions, he can make a recent lapse become an exception rather than a precedent for his future behavior, he might feel relieved. Very likely this knowledge will diminish his feelings of fatalism and hopelessness, and make him more motivated to contemplate bargaining with his future selves in order to obtain the larger, later reward.

I therefore propose that Ainslie's idea of "bargaining with your future selves" should be complemented with the idea of "shaping your past selves." The result of such a complementation is that an action can work in two ways at the same time, that is, as a precedent for future behavior and as a shaper of past behavior. This means, to use Ainslie's terms, that the behavior in question is not pushed, but pulled (pp. 19, 69). However, it is now pulled more strongly, for two forces are operating simultaneously. In Ainslie's metaphor, a future reward is pulling my present behavior into the future. To this I have added the metaphor of a current reward that is pulling my past behavior into the present. The resultant force is greater than either of its components, and it may well recruit strengthened motivation (cf. Peijnenburg 2004; forthcoming).

## Problems with internalization

Howard Rachlin

*Psychology Department, State University of New York at Stony Brook, Stony Brook, NY 11794-2500.* **howard.rachlin@sunysb.edu**

**Abstract:** Ainslie's *Breakdown of Will* contains important insights into real world self-control problems, but it loses testability to the extent that it internalizes concepts whose meaning lies in overt behavior and its consequences.

Most psychologists who think about self-control tend to stop when they have postulated two forces: a primary impulsive tendency to consume an immediate reinforcer, and a more far-seeing tendency ("the will," in Ainslie's terms) to resist such consumption when it interferes with long-term goals. *Breakdown of Will* (Ainslie 2001) shows conclusively that such a two-force conceptual scheme is totally insufficient to describe almost any real-life motivational dilemma. In our society of plenty, the far-seeing tendency itself needs to be controlled. Otherwise, as Ainslie clearly points out, we will be just as badly off as we would have been if we simply gave in to all our impulses in the first place; indeed, we might be worse off. This fascinating book contains a rich analysis of human motivation and many deep and insightful descriptions of motivational dilemmas.

Having said this, it might sound churlish to complain. Yet I do. Although Ainslie takes care to relate the phenomena he discusses to hyperbolic discounting – a fundamentally behavioral conception – he tends to treat hyperbolic discounting itself as an internal, nonbehavioral (or at least non-overtly behavioral) process. Consequently, some of the discussion takes the form of a literary essay (albeit finely wrought), rather than a scientific analysis. (See particularly the discussion of indirection, pp. 187–96.)

At the root of this problem is Ainslie's attitude towards mental life in general; it is not behavioral enough. (I daresay most of the other commentaries will complain that it is too behavioral.) There is a paucity of empirical research described or cited and few suggestions about how such research could be conducted, especially in the later chapters. Instead, an internal arena is imagined with behaviors, discriminative stimuli, and rewards – all concepts originally constructed to describe the interaction of the behavior of whole organisms with their environments – interacting and competing over time. This internalization of fundamentally external concepts forces Ainslie to resort to internal "thought experiments" like Newcomb's problem (p. 134), rather than real experiments, as evidence for his theory.

Before discussing Newcomb's problem, let us consider a more fundamental concept – the interaction of hyperbolic discount functions. Ainslie writes as if they were internal forces – as if each person has a set of them that he consults (consciously or unconsciously) whenever he has to decide how to behave. But hyperbolic discount functions are most usefully conceived not as internal forces prior to behavior, efficiently causing behavior (that may or may not be inhibited), but as descriptions or summaries of actual overt choices. If anyone has hyperbolic discount functions it is the scientific observer, not the actor. The same goes for a person's intentions. Intentions are not singular internal events occurring just prior to overt actions and causing those actions, but actual patterns of behavior – behavior of the whole organism (to use Skinner's phrase) over time.

Newcomb's problem is interesting only if you believe that intentions are something hidden deep inside a person, normally accessible only by introspection. Newcomb's "powerful being" is a mind reader who can divine those intentions and thus reward or punish the person for intending one thing and doing another. But, if our intentions are patterns of behavior extending into our pasts as well as futures, anyone who knows us sufficiently well (our friends, relatives, perhaps even psychoanalysts) could discover them as well as we could ourselves. These are our true mind readers – better than Newcomb's "powerful being" (unless she can use her power to become invisible and has the time to follow us around wherever we go).

Imagine, nevertheless, that I actually had internal intentions and that my vacation starts in two weeks. Today I intend to go to the beach, but a week from now I change my mind and intend to go to the mountains. Then, on the day before my vacation, I change my mind again and intend to go to the beach. However, the next day, I actually go to the mountains – as I have done on 60% of my previous vacations. Did I really do the opposite of what I (internally) intended at the time or did I just have a weak intention to go to the mountains, as instantiated in my past behavior, and act consistently with that intention? Alternatively, suppose, despite my past tendency to go to the mountains, I went to the beach this time and on every vacation thereafter for the next ten years. You might look back then and say that I really did intend to go to the beach this time. Or, you might not. What my intentions actually are is a matter of how they may best be used to predict my behavior. Identifying them with the operation of some cognitive mechanism in my brain or what I may or may not say to myself or to other people will detract from such use. What I say to other people about my past, present, and future behavior is evidence of my intentions, but it is not evidence of an internal state; it is evidence of my past, present, and future behavior (including verbal behavior). Newcomb's "powerful being" could no more discover my intentions by looking inside my head than she could discover the path of a leaf as it falls by looking inside the leaf. Newcomb's problem is a conceptual as well as a physical impossibility.

Personal rules are a kind of intention in the sense described above. They are not internal forces bargaining with impulses in an internal arena, striking a deal and only then causing overt behavior; personal rules are descriptions of patterns in a person's behavior over time accessible to observers (as well as the behaving person). Rules with occasional violations (which Ainslie discusses so insightfully) are just more abstract rules, hence more abstract patterns. It is true, as Ainslie writes (p. 81), that a theory that says rules are behavioral patterns ought to specify how those patterns are formed and maintained. I have claimed (Rachlin 2000) that, like simpler behavioral patterns, highly abstract patterns may be shaped by extrinsic reinforcement (by parents and society at large as well as by the nonhuman environment). The habit of developing abstract and temporally extended behavioral patterns may itself be shaped. The external reinforcers form a kind of scaffolding for a structure like an arch that will stand by itself when finished. Such patterns may then be maintained, because once fixed, they prove to be of intrinsically of high value and are costly to disrupt. A girl may learn to play an instrument by external reinforcement but keep playing it by the intrinsic value of its pattern. Admittedly, this is woefully insufficient as a theory. We need empirical tests and evidence. But such a conception of personal rules is testable. Thought experiments are fun and often illuminating (unlike Newcomb's silly problem), but they cannot be used as evidence for a theory of self-control, however brilliantly conceived that theory may be.

ACKNOWLEDGMENT

This commentary was prepared with the assistance of a grant from NIH.

## Behavioral (pico)economics and the brain sciences

Don Ross[a,b] and David Spurrett[c]

[a]*Departments of Philosophy and Economics, University of Alabama at Birmingham, AL 35294-1260;* [b]*School of Economics, University of Cape Town, Rondebosch 7701, South Africa;* [c]*School of Philosophy and Ethics, University of KwaZulu-Natal, Durban 4041, South Africa.*

**dross@commerce.uct.ac.za spurrett@ukzn.ac.za**

**http://www.uab.edu/philosophy/ross.html**

**http://www.nu.ac.za/undphil/spurrett/**

**Abstract:** Supporters of Ainslie's model face questions about its integration with neuroscience. Although processes of value estimation may well turn out to be locally implemented, methodological reasons suggest this is less likely in the case of subpersonal "interests."

Ainslie's (2001) model of the function and pathologies of the will exploits two main ideas. The first, hyperbolic discounting of delayed rewards, predicts the intertemporal inconsistency which is the problem facing the will. The second, bargaining among subpersonal interests, is the motivational competition arising, given intertemporal inconsistency, that in turn *is* the will.

We think Ainslie's model is an elegant, powerful, and exciting contribution to behavioral economics. What about the brain? Anyone concerned with how the sciences hang together will want to know more about the relations between the sort of behavioral science Ainslie's *Breakdown of Will* exemplifies, and what the brain sciences tell us about neural processes of reward estimation. The recent and rapid rise of neuroeconomics makes questions about these relationships especially pressing.

Consider discounting first. Besides discounting for delays, agents have reason to discount for decreased likelihood, for inflation, and perhaps for other separations between themselves and rewards that are not simply temporal. Behavioral data is itself equivocal on the extent to which there is a single discounting system or several. Ostaszewski et al. (1998), for example, found dissociation between discounting for delay and for inflation.

In any event, what appears basic from a behavioral perspective may not be implemented in a simple or unified way in the brain. Observed hyperbolic temporal discounting need neither be produced by a dedicated "discounting module" nor involve the activity of any neural subsystem that itself computes a hyperbolic function. Recent research in neuroscience suggests a range of possibilities, of which we draw attention to three. Montague and Berns (2002) show that steeper than exponential discounting can arise from the combination of exponential discounting of the value of a future reward with growing uncertainty that the reward will arrive the further into the future it is expected. McClure et al. (2004) propose that the existence of distinctive neural systems for appraising imminent and delayed rewards may explain an overall pattern of steeper than exponential discounting. Finally, Tanaka et al. (2004) argue both for separate neural systems differentially recruited for appraising immediate and delayed rewards, and for the existence of structured neural maps of multiple time scales in brain regions involved in reward prediction.

These results are neither straightforwardly complementary nor are they in definite conflict, partly because comparison of their re-

sults and methods is impeded by differences between the specific task designs in each case. Ongoing research will extend and refine our understanding of the implementation of value estimation for rewards at different times, including how the implementation is fractionated into multiple neural subsystems. This fractionation, and the consequent possible dissociations that might, in turn, be experimentally investigated, can be expected to help explain various behavioral pathologies.

These issues draw attention to a general methodological question raised by the ongoing, multiplex integration of economics with cognitive science, through neuroeconomics and through Ainslie's "picoeconomics" simultaneously. Economic theory has traditionally not presumed or imposed any particular ontological assumptions about the behavioral units over which it generalizes (Davis 2003; Ross 2005). These are, in different economic models, variously whole biological people, time slices of people, firms, nations, animals, and even species (Vermaij 2004). The basis for this ontological permissiveness – which is the source of economic theory's generalizing power – lies in the fact that an economic agent is just a utility or fitness-maximization function. Which of these there are in the world isn't typically something we can simply observe. Instead, agents are inferred in the construction of economic models. We observe a class of states in some system that seem relatively resistant to perturbation, and infer that it represents an equilibrium in some dynamics or comparative statics of that system. Then we posit agents whose strategic or competitive interaction would yield the equilibrium in question.

This method is fairly clearly the one by which Ainslie supposes that we should infer subpersonal interests in picoeconomic models. The alternative possibility would be to independently identify brain regions (groups of neurons) that appear to track value and contribute to behavioral syntheses in relatively modular ways. However, various remarks of Ainslie's seem to set him against this approach. For example:

> [I]nterests are limited in their duration of dominance, but not necessarily in their access to any of the functions that compose the "self" in any of its definitions . . . [L]ike parties trying to rule a country, internal interests gain access to most of a person's resources when they prevail. The person who wants to stay up later at night and the person who wants to rest in the morning are indeed entire personalities, in the sense that they have the whole person's psychic apparatus at their disposal; and yet they are clearly in conflict with one another. When an intelligent person is acting in his long-range interest not to smoke, he may use that intelligence to devise better stratagems to precommit his future behavior; but when he acts in his short-range interest to have a cigarette, he can marshall that same intelligence to evade these devices. (Ainslie 1992, p. 94)

Thus, in the tradition of behaviorism that has dominated economics for many years, Ainslie regards interests as descriptions of molar patterns in behavior. We suggest that this kind of theoretical interpretation is best understood by reference to Dennett's (1987) so-called "intentional stance" (Ross 2005). As Dennett has long argued, subpersonal agents identified in this way should not generally be expected to reduce to units individuated by the "bottom-up" strategies of neuroscience.

We would not favor any attempt at a sweeping philosophical "solution" to the question about how picoeconomic interests will turn out to relate to neuroeconomic units. It seems to us correct to say that neuroeconomics should identify constraints on the informational dynamics in terms of which intrapersonal bargaining games go on. But this doesn't directly address questions about potential mismatches in individuation principles over subpersonal agents. For now, we simply warn against premature attempts to locate picoeconomic interests directly in neuronal structures. This straightforward sort of "bottom-up" approach to theory construction has seldom been wholly successful in the behavioral sciences, and it has historically served economists especially poorly.

## Freud meets Skinner: Hyperbolic curves, elliptical theories, and Ainslie Interests

Federico Sanabria and Peter R. Killeen

*Department of Psychology, Arizona State University, Tempe, AZ 85287-1104.*
**Federico.Sanabria@asu.edu Killeen@asu.edu**
**http://www.asu.edu/clas/psych/research/blab/**
**http://www.asu.edu/clas/psych/people/faculty/pkilleen.html**

**Abstract:** Ainslie advances Freud's and Skinner's theories of homunculi by basing their emergent complexity on the interaction of simple algorithms. The rules of competition and cooperation of these interests are underspecified, but they provide a new way of thinking about the basic elements of conditioning, particularly conditioned stimuli (CSs).

> *A mere inconsistency in conduct from one moment to the next is perhaps no problem, for a single self could dictate different kinds of behavior from time to time. But there appear to be two selves acting simultaneously and in different ways when one self controls another or is aware of the activity of another.*
> –B. F. Skinner (1953, p. 284)

If Freud dropped the ball when tackled by modern critics (see Ainslie 2001, p. 9), Ainslie has recovered it and is making great forward progress. Why has it taken so long to meld the best of Freud and Skinner? By drawing an image of the human psyche as a negotiation, not among Skinner's two selves or Freud's three, but among a host of economic interests imbued with simple operating characteristics, Ainslie accomplishes much more with the metaphor than did the giants on whose shoulders he stands.

In complexity theory, gross outcomes issue from competition among many elements governed by simple algorithms. The complex behavior of ant colonies arises from a network of such algorithms (Hutchinson & Gigerenzer 2005). In Ainslie's system, the elements are *Interests*, and their primary operating characteristics are the central theme of the book – hyperbolic discounting. Just as carriers of AI "genes" mix and morph, so do Interests in Ainslie's mental life, often overriding one another, sometimes riding one another, with relative strengths depending on proximity to rewards. Because such concatenation would usually be unproductive, Ainslie needs an ego (Fodor's [1983] executive module?) to play cop. But no *ex machina* machinery need apply for top ego-cop in this self-organizing theory. Instead, Ainslie leaves it to the Interests to negotiate, replacing the homunculus with a host of agents. This is not so bad as it sounds, if their operating algorithms are few and simple. But this part needs work. Let us hope that the present précis by Ainslie will draw workmates to help replace this ellipsis with models and data.

A recurring theme in the book is the relation between hyperbolic discounting and the "rational" standard, exponential discounting. Without a story for that relation, Ainslie's theory is a "Dutch Book." Accumulating hyperbolically discounted reinforcers may provide the needed approximations to temporally consistent choices. If organisms in their niche always "bundled" reinforcers to solve intertemporal inconsistencies, researchers would rarely speak of hyperbolic discounting, except when extraneous agents "unbundled" discount functions. As Ainslie points out, however, inconsistencies are too common for exponential discounting to be the rule. It appears that the neat ramp up to the rational tower of temporal consistency is a kluge of multiple ramps of discordant slope. But hyperbolic discounting is insufficient. We need a chemistry of intra-individual interests. When are they bundled? When broken down? When parasitized? Without clear theoretical groundings, Ainslie's models (p. 208, note 14) reduce to mathematical prosthetics for exponential discounting. We need to identify adaptive biological and behavioral processes that deliver hyperbolic discounting. Some candidates are the following:

***Reinforcer expiration.*** The hyperbolic shape of temporal discount functions may be a fixed property of behavior that evolved in contexts where preference reversals optimized reinforcement rate (Logue 1988). Such a scenario often involves non-constant

rates of reinforcer expiry. When predators compete, prey that are not attacked quickly may not be attacked at all, and be left either to maintain the prey population, or to feed a competitor. Rational impulsiveness becomes irrational only in modern times, when banks and groceries ensure longevity of some reinforcers. Other reinforcers, such as sexual partners, maintain selection pressure for alacrity.

***Smaller-sooner versus larger-later?*** Temporal inconsistencies may rarely occur naturally, but may be manufactured by the "smaller-sooner versus larger-later" (SSLL) experimental paradigm. SSLL appears to be designed to model the choices between temptations and virtues that modern human-regulated environments offer, yet it fails to capture the relevant features of the environment of nonhuman organisms – an environment similar to that in which human behavior presumably evolved. Foragers, for instance, sequentially exploit depleting food patches, continuously facing the choice between staying and leaving the currently exploited patch. Blue jays nearly maximize food intake when these contingencies are simulated in the laboratory, but not when facing an economically equivalent SSLL preparation (Stephens & Anderson 2001).

Whether shaped through natural selection or provoked through experimentation, hyperbolic discounting may be the expression of more general – and adaptive – behavioral processes. Let us consider two candidate processes: sensory/temporal discriminability, and Pavlovian conditioning.

***Relative discriminability.*** Adding a constant delay to both reinforcers may turn an "impulsive" subject into a "self-controlled" one; it also makes both delays less discriminable, while the discriminability of amount remains unchanged. Human participants in temporal discounting experiments may be subject to an analogous – verbally mediated – phenomenon: Most humans can distinguish goods attainable with $10 versus $100, but not with $910 versus $1,000. The role of the discriminability of reinforcement delays and amounts on intertemporal choice is yet unknown, but the ubiquity of Weber's law suggests its potential significance.

***Approach gradients?*** A key property of signals of reinforcement is that they become both conditioned reinforcers, or CRs, and conditioned stimuli, or CSs, eliciting approach (Hearst & Jenkins 1974). Is this the behavioral substrate of desire, of appetite? If so, then Ainslie's hyperbolic interests, Skinner's CRs, and Pavlov's CSs are the same entity. A theory of one is a theory of all. Ainslie offers the opportunity to embed the hoary study of CRs in the neomodern study of genetic algorithms and artificial life. This would reinvigorate conditioning theory, promising a much richer and more certain application to the human condition.

The irrationality of hyperbolic discounting may be an awkward by-product of the demands of (a) rapid cultural evolution on human behavior, and (b) experimental attempts to generate such distortions in nonhuman behavior. What would life be like if we discounted exponentially? Hyperbolae may remain the best way to wire *homo economicus* in this post-modern era of Internet foraging. Simulations, anyone? The winners in such round-robin Axelrod competitions will probably find themselves equipped with modules that owe more to Skinner than to Freud; but it would not be inappropriate to call the modules "Ainslies."

ACKNOWLEDGMENTS

Preparation of this commentary was supported by NSF IBN 0236821 and NIMH 1R01MH066860 grants.

## On the coexistence of cognitivism and intertemporal bargaining

Keith E. Stanovich

*Department of Human Development and Applied Psychology, University of Toronto, Toronto, Ontario M5S 1V6, Canada.* **kstanovich@oise.utoronto.ca**
**http://tortoise.oise.utoronto.ca/~kstanovich/index.html**

**Abstract:** Although Ainslie rejects cognitivism as providing an explanation of willpower, a type of nonhomuncular cognitivism is hiding in his own proposal. The key mental mechanism of aggregating individual decisions (bundled reframings) involves representation and decoupling operations encompassed within the analytic system of dual-process mental architectures.

Like others (e.g., Dennett 2003), I am immensely impressed with Ainslie's subtle interpretation of the concept of will as intertemporal bargaining (Ainslie 2001). I would like to explore the idea of marrying his ideas of intertemporal bargaining with cognitivism in the form of dual-process theories of reasoning (Evans 2003; Kahneman & Frederick 2002; Sloman 1996; Stanovich 1999; 2004).

In his book, Ainslie is perhaps rightly scornful of traditional cognitivist views of the will that posited little more than a homunculus standing outside the emotions. But more contemporary dual-process accounts share more features with Ainslie's intertemporal bargaining view.

First, dual-process models are closer to multi-entity bargaining views than it may seem, because such models are really about a multitude of processes, not just two. What has been termed, perhaps misleadingly, System 1 (Kahneman & Frederick 2002; Stanovich 1999), is really a whole *set* of autonomous processes – processes given the term autonomous because: (1) their execution is rapid, (2) their execution is mandatory when the triggering stimuli are encountered, (3) they are not dependent on input from high-level control systems, and (4) they can operate in parallel.

In contrast to the autonomous set of systems stands the nonautonomous system(s) – sometimes going under the name of analytic processing – with the opposite set of properties. Analytic cognitive processes are serial, rule-based, often language-based, and computationally expensive. The analytic system carries out two critical processes. First, it is responsible for sustaining the decoupling of representations from the world so that cognitive simulations can be run which test the outcomes of imaginary actions (Currie & Ravenscroft 2002; Dienes & Perner 1999; Nichols & Stich 2003; Perner 1991). Secondly, the analytic system can override responses triggered by the autonomous system.

Hyperbolic discounting provides a wonderful and heretofore missing explanation of where the analytic system gets the motivational force to override the autonomous systems that are responsible for preference reversals in conflicts between short-term and long-term interests. The bundling of rewards extended over time turns hyperbolic curves into exponential ones (at least temporarily, subject, of course, to the many other subtle bargaining effects that Ainslie's book details so ingeniously). However, there is much reference in Ainslie's book to bundling choices together, to choosing on the basis of principle, and to aggregating choices into categories. Now this kind of talk hides a lot of cognitivism – but in this case I think it is the good kind of cognitivism and not the bad kind that Ainslie spurns at the beginning of his book. Bundles, categories, principles – these are representations, often formed using the quasi-linguistic representational formats of the analytic system. The bundles or principles represent an alternative framing of the situation (one that yields exponential discounting rather than hyperbolic discounting). If the bundled frame dominates, then the long-term interest wins out. If the unbundled, single-choice frame dominates, then the short-term interest wins out. This type of representational redescription (see Karmiloff-Smith 1992) is much facilitated by language (and the discrete categories it provides). Using linguistic rules to rapidly implement new goal hierarchies is a quintessential function of the nonautonomous analytic system.

Likewise, I suspect that metarepresentation (another unique

analytic system function not within the capability of autonomous systems) would be implicated if we fully unpacked the mechanisms behind bundling, aggregating, or implementing a rule. At some point before the rule is fully instantiated, the first-order preference for the short-term reward has undergone a critique (the person realizes that they prefer to prefer otherwise).

Finally, Ainslie is clear that the reward bundling process critically implicates a self-prediction process ("a big part of this picture is her expectation of how she'll choose at later times," p. 131). However, these self-predictions are cognitive representations used in simulations in what some investigators call the Possible World Box (PWB; Nichols & Stich 2003), and sustaining the PWB is cognitively effortful (Glenberg 1997; Stanovich 2001; 2004). Ainslie has given us an insightful explanation of the motivational force that sustains the PWB.

I suspect that such dual-process explanations generalize to situations and effects not involving intertemporal discounting (e.g., probability matching, conjunction fallacies, various non-Bayesian responses). I have argued (Stanovich 2004) that many fallacies in the heuristics and biases literature arise from conflicts between autonomous systems more geared to genetic fitness maximization (i.e., to subpersonal entities not maximizing at the level of the individual) and analytic system procedures geared toward personal utility maximization. Ainslie edges toward this view when he posits (p. 45) that hyperbolic discounting might have been fitness maximizing at the expense of the individual organism, and when he allows for bundling to be facilitated by "executive processes that find bright lines" (p. 99).

Ainslie's view is consistent with dual-process theorists (Evans & Over 2004; Stanovich 2004) who emphasize that the bundling of choices into rules is evolutionarily unprecedented ("certainly a novelty in nature . . . a rather artificial process unlikely to have arisen in lower animals"; Ainslie 2001, p. 146), but he parts company with most of us in his view that "we apparently used the same intelligence that created this impulse problem to find a way around it"; p. 146). Autonomous processes create the problem of making us prone to responses that serve ancient interests of genetic fitness but not current personal utility maximization, and they lack a host of features necessary to overcome the problem, including: the ability to decouple from primary representations, the ability to create possible worlds and run simulations in them, and the ability to metarepresent. These processes – necessary to perceive an alternative world that better serves personal long-term interests and to take actions that bring about that world – are fundamentally different from the ballistic and automatically triggered features of autonomous systems.

In short, Ainslie adds the motivational component missing in most dual-process theories; whereas the latter deal extensively with the linguistic bracketing, cognitive decoupling, and simulation operations that are the mechanisms lying underneath the bundling and aggregating of choices. I see great potential for enriching dual-process theory with the insights from Ainslie's very nuanced attempt at a scientifically valid elucidation of the ancient concept of willpower.

## "To do or not to do?" Modeling the control of behavior

John D. Swain[a] and James E. Swain[b]

[a]*Physics Department, High Energy Group, Northeastern University, Boston, MA 02115-5000;* [b]*Child Study Center, Yale University School of Medicine, New Haven, CT 06520.* **john.swain@cern.ch** **james.swain@yale.edu**

**Abstract:** The author of this fascinating book explores the problem of decision-making. As a basis, he uses hyperbolic discounting theory to discuss many basic assumptions related to self-control. In an accessible conversational tone, he succeeds in capturing many current problems in decision science and presents a rational framework for further work.

Rewards in the far future tend to be valued less than rewards in the more immediate future. The way in which future value is decreased seems to constitute a "breakdown of will," in that choices are often made that seem not to favor the overall best outcome. Assignment of reduced value to a delayed reward is expressed in terms of "discounting," and arguments are made that rational discounting should be done according to an exponential function. Such a distribution is well known in essentially all the sciences and follows from expressions of the form "rate of change of $V$ with respect to time is proportional to the value of $V$ itself at that time." It embodies an assumption that there is no change in the form of the solution if time is delayed by some constant.

Ainslie (2001) points out the rather striking fact that people do not discount rewards according to a relationship of this sort.

> If I asked a roomful of people to imagine that they've won a contest and can choose between a certified cheque for $100 that they can cash immediately and a postdated certified check for $200 that they can't cash for three years, more than half of the people usually say they would rather have the $100 now. If I then asked what about $100 in six years versus $200 in nine years virtually everyone picks the $200. But this is the same choice seen at six years' distance. (p. 33)

An exponential discounting function is incompatible with this behavior, so clearly, some other sort of discounting must take place favoring early rewards over later ones more strongly. The beauty of this sort of quantitative measure of reward and time is that the discounting function can actually be computed (or, more properly, estimated) from experiments. The result is that the data are well described, not by an exponential discounting, but by a hyperbolic one, that is, the perceived value at time $t$ from now falls off as $1/t$.

Such a hyperbolic discounting has significant qualitative differences both at early and late times. At very early times, it in fact explodes, requiring a modification to render it finite, so that $1/t$ must be replaced by an expression of the form $1/(t + a)$. It very strongly favors immediate rewards. At very late times, however, it falls off more slowly than an exponential decay, making very long-term rewards more appealing than they would be for an exponential discounting. These twin features, argues Ainslie, can go a long way towards explaining both the fact that we tend to favor immediate gratification over mid-term gratification, and yet we can also plan for events far into the future.

From the point of view of the physical sciences, it is interesting to examine the origins of the distributions in order to see why hyperbolic discounting might be favored. Any discounting formula must come from an expression of the form "rate of change of value is some function of value and time." Exponential discounting, as described earlier, assumes no special time (so the function referred to is independent of time), and it also assumes that there is no special value – the rate of discounting is proportional to the value already there. Hyperbolic discounting assumes, on the other hand, that discounting is proportional to the square of the value, and although it introduces no special time, the solution in fact favors the special time $t = 0$, or "now."

Under conditions of great stability, where there is no special time, and "now" is much the same as "tomorrow" or "next year," and there is no special urgency attached to a large versus small value, one might well expect a simple exponential discounting, and indeed, this is what one sees in finance with respect to interest rates. Constant interest rates reflect stable times, whereas interest rates that change with time represent deviations from strict exponential discounting.

In nature, for the most part, there is little reason to expect this sort of stability. Lifespan is certainly finite, and most people would have little interest in waiting for 1,000 years versus 1,003 for the experiment with the cashiers' cheques. Even more pressing is the evolutionarily critical ability to reproduce, and Ainslie does suggest that a gene-driven urge to somehow reproduce "now," for "tomorrow" we could be dead, would seem to motivate hyperbolic discounting. Ainslie also hints at the role of nonlinearities, and although he only begins the discussion of what could be a very wide

field of inquiry, it is easy to see how nonlinearities could come into play. Suppose that instead of the offers with $100 or $200, one had 100 or 200 seeds, and further suppose that by planting the seeds one could have 100 seeds yield 1,000 in one year. Now delaying acceptance of 100 seeds to get 200 seeds three years later is clearly foolish – one will be well ahead of the game with 100 seeds now! Here, however, one sees a difference with seeds in that they do not have so much a fixed value, as a value that itself can increase over time. Similar effects can arise for coupled resources and in cooperative systems. In a leading approximation, the simplest nonlinear modification of the equation leading to exponential discounting will generically lead to hyperbolic discounting.

How far hyperbolic discounting will ultimately go to resolve deep questions about the human will, remains to be seen, but Ainslie makes it clear that, at the very least, defining rational behavior as that which would correspond to exponential discounting (and thus to assumptions about uniformity of conditions in time and the lack of any nonlinearities) is flawed. More complex discounting algorithms are certainly conceivable, but both from general arguments and a wealth of experimental data, it seems that hyperbolic discounting goes a long way towards capturing the basic spirit of these.

This is a well organized and reasonably priced, accessible book – useful for any behavioral scientists interested in a deeply considered introduction to the topic of decision science.

## Reference point-dependent tradeoffs in intertemporal decision making

X. T. Wang and Jeffrey S. Simons

*Psychology Department, University of South Dakota, Vermillion, SD 57069.*
**xtwang@usd.edu jsimons@usd.edu**
**http://www.usd.edu/~xtwang/ http://www.usd.edu/~jsimons/**

**Abstract:** We agree with Ainslie's general approach to intertemporal choices and self-control. However, we argue that a concept of "will" is superfluous in explaining tradeoffs between SS (smaller and sooner) and LL (larger and later) rewards in a framework of temporal goal setting and goal aggregation. We provide an alternative framework of reference point-dependent tradeoffs between SS and LL options.

Ainslie (2001) brings human choices out of the realm of rational maximization of economic goods and into a psychological world of motivation, temptation, and risk preferences. We agree with his general approach to studying risky choices and self-control in a framework of temporal goal setting and goal aggregation. Ainslie suggests that "will" can be viewed as being the effect of aggregating goals over time to determine choice. This hypothesis affords some interesting reasons as to why people would prefer smaller and sooner options (e.g., cognitive deficits that make the aggregation of goals difficult; a foreshortened sense of the future; experience with unpredictable environments; self-efficacy in achieving long-term goals, etc.).

However, we argue that in a framework of goal settings and goal aggregation, a concept of will as an explanatory construct is superfluous. We propose a new conceptual framework of reference point (goal or a minimum requirement) dependent tradeoffs between SS and LL rewards to account for intertemporal decision-making and self-control.

Ainslie's description of will as aggregated choice is a potentially rich and informative perspective. However, Ainslie's conceptualization of the will is primarily descriptive, yet the construct of will in psychology is mainly promoted as explanatory. He defines "strong" will as the aggregation of future choice points to facilitate choosing longer, later (LL) over shorter, sooner (SS). He suggests that "strong" will manifests when the SS/LL choice is viewed as a class of choices and that choosing SS at one point in time is perceived as promoting SS choices at each successive time point. Ainslie states that individual differences in aggregation rules lead to adaptive or maladaptive consequences of the functioning of the will. What we do not know is how people develop variations of these aggregation rules, how some are able to view the aggregation of choices *against* SS as more reinforcing than the current choice *for* SS. The answer to that question provides explanatory power and yet this seems largely untouched in Ainslie's conceptualization of the will.

Contrary to the assumption indicated in Ainslie's book, that LL rewards are always superior to their SS alternatives, we intend to demonstrate that some ostensibly irrational and impulsive behaviors in favor of SS rewards over LL ones can be both normative and adaptive, given that risky choices are bounded by goals and deadlines in life.

The last three decades have witnessed great theoretical and empirical developments in the studies of reference points in human decision-making regarding risk (e.g., Heath et al. 1999; Kahneman & Tversky 1979; Lopes 1987; Tversky & Kahneman 1981) and in foraging behavior of nonhuman animals (e.g., Kacelnik & Bateson 1997; Stephens & Krebs 1986).

In making intertemporal decisions between SS and LL rewards, the process of approaching a goal (G) can be viewed as a process of status quo (SQ) improvement, whereas the process of falling towards a minimum requirement (MR) can be seen as a process of SQ deterioration. As illustrated in Figure 1, when faced with SS and LL alternatives, the choice becomes a tradeoff between the amount and the delay of rewards with reference to distances to the upper- and lower-bound reference points (i.e., G and MR).

For an upward expected SQ over time, SS should be preferred to LL (SS > LL) if SS can reach a goal earlier. SS (or any choice) should be preferred whenever it will be sufficient for reaching the goal state. The upper middle arrow is LL in respect to the upper left arrow, but SS in respect to the upper right, yet in either case it should be preferred because it moves the person past the goal. Essentially, the crucial determinant is not maximizing value but minimizing the goal discrepancy as quickly as possible (cf. Carver et al. 1996). Outcomes that fall both below or both above a reference point (a goal or a minimum requirement) are expected to be more similar in their psychological values, whereas outcomes that are located on different sides of a reference point are expected to be markedly different in psychological values.

For a downward expected SQ over time, LL options should be

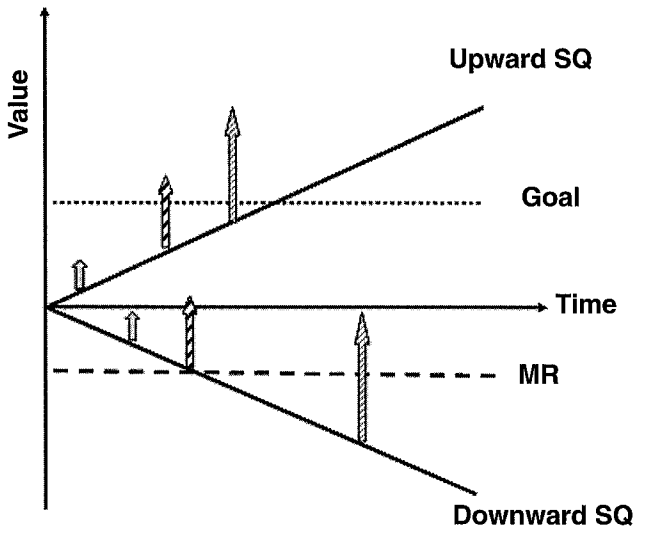


Figure 1 (Wang & Simons). Expected future gains on top of either an upward status quo (SQ) trajectory or a downward status quo trajectory. The length of the arrows represents the amount (value) of rewards (gains) at different time points. MR = minimum requirement.

preferred to the SS alternatives (LL > SS), provided that the delay will not allow the person to fall or remain below the MR. However, SS options should be preferred when they can keep the SQ above the MR or bring the SQ above the MR sooner than LL alternatives. In the case illustrated in the lower part of the figure, the medium gain would be superior to the largest gain (SS > LL) because survival cannot be delayed. A starved man needs any food that can feed him instead of a delayed larger supply. As the SQ approaches the MR in time, the temporal difference is vital, but the amount difference is functionally null.

This analysis suggests that some impulsive behavior such as drug use and unprotected sex may not be simply a result of intoxication or cognitive deficit, but adaptive reactions to perceived goal distance and to subjective estimation of SQ trajectory, which may or may not be accurate. That is, if one is below, or perceived to be falling below a MR, the option that most quickly returns the person to above the MR should be favored. Though the projected outcome of the LL may be superior, the individual is unable to be sustained below the MR to reach the LL choice point in time.

Hyperbolic discounting functions provide a general mathematical expression of psychological mechanisms of intertemporal decision-making. However, the functions themselves are not psychological mechanisms and seem not to be congruent with a framework of reference-dependent decision-making.

# Author's Response

## A bazaar of opinions mostly fit within picoeconomics

George Ainslie

*Veterans Affairs Medical Center, Department 116A, Coatesville, PA 19320 and Temple Medical College, Philadelphia, PA 19140.*

**George.Ainslie@va.gov** **www.Picoeconomics.com**

**Abstract:** The will has generated a wider range of opinions than most phenomena, lacking as it does both an animal model and consistent behavioral correlates. It has even been held not to exist. The commentators approached my intertemporal bargaining (picoeconomic) model from many angles. Doubts about the existence of the underlying phenomenon, hyperbolic discounting, were still raised by some, but other commentators added to the evidence for it, which I regard now as overwhelming. Where mechanisms of self-control were specified, I found it possible to place them within a picoeconomic framework.

### R1. Introduction

My purpose has been to show the feasibility of a bottom-up model of choice-making, one that starts with the simple reward-seeking processes observable in most animals and combines these processes by simple principles so that, once a capacity for foresight and self-perception is added, it predicts the familiar nuanced experience of an autonomous ego. The commentators have made useful suggestions about each stage of this model, and have given me occasion to clarify many aspects of it in ways I had not thought of before. Most of their analyses start with the seminal empirical finding of hyperbolic discounting, which still seems to be controversial. Section R2 concerns the intertemporal conflict implied by this finding, and several commentators' critiques of the idea that internal interests based on conflicting rewards create a self in the form of a marketplace (discussed further in sect. R3). One commentary encouraged my alternative approach to classical conditioning (sect. R4). The greatest number of commentators addressed the capacity of a modified repeated prisoner's dilemma game in this marketplace to generate strength of will (sect. R5), a mechanism that also gives a rationale for the experience of freedom of will (sect. R6). My suggestion that thought experiments can provide valid data on this kind of recursive internal process was addressed in the case of the least straightforward of my examples, Newcomb's problem (sect. R7). Finally, since most of the commentators deal in one way or another with the empirical status of my approach, I review the present status of research on this topic (sect. R8). The main topic of the last third of the target book, hyperbolically driven impatience for premature satiation of appetites and its likely consequences, was, perhaps wisely, left alone.

### R2. Hyperbolic discounting

The most important consequence of hyperbolic discounting is that it may cause the value of a small, sooner (SS) reward to spike above that of a larger, later (LL) one temporarily, when the SS reward is imminently available. **Bridgeman** says that the hyperbolic model will not work because, among other problems, hyperbolic curves go to infinity at zero delay; but the empirically derived curve I propose does not go to infinity. **Arló-Costa** suggests that Rubinstein's (2003) "similarity relations" mechanism contradicts hyperbolic discounting. However, those experiments mostly show that subjects ignore very small differences, which at most suggests a supplementary principle to the robust hyperbolic curves that have been observed in both human and nonhuman experiments. One of Rubinstein's experiments unwittingly replicates Kirby and Guastello (2001): A preference for $997 now over $1,000 a month later, but not for four $997's at monthly intervals over four $1,000's, each a further month later, does not refute hyperbolic discounting, but rather demonstrates the mechanism by which it permits willpower – the predicted increase in preference for LL rewards when choices are bundled into series (*Breakdown of Will*, Ainslie 2001, pp. 82–84).

**Green & Myerson** refer to data, much of it their own, that support a specifically hyperbolic shape. They argue for a credible refinement that improves its already superior fit with choice data by raising the denominator of the value equation to a power, an adjustment first suggested for the matching law in general by Baum (1974). I have not examined this suggestion at any length because, as Green & Myerson point out, it does not change the strategic implications of the basic hyperbolic curve for intertemporal bargaining. However, this added bit of precision clearly supports the basic hyperbolic shape of the discount curve, as does the finding of Green et al. (in press) that the choice between two non-immediate rewards is also evaluated hyperbolically. Ainslie and Haendel reported temporary changes of preference between two non-immediate alternatives as early as 1983, but the parametric work of Green and his colleagues argues much more strongly for hyperbolic discounting in all delay periods (cf. Ainslie & Haendel 1983).

Does this mean that we are sometimes observing people's inborn discounting tendency in the raw? Almost certainly not. People report discounting large amounts of money less steeply than small ones, while rats and pigeons, presumably making something close to raw choices (without acquired cognitive overlay), do not change their observed discount rates by amount of reward, or even discount smaller amounts less than greater amounts (Wogar et al. 1993). Older people discount rewards much less steeply than younger ones (Green et al. 1994); although this effect has not been looked for in pigeons, it is hard to imagine it there. The impatience factor varies by hundreds within single human experiments, but only sevenfold in nonhumans (reviewed in Ainslie & Monterosso 2003, p. 42). The large but variable increases in human patience are best ascribed to learned impulse control, modulating but not obliterating the original discount function. I have argued that the modulation of observed discounting comes from the action of committing maneuvers on undiminished impulses, rather than an ability to bend the function itself (Ainslie 1992, pp. 123–25): To bend the function itself would be to make distant rewards more valuable, in effect to coin value, a skill that we would expect any subject to have learned maximally if it were possible. It would be as if a child could suddenly learn to make her birthday seem only half as far away. Such a change should certainly be highly rewarding, in contrast to the painfulness that usually characterizes impulse control.

The most basic objection to hyperbolic curves is that they ought not to exist (**Bach**), perhaps because evolution ought not to have selected for them. But I have cited numerous controlled experiments showing that something very close to that form does exist, in both humans and nonhumans. It is clear why animals in the wild benefit from immediate obedience to instinctual urges, enforced by a steep devaluation of future possibilities, just as **Sanabria & Killeen** suggest; but that can be achieved by exponential curves as well, without the potential problem of intertemporal conflict. I have suggested that hyperbolic curves might betray the individual into preserving its genes by fighting or reproducing in opposition to its own interests (as **Stanovich** also notes), or alternatively, that these curves were dictated by some basic scheme of sensory organization and are harmless in less foresighted organisms (*Breakdown of Will*, pp. 45–47), but this is pure speculation.

**Swain & Swain** point out that hyperbolic discounting is a simple modification of the exponential kind – making decreases in value with delay proportional not to momentary value, as in exponential discounting, but to the square of momentary value.

$$\frac{dV(t)}{dt} = -k\,V(t) \times V(t) \text{ instead of } \frac{dV(t)}{dt} = -k\,V(t) \text{ the exponential formula}$$

(amplification in John Swain, personal communication). This nonobvious relationship which hyperbolic curves have with exponential ones may or may not turn out to play a role in their evolutionary origin.

**Bach** observes, by way of rebutting hyperbolic theory, that countless curves can produce temporal reversal. Most of them, however, would lead to the same limited warfare relationship among successive motivational states, which is the basis of intertemporal bargaining theory. I have advocated the hyperbolic shape not because it is unique in this regard, but because it is so well supported by data. Bach adds that different modalities of reward might have differently shaped curves, thus suggesting an alternative mechanism for temporary preferences that would have more serious theoretical consequences. If alcohol were discounted more steeply than the absence of hangover, for instance, then even an exponential discounter might prefer heavy drinking temporarily, in the period before it was available. However, virtually all amount-versus-delay experiments have kept within single modalities of reward.

The most widely accepted argument against hyperbolic discounting as a mechanism for impulsiveness cites the frequent experience (and some fMRI data; McClure et al. 2004) that an emotional rush based on immediate availability precedes impulsive consumption. This observation has led to the proposal that temporary preference for the impulse is based on this rush, and that except for this brief period, discounting is exponential (Laibson 1997). **Harrison & Lau** make a similar objection, that the temporary preference phenomenon in human subjects is an artifact of not using a "front end delay" (SS reward at one month, say, instead of immediately). They report, in contradiction to **Green & Myerson**, that when they used such a delay they elicited preferences for various amounts of money at various delays that did not imply hyperbolic discounting or temporary preference in most subjects. However, these experiments offered large amounts of money ($450 to $1,840) and listed the possible choices to subjects with the equivalent annual discount rates next to them (Harrison et al. 2005). Their subjects could well have been choosing among the listed discount rates rather than going by the hard-to-calculate SS versus LL amounts. It has always been clear that people can learn to obey conventional exponential curves in some circumstances, especially in sizeable financial transactions (Ainslie 1991). People who cannot, will be money-pumped: Someone who has learned the knack of exponential evaluation will buy their overcoat, for example, every spring, and sell it back to them every fall. The remarkable finding is that so many experimenters report a hyperbolic pattern, even in financial preferences.

As for the "conditioned" rush that often precedes impulsive choices, it may be a reward-seeking process that attempts to *make* a reward happen immediately (see *Breakdown of Will*, pp. 65–69). It is apt to arise in Darwin-James-Lange fashion when you do not have complete confidence in your resolution not to indulge in this kind of reward (Ainslie, in press a).

## R3. The marketplace model

Several commentators object to my assertion that an ego or self need not exist as an organ, but only as an alliance of processes motivated by long-range reward. **Bach** criticizes my making cognition the servant of reward, "something's being desired because it is rewarding" rather than "something's being rewarding because it is desired." He prefers "magnitudes assigned to particular rewards" by a governing cognitive process. Certainly, that is the way that our culture – and perhaps all cultures – describe the experience of higher-level judgments. The impossibility of predicting decisions from the apparent contingencies of reward has led to the theory that choice comes ultimately from a transcendent ego; but if this ego is not to be exempted from the chain of

causality, it must itself have a basis for making choices. Judgments are called higher-level when they buffer decisions from relatively fast-paying rewards. My argument has been that the tails of hyperbolic discount curves at long delays describe adequate motivation for these judgments in all the complexity that we ascribe to them. I suspect that it is impossible by spontaneous introspection to discriminate intertemporal bargaining in an internal marketplace from the organ of judgment that is commonly imagined – and, furthermore, that analysis of this introspection by means of thought experiments provides evidence for the bargaining model. True, we often "assign" value to a goal, but in doing so we are not just thinking, but shifting motivational weights, and we must apply weight in order to do this. The function of cognition is to estimate where to put the fulcrum.

**Bridgeman** makes the opposite objection, that my game-theoretic mechanism for integrating motives entails too many free parameters to be of any theoretical use. I am not sure that my model is clear to him. I do not propose that the data that goes into previous choices is discarded, for example, but that all elements of your choice history can be part of the self-prediction necessary for intertemporal bargaining. The experimental phenomena that the theory predicts have been relatively elementary – that organisms are sometimes motivated to limit their own future choices, for instance, or that choosing in bundles increases the tendency to choose LL alternatives – but that has to be true of all models of human choice. The complexity of determinants makes certainty impossible in principle, and gives rise to the impression of free will (see sect. R5). Yet some modeling is still possible. The complex theoretical problems of internal motivational conflict seem to be closely analogous to those found in macro- and microeconomics, which have been rendered much more tractable by a game-theoretic approach (Smith 1992).

**Ferrero** correctly points out that what I call interests – processes selected (shaped) by what rewards them – are not "replicators in a process of natural selection," competing by heritability. Natural selection is a mathematical effect governing the survival of organisms over multiple generations; the fitness that is inferred from survival follows an exponential curve (Lotka 1957, p. 123). I am describing selection by an organic process, reward, that is coherent in the respect that it determines the relative dominance of all mutually substitutable processes according to the same hyperbolic attenuation curve that governs perception (Gibbon 1977); under its control, the learned processes that I call interests may wax and wane, but do not die to be replaced by new generations. What are transient are the momentary configurations of motivation that determine which of these processes prevail at each point in time – my "successive selves."

I agree that interpersonal analogies are just illustrations, but since interests can be expected to share some properties with whole people, such analogies may clarify. A legislature contains people who serve particular interests, and who must both *overtly* bargain with others to establish rules and *covertly* plan strategies – to augment their own power, diminish that of competitors, and while in power protect as much of their gain as possible against periods when they will be out of power. My argument is that the shifting preferences created by hyperbolic discount curves create similar incentives for competing interests – which, remember, are just reward-seeking processes, each defined by the reward it seeks. A coherent ego grows – is learned – insofar as it is valuable to enough interests, and comes into play only when coming into play is more rewarding than not coming into play. **Ferrero** seeks to differentiate "really strategic thinking" from "reflection on the import of one's present action in the context of one's continued existence," but I would accept as strategic the latter process, involving, as he says, an appreciation of precedents. Whether you regard an agent as having "identity over time" or comprising successive states in a limited warfare relationship is much like the question of whether you regard an atom as solid or mostly empty. All these views are true in their way, but the former of each pair just extends a subjective impression, whereas the latter specifies a model.

**Griffin & Dennett** warn that the interests I propose may be "too smart," perhaps becoming the bad sort of homunculi, "bogeymen [that] duplicate entire the talents they are rung in to explain" (Dennett 1978, pp. 123–24). However, I mean them to be not necessarily more than conventions of the observer. Interests are ways of grouping the processes that are shaped by one reward as opposed to those shaped by its alternative(s), and then only if the reward is increased by strategically blocking the alternative(s). There is no reason for their generality not to remain fluid. We could speak of a dieting interest or a long-range fitness interest, or even a self-control interest, or, moving the other way on a specificity scale, of a cholesterol-reducing interest or even a keeping-my-appetite-for-today's-dinner interest. The opposing interests must necessarily be short range, a snacking interest, or a taking-the-edge-off-my-appetite interest or an eating-that-particular-ice-cream-cone interest. The concept is useful only when the contingencies of one reward create an incentive to forestall or avoid being forestalled by the other(s) – thus, it would be pointless to distinguish a chocolate interest from a vanilla interest.

I say "not *necessarily* more" than observers' conventions, because the fluidity of the concept does let us recognize that some interests can assume an identity and function as personalities. That is, the observers that group together the processes which seek a particular kind of reward might include this group itself: The interest might include an observing process that advances the group by scrutinizing its properties and distinguishing members from non-members. The extreme example would be the outlawed traits that combine to escape the surveillance of the main personality in the dissociative disorders (Ainslie 1999). However, this should happen only when there is differential reward to form and defend separate identities – perhaps the same reward contingency that usually leads people to form a single sense of self. As **Ross & Spurrett** suggest, interests can be seen as Dennettian intentional systems, which can also divide and combine without a priori limits. The limited warfare relationship motivated by hyperbolic discounting then serves variously as the wedge that divides them and the glue that combines them.

There are undoubtedly properties of bargaining that are implicit in people's hardware, in the same way that our hardware prepares us to communicate with nouns and verbs in whatever the language we wind up learning. As **Griffin & Dennett** point out, we cannot observe or even infer most of these properties yet. However, an effort has begun to find mechanistic models that might resonate with them, along the lines suggested by **Ross & Spurrett**

(Ainslie 2005b). I certainly agree with Ross & Spurrett that we have no basis for speculating about "how picoeconomic interests will turn out to relate to neuroeconomic units," but that "neuroeconomics should identify constraints on the informational dynamics in terms of which intrapersonal bargaining games go on." Nevertheless, neurophysiology has already described phenomena that are at least consistent with the implications of hyperbolic discounting. For example, the activity in brain centers that responds proportionately to rewards is also often found to be proportional to the surprisingness of rewards (Berns et al. 2001). Although this activity may be only informational, it also may be involved in the habituation of predictable rewards, to which I have attributed the impracticality of ad lib self-reward (see *Breakdown of Will*, pp. 164–74). More counterintuitive are Berridge's (1999) findings that both human and nonhuman subjects will work to obtain events that they report or show other evidence of not liking, if and only if the events are imminently available. He suggests that the "incentive salience" which selects for such choice does so through classical conditioning, but it is much more likely to do so by inducing genuine, but temporary, preference for brief, imminent reward (see *Breakdown of Will*, pp. 51–54). As a final example, Iacoboni et al. (1999) have reported neuronal activity in areas governing particular movements that occurs when a person observes someone else making similar movements. This finding suggests a fundamental preparedness to model others' experiences as your own, which would be at the service of the primary incentive for empathy (see *Breakdown of Will*, pp. 179–86).

## R4. Aversion with only one factor

**Sanabria & Killeen** were alone among the commentators in mentioning the opportunity to revise "the hoary study of CRs [conditioned responses]" afforded by hyperbolic discounting. I judge such a revision to be among its most important implications. It has no obvious bearing on the learning of information by simple connectionism, sometimes called stimulus–stimulus conditioning (Rescorla 1988). However, for the selection of responses, the potential of brief temporary preferences to lure organisms into response processes that are aversive overall, could let us do without the *deus ex machina* of a second, "conditioned" selective principle that is so often invoked to explain aversive, involuntary, or maladaptive processes. This is not to deny that classical conditioning exists as a procedure, but to point out that the combination of prepared responses and hyperbolically discounted rewards could explain its observed patterns, and more parsimoniously (*Breakdown of Will*, pp. 51–61; target article précis, sect. 4.1). After all, conditioning and reward-based choice depend on the same selective factors, but heretofore, rewards have been equated just to satisfactions, with only unconditioned stimuli seen as able to select aversive processes. The model of aversion as a rapidly cycling addiction comprising reward and inhibition of reward lets us add conditioned processes to the marketplace of rewarded behaviors.

## R5. Intertemporal bargaining

Because of its subtlety, the process of will has been described from several different angles, which seem at first glance to demand incompatible theories. The true picture is probably more like one of the blind men describing an elephant. I have described it in terms of expected reward, but it is also a process that could be seen as one of assigning reward values (implied in **Bach**), goal-setting (**Wang & Simons**), metarepresentation/representational redescription (**Stanovich**), or the converse of expected regret (**Connolly & Reb**). Bach does not actually address the problem of will, so I will not belabor the question of what factors constrain the assignment of reward values.

**Wang & Simons**' reference point model deals with the choice of instrumental behaviors toward a given goal, not the competition of goals themselves. As such, it presumes some means of goal-setting, and carries on from there. But a goal is not the same thing as an amount of reward. If the only significant reward in a situation comes from the kind of all-or-nothing goal depicted in their figure, which you either "move past" or you do not, then the upper mid-length (striped) arrow is the only sensible choice. By the definition of the problem, the greater length of the later arrow would not be more rewarding but superfluous. If there is a second goal of avoiding a disaster – their "MR" – then, obviously, no response that permitted the disaster would be rewarded. The competition between responses would then be based on each option's chance of getting the positive goal (how probable, how rewarding, how soon?) and its risk of incurring disaster (how probable, what loss of reward, how soon?). The upward and downward baselines would have an impact to the extent that they affected these dimensions. Hyperbolic discount curves predict that the distance to the changes of reward from either goal or disaster would disproportionately influence the person's choice of options when this distance was short, unless she had made her choices the subject of an effective personal rule.

**Stanovich** suggests that "cognitivism in the form of dual-process theories of reasoning" is compatible with intertemporal bargaining. I agree. Recursive self-prediction certainly involves cognitive processes, often subtle ones, as do many other interpretations that are central to the growth of farsighted mental processes. My criticism of cognitivism has to do with the seemingly doctrinaire avoidance of any selective principle (like reward) that characterizes some writings in this school – the insistence that ego functions are above that kind of thing, which might otherwise reduce them to mere mechanisms. (Is this implicit in **Bach**'s commentary?) Stanovich largely covers the informational processes that I agree must be involved in conflictual choice-making, and recognizes that one process is selected over another on the basis of "motivational force." Even the point of disagreement that he names is a misunderstanding, the product of an ambiguity on my p. 146: When I speak about human intelligence creating an impulse problem, I am referring to the inventions that have given us so much freedom of choice, not the source of urges themselves.

However, I would like **Stanovich** and the many other advocates of dual-process theories to consider the possibility that the commonly listed pairs of autonomous and analytic, or visceral and cerebral, or hot and cool, processes do not represent dichotomies but two zones on a continuum, and not the most distant zones at that. Between the autonomous and analytic processes there are those of middling rapidity, partial responsiveness to triggering stimuli, and marginal controllability by high-level control systems. Hot anger shades into grudge and burning love into attachment; the

fear of imminent collision morphs through the fear of flunking today's test to the fear that civilization is declining. Probably all delayed incentives must motivate with immediate emotional representations – the process that is conspicuously absent in the amotivation of severe depressives, and that runs amok in the grandiose planning of manics. That is, although analytic processes are often alternatives to autonomous ones, they also harness these necessary sources of reward more or less extensively. The brain imaging studies that have been held to demonstrate two alternative processes actually show activity in analytic centers during all contingencies of reward (always token rewards so far, however); they do show activity in the supposedly visceral centers only in imminent contingencies, but the designs have not been such as to discriminate on-off dichotomies from continua (e.g., McClure et al. 2004; Tanaka et al. 2004).

The key reason to hypothesize a continuum of analyticness (which I imagine to vary with duration of dominance, but that is not essential) is that a given process may sometimes compete with a higher function, and sometimes with a lower one. Getting drunk is not as autonomous as emitting a tic, and a tic not as autonomous as succumbing to panic; for that matter, compulsive self-control could be said to be more autonomous than flexible choice. In other words, there can be more than one tier of metapreferences, for instance in my example from *Breakdown of Will*, p. 63:

> Take a patient trying to overcome bulimia nervosa. Her longest range interest is to eat flexibly, but she has a compulsion to strictly limit her intake. This compulsion arises from a perception that she has an addiction to eating, which periodically breaks out in the form of a binge. Let's say, to create a crude example, that her bingeing is sometimes preyed upon by an obsessional worry that her food is spoiled, an itch range [of reward durations] urge that spoils the pleasure of eating – and that she then stubs her toe, and has such pain that she can't even entertain her worry – we have what could be called a chain of predation containing . . . experientially distinct ranges of preference.

All but the longest range interest could be called autonomous in some sense, and all but the pain have some component of analysis.

I digress for a moment to mention **Bach**'s three motivational factors as alternatives to hyperbolic overvaluation in "the motivational force of a particular desire": persistence of coming to mind, insatiability, and "resistibility to the second-order desire to get rid of it." These factors can be created parsimoniously by the rapid cycling of brief immediate reward and longer obligatory nonreward that I propose for wanted but disliked behaviors (itches; *Breakdown of Will*, pp. 51–54 and target article Précis, sect. 4). In the preceding paragraph, tics and obsessional worries play this role. That is, all it takes to put the pesky video game into a purely reward-based model is an urge that is strong enough to command your attention briefly with the promise of immediate payoff, that is quickly disposed of, but that also regenerates quickly. Once the urge succeeds in getting you playing, of course, the math of reward is more like that of a binge, and the temporary preference considerably longer.

I agree with **Connolly & Reb** that there is, first of all, a simple "outcome regret" that does not follow a failure of self-control. I have speculated that its proportionality with the nearness of the miss comes from the person's urge to break her rules for constructing beliefs (Ainslie 1992, pp. 321–23). It is less threatening than the "regret associated with the choice process itself," an experience that is often called guilt. Connolly & Reb discuss anticipated regret as an important incentive for self-control, and I agree with this, also; but fear of the choice-process kind of regret is really the same incentive as the expectation of a bundle of future rewards, seen from the converse angle as the threat of losing them. A highly meaningful failure of will can induce a guilt (or regret) that is close to panic, since it may lower a person's expectation of self-control generally. What Connolly & Reb call increased regret salience, I would call increased personal rule salience, the perception of more of your behaviors as test cases for your will. Increasing the specificity of your intentions has indeed been shown to increase your self-control (Gollwitzer, in press), and may well entail both increased expectations of bundled rewards and increased fear of losing them. But increased reliance on willpower makes you more compulsive (see *Breakdown of Will*, pp. 151–53), a condition noted for a preoccupation with avoiding regret.

**Griffin & Dennett** point out that little is known as yet about the constraints on bundling. In particular, "Does each bit of the bundle have the same value?" There are some data showing that, with small bundles in controlled experiments, the value of the bundle is the simple sum of the values of the component rewards, discounted hyperbolically (Mazur 1986). It is also clear that in naturalistic situations, the value of a stretch of experience as measured by willingness to repeat it is *not* the sum of the values of its component moments, but receives disproportionate weight from the most intense and the most recent parts (Kahneman 1999, pp. 19–22). Because no choices repeat exactly, and most choices share some feature or other in common, the make-up of a bundle clearly depends on personal bookkeeping, perhaps mediated by socially taught language, but I think not necessarily so.

What constrains your perception of precedents and hence bundles? In the simplest model, it is your prediction of what you will see as a precedent in the future, arrived at by whatever cognitive processes and influenced by whatever lawyerly pleading you may have learned by trial and error to prevent a realization of having fooled yourself. An executive "I" may earn its keep by expertise at discerning bundles, but the process can also happen without central direction, from incentives analogous to those that moved generations of unsupervised judges to develop the English common law. If the pianist claims an exception to his no-wine rule because "It's just a chamber rehearsal," and then realizes, "but I *knew* that the esteem of my colleagues was still at stake," he has made an error. This error will cost him some of the expected reward of reputation, and not just from this episode; the recognition that "I *knew* that" identifies a lapse, which makes his no-wine rule and maybe his will in general less credible. **Griffin & Dennett** are right that he may view a lapse involving his colleagues as less serious than a lapse involving the public, or, if he has alcoholic tendencies, he may view all lapses involving wine as equally dangerous. But these views are not at his arbitrary disposal; they are predictions about the fact of his subsequent views, which determine the stake available to bet against impulses. Although there is wiggle room in both the prediction and the determination of these views (the effect of rationalization), both are ultimately constrained by the reward they deliver; too much wiggling turns them into useless self-delusions and leaves him at the mercy of his impulses.

**Peijnenburg** deals with one kind of wiggle room. She

says that she, and I, are dealing with the "easy," diachronic form of *akrasia*; but my position is that relying on personal rules, while fortifying you against easy *akrasia*, also creates a potential for the hard, synchronic form. That is, bundling choices changes the basis of impulses from the simple proximity of a possible SS reward to the existence of loopholes or other weaknesses in the personal rule. All the examples that authors have described of doing what you simultaneously believe not to be best can be seen as obeying an impulse in violation of a personal rule. You "really" wanted the long-range rewards defended by your rule, but in the case at hand the rule was not strong enough. Or you found a rationalization that distinguished the case at hand, which might or might not still be counted as *akrasia*. I would call Peijnenburg's example of the married woman with the single lapse of fidelity a case of rationalization. The woman is asserting that an occasional lapse is best called an exception, a stance that, taken after the fact, may preserve adequate credibility for her rule. I agree that this is probably the wisest interpretation for the woman to make. The optimal degree of tolerance for lapses has been discussed in addiction psychology, with the conclusion that the most rigid rules, "zero tolerance," invite an "abstinence violation effect," the complete collapse of resolve in the relevant area after a lapse (Curry et al. 1987). On the other hand, to plan such an exception in advance would be to start down a slippery slope. In order to preserve the rule in question by "shaping your past [self]," the lapse had better have been a surprise, as it was in this example; and even so, repeated appeal to such shaping would motivate short-range interests to inhibit your self-awareness, so that more lapses could be surprising (see Ainslie 1999). Because the lapse must not be planned in advance, Peijnenburg's shaping effect looks as if it is operating retroactively, but actually it is a current interpretation (as of the morning after) that is designed to affect future choices (by repairing the expectation of being a faithful wife). To affect choice, all reward differentials must be prospective.

## R6. Freedom of will

**Bridgeman** argues that choice must be strictly determined by prior causes, which is also my position. I did not develop it in the target Précis, but I did in the book (*Breakdown of Will*, pp. 129–34). I argue that there is a genuine distinction to be made between choices that are the straightforward product of identifiable incentives, and those that are subject to the recursive self-prediction that arises in intertemporal bargaining. We feel driven by the former, but are open to surprise by the latter, because of the sensitive dependence of its outcome on small shifts of predictive cues. I claim that the location of recursiveness within the will itself adds the sense of ownership, the lack of which has led other authors to reject such a process as the mechanism of free will. This solution involves neither illusion nor indeterminacy:

> Of course, mere dependency on internally fed back processes doesn't create the feeling of being a self: "If chaos-type data can be used to justify the existence of free will in humans, they can also be used to justify the existence of free will in chaotic pendulums, weather systems, leaf distribution, and mathematical equations (Sappington 1990)." That is, even information-rich processes that don't somehow engage what feels like our self will still be experienced as random, "more like epileptic seizures than free responsible choices (Kane 1989, p. 231)." So far, chaos theory has not been given an element that internalizes the process. [However,] I'm arguing that intertemporal bargaining supplies that extra element: that your own motivation – in many cases emotion – is what you're predicting. (*Breakdown of Will*, p. 233)

If recursive self-prediction turns out to be coterminous with what we experience as free will, I have no problem with calling it free will. Were the experience specifically one of indeterminacy it would be an illusion, of course; but as Bridgeman and many others have pointed out, "we don't confuse lack of control with freedom." The distinction between free will and the alternative usually imagined, what might be called stimulus-boundness, is not an illusory one.

## R7. Thought experiments

**Rachlin** points out that I am violating the behaviorist convention against internal models by trying to discern mental mechanisms. The original purpose of this convention was to promote the productive methodology of controlled experimentation in the face of an entrenched tradition of introspection (Boring 1950, pp. 641–43). It succeeded because the methodology proved very fruitful indeed, and described a number of regularities in behaviors upon which external events were contingent. However, it never demonstrated that mental mechanisms did not exist or even that they were unimportant, just that a great deal could be learned while obeying a discipline that disregarded them. Ironically, exploration of one of the most robust of these regularities, Herrnstein's matching law (1997), has produced evidence consistent with a model of reward as an internal event, one that is less dependent on external events than has been heretofore assumed by any branch of motivational science. Hyperbolic discounting has removed theoretical obstacles to regarding such disparate phenomena as appetites, emotions, involuntary behaviors, and ego functions as reward-dependent behaviors competing in a marketplace, rather than as transferred reflexes or transcendent cognitions (see *Breakdown of Will*, pp. 48–70, 90–104). I would argue that these theoretical obstacles were the last pivotal ones, and that the behaviorist convention is not an obstacle even theoretically. It is a self-control device for theorists, designed to ensure consensual validation of findings. The degree to which an individual, or a field, is willing to risk failures of consensual validation in order to increase the sensitivity of observation is ultimately a matter of taste.

In the thirty years since the core of hyperbolic discounting theory was first published (Ainslie 1975), no one has raised a reason why it *could not* be true; the intertemporal bargaining that it predicts is at the very least a possibility. Any discipline that rules this out because of the difficulty of studying it will be left with a restricted field of view. Such a discipline will be ignoring suggestive evidence from many sources: pattern-matching, interpersonal analogs, computer models, . . . and thought experiments (*Breakdown of Will*, Ch. 8). The trouble with introspectionism was that the designs of Titchener and his generation did not elicit observations that all could share – the equivalent of "native speaker" recognitions in linguistics or the perceptual shifts of Gestalt psychology. Thought experiment is a way of producing highly specific shared experiences. It has been developed in recent years mostly by the philosophy of mind,

a field that seems to have a discipline opposite to behaviorism – *never* rely on controlled experiment. I offer it as a complementary way to test theories that purport to explain our experience.

The function of thought experiments is to clarify what knowledge we already have by testing its implications in simplified situations. Entire fields have been created by this method – the exemplar was plane geometry – but sufficient examples are scarce in motivational theory. **Rachlin** and **Arló-Costa** criticize my application of Newcomb's problem, the former because, aside from his behaviorist convictions, he does not think that the counterfactual premise evokes a meaningful experience, the latter because he does not see how hyperbolic discounting predicts a rational solution. Insofar as a reader cannot relate to a thought experiment, or thinks she should not, she will need to see the reports of an adequate sample of others, which are actually available in the Newcomb case and confirm a widespread disposition to choose one box (Shafir & Tversky 1992). However, what the introspection, or sample, demonstrates is not necessarily the rationality of one solution, but something about the nature of the solving. If the reader, or sample of undergraduates, feel like choosing one box, this is evidence that they are going by a consideration beyond what the conventional view of utility (rational choice theory, or RCT) predicts. That is the service of this thought experiment – to demonstrate that RCT is missing a piece here. The only hypothesis about why RCT does not describe people's one-box disposition has been magical thinking – your belief that by faking the diagnosis of being a one-boxer you will cause the $1 million to be in that box. Intertemporal bargaining offers an alternative hypothesis, that people are intuitively familiar with the kind of *diagnostic* thinking that is also *causal* and apply it to the Newcomb presentation.

Perhaps the evocativeness of the Newcomb problem is best explained if we bypass its externalizing dollar values and guess how someone might interpret it in terms of internal utility: The first box contains your aggregated expectation of reward for resisting your greatest temptation over time, the second your reward for giving in to it this time. The omniscient being dwells somewhere in your own consciousness. (At least there is something there that seems omniscient, always saying, "I told you so.") It is literally possible to get the contents of both boxes, but this situation, involving as it does your greatest temptation, is not new to you. You know by now whether you are a two-boxer in this situation, and this knowledge governs both your choice (knowing you are a two-boxer undermines any resolve not to give in to your temptation) and the contents of the first box (empty for the two-boxer, full for the one boxer). The missing piece in RCT as applied to Newcomb's problem, my hypothesis goes, is the role played in your choice by what you know (omnisciently, if we are to be strict). Knowing that you are a one-boxer fills the box with expectations of reward (= the present value of the aggregated actual rewards to come); knowing that you are a two-boxer empties it. Bundling choices converts not only intertemporal conflicts to conflicts of principle, it converts actions to traits: Insofar as you have always chosen two boxes you *are* a two-boxer, a trait that governs your choices unless some factor happens to balance one of them closely. In real life there are many things you can do to overcome a two-boxer trait, of course, or to wreck a one-boxer trait, but Newcomb's problem serves to illustrate the pure extremes. Incidentally, **Arló-Costa** has me wrong: Two-boxers are irrational, as are the people who renege on drinking Kavka's toxin (*Breakdown of Will*, pp 126–29).

## R8. The state of research

**Sanabria & Killeen** peg the current state of picoeconomic knowledge accurately. As they point out, Skinner perceived that strategically related selves could be shaped within a person by operant reward. However, hyperbolic discounting was necessary to explain why there should be separate selves, not a unified organization of behaviors all seeking the same goals. The explanation "needs work. . . . We need a chemistry of intra-individual interests. When are they bundled? When broken down? When parasitized?"

Chapter 8 of *Breakdown of Will* reviewed progress toward answers, as of 2001 (pp. 117–40). Controlled research on the components of motivation continues, some of it by the commentators on the Précis. Interpersonal analogs of the hypothesized intertemporal processes can be run (Monterosso et al. 2002), but they are noisy, and complicated by the unavoidable admixture of intertemporal bargaining within the participants (see Ainslie, in press b). A rereading of the psychoanalytic literature, substituting hyperbolic curves for repression as the source of impulses, suggests some dynamics. But correspondences with this once-dominant distillation of psychological anecdotes enrich intertemporal bargaining theory more than they test it. For tests of the coherence and naturalism of intertemporal bargaining theory, as well as its downstream consequences, I am returning to computer simulation (Ainslie 1992, pp. 274–91; 2005b) just as **Sanabria & Killeen** suggest. A team at the University of KwaZulu-Natal is developing of an automated working model of a picoeconomic agent, hoping that the hypothesized need for surprise suggests a rationale for gambling behavior. As for fitting this theory to the real world, the problem is coming within the reach of neurophysiology, but only just. Finally, the potential of the much maligned thought experiment to test the actual occurrence of predicted processes remains controversial, but I believe that it is far from nil (Ainslie 2005a). We witness the operation of our own wills on a daily basis, and may be prevented from reporting it clearly more by our assumptions than by any lack of its visibility.

NOTE

The author of this response is employed by a government agency and, as such, this response is considered a work of the U.S. government and not subject to copyright within the United States.

## References

**Letters "a" and "r" appearing before authors' initials refer to target article and response respectively.**

Ainslie, G. (1974) Impulse control in pigeons. *Journal of the Experimental Analysis of Behavior* 21:485–89. [aGA]

(1975) Specious reward: A behavioral theory of impulsiveness and impulse control. *Psychological Bulletin* 82:463–96. [arGA]

(1986) Beyond microeconomics: Conflict among interest in a multiple self as a determinant of value. In: *The multiple self*, ed. J. Elster, pp. 133–75. Cambridge University Press. [aGA]

(1991) Derivation of "rational" economic behavior from hyperbolic discount curves. *American Economic Review* 81:334–40. [rGA]

(1992) *Picoeconomics: The strategic interaction of successive motivational states within the person*. Cambridge University Press. [arGA, DR]
(1995) A utility-maximizing mechanism for vicarious reward: Comments on Julian Simons "Interpersonal allocation continuous with intertemporal allocation." *Rationality and Society* 7:393–403. [aGA]
(1999) The dangers of willpower: A picoeconomic understanding of addiction and dissociation. In: *Getting hooked: Rationality and addiction*, ed. J. Elster & O-J. Skog, pp. 65–92. Cambridge University Press. [rGA]
(2001) *Breakdown of will*. Cambridge University Press. [arGA, HA-C, KB, BB, TC, LF, LG, RG, GWH, JP, HR, DR, FS, KES, JDS, XTW]
(2005a) Can thought experiments prove anything about the will? Presented at the conference, *Mind and World II*, University of Alabama, Birmingham, March 18, 2005; proceedings to be published by Routledge. [rGA]
(2005b) Emotion as motivated behavior. In: *Agents that want and like: Motivational and emotional roots of cognition and action. Proceedings of a Symposium at the Artificial Intelligence and Simulation of Behaviour '05 Convention, Hatfield, U.K.*, AISB ed. L. Canamero, pp. 1–8. [rGA]
(in press a) A selectionist model of the ego: Implications for self-control. In: *Disorders of volition*, ed. N. Sebanz & W. Prinz. MIT Press. [rGA]
(in press b) You can't give permission to be a bastard. Commentary on Henrich et al. (2005), "Economic man" in cross-cultural perspective. *Behavioral and Brain Sciences* 28(6). [rGA]
Ainslie, G. & Haendel, V. (1983) The motives of the will. In: *Etiology aspects of alcohol and drug abuse*, ed. E. Gottheil, K. Druley, T. Skodola & H. Waxman, pp. 119–40. Charles C. Thomas. [arGA, GWH]
Ainslie, G. & Herrnstein, R. J. (1981) Preference reversal and delayed reinforcement. *Animal Learning and Behavior* 9:476–82. [aGA]
Ainslie, G. & Monterosso, J. (2003a) Building blocks of self-control: Increased tolerance for delay with bundled rewards. *Journal of the Experimental Analysis of Behavior* 79:83–94. [aGA]
(2003b) Hyperbolic discounting as a factor in addiction: A critical analysis. In: *Choice, behavioural economics, and addiction*, ed. R. R. Vuchinich & N. Heather, pp. 35–62. Pergamon. [rGA]
American Psychiatric Association (1994) *Diagnostic and statistical manual of mental disorders, 4th edition*. APA Press. [aGA, LG]
Aristotle (1984) *The Complete Works of Aristotle*, ed. J. Barnes. Princeton University Press.
Arló-Costa, H. & Helzner, J. (2005) Comparative ignorance and the Ellsberg's phenomenon. In: *ISIPTA '05: Proceedings of 4th International Symposium on Imprecise Probabilities and Their Applications*, ed. F. G. Cozman, R. Nau & T. Seidenfeld, pp. 21–31. Brightdoc. Available at: http://www.sipta.org/isipta05/proceedings/index.html. [HA-C]
Baier, A. (1991) *A progress of sentiments: Reflections on Hume's treatise*. Harvard University Press. [aGA]
Batson, C. D. & Shaw, L. L. (1991) Evidence for altruism: Toward a pluralism or prosocial motives. *Psychological Inquiry* 2:159–68. [aGA]
Baum, W. M. (1974) On two types of deviation from the matching law: Bias and undermatching. *Journal of the Experimental Analysis of Behavior* 22:231–42. [rGA]
Baumeister, R. F. & Heatherton, T. (1996) Self-regulation failure: An overview. *Psychological Inquiry* 7:1–15. [aGA]
Becker, G. & Murphy, K. (1988) A theory of rational addiction. *Journal of Political Economy* 96:675–700. [aGA]
Beecher, H. (1959) *Measurement of subjective responses*. Oxford University Press. [aGA]
Bell, D. E. (1982) Regret in decision making under uncertainty. *Operations Research* 30:961–81. [TC]
Benzion, U., Rapoport, A. & Yagil, J. (1989) Discount rates inferred from decisions: An experimental study. *Management Science* 35:270–84. [GWH]
Berlyne, D. E. (1974) *Studies in the new experimental aesthetics*. Hemisphere. [aGA]
Berns, G. S., McClure, S. M., Pagnoni, G. & Montague, P. R. (2001) Predictability modulates human brain response to reward. *Journal of Neuroscience* 21:2793–98. [arGA]
Berridge, K. C. (1999) Pleasure, pain, desire, and dread: Hidden core processes of emotion. In: *Well-being: The foundations of hedonic psychology*, ed. D. Kahneman, E. Diener & N. Schwartz. Sage. [rGA]
Berridge, K. C. & Robinson, T. (1998) What is the role of dopamine in reward: Hedonic impact, reward learning, or incentive salience. *Brain Research Reviews* 28:309–69. [aGA]
Boring, E. G. (1950) *A history of experimental psychology*. Appleton-Century-Crofts. [rGA]
Bratman, M. E. (1999) *Faces of intention: Selected essays on intention and agency*. Cambridge University Press. [aGA]
Carson, S. M., Moses, L. J. & Hix, H. R. (1998) The role of inhibitory processes in young children's difficulties with deception and false belief. *Child Development* 69:672–91. [RG]
Carver, C. S., Lawrence, J. W. & Scheier, M. F. (1996) A control-process perspective on the origins of affect. In: *Striving and feeling: Interactions among goals, affect, and self-regulation*, ed. L. L. Martin & A. Tesser, pp. 11–52. Erlbaum. [XTW]
Chapman, G. B. & Elstein, A. S. (1995) Valuing the future: Temporal discounting of health and money. *Medical Decision Making* 15:373–86. [LG]
Chung, S. & Herrnstein, R. J. (1967) Choice and delay of reinforcement. *Journal of the Experimental Analysis of Behavior* 10:67–74. [aGA]
Clum, George A., Clum, Gretchen A. & Surls, R. (1993) A meta-analysis of treatments for panic disorder. *Journal of Consulting and Clinical Psychology* 61:317–26. [aGA]
Coller, M., Harrison, G. W. & Rutström, E. E. (2003) Are discount rates constant? Reconciling theory and observation. Working Paper 3–31, Department of Economics, College of Business Administration, University of Central Florida. [GWH]
Coller, M. & Williams, M. B. (1999) Eliciting individual discount rates. *Experimental Economics* 2:107–27. [GWH]
Connolly, T. & Reb, J. (2003) Omission bias in vaccination decisions: Where's the "omission"? Where's the "bias"? *Organizational Behavior and Human Decision Processes* 91:186–202. [TC]
Connolly, T. & Zeelenberg, M. (2002) Regret in decision making. *Current Directions in Psychological Science* 11:212–20. [TC]
Currie, G. & Ravenscroft, I. (2002) *Recreative minds*. Oxford University Press. [KES]
Curry, S., Marlatt, A. & Gordon, J. R. (1987) Abstinence violation effect: Validation of an attributional construct with smoking cessation. *Journal of Consulting and Clinical Psychology* 55:145–49. [rGA]
Davidson, D. (1969) How is weakness of the will possible? In: *Moral concepts*, ed. J. Weinberg, pp. 93–113. Oxford University Press. [JP]
Davies, N. (1981) *Human sacrifice in history and today*. Morrow. [aGA]
Davis, J. (2003) *The theory of the individual in economics*. Routledge. [DR]
Deluty, M. Z., Whitehouse, W. G., Mellitz, M. & Hineline, P. N. (1983) Self-control and commitment involving aversive events. *Behavior Analysis Letters* 3:213–19. [aGA]
Dennett, D. C. (1978) Artificial intelligence as philosophy and as psychology. In: *Brainstorms: Philosophical essays on mind and psychology*, ed. D. C. Dennett, pp. 109–26. Bradford Books/MIT Press. [rGA]
(1987) *The intentional stance*. MIT Press. [DR]
(2003) *Freedom evolves*. Viking. [KES]
Dienes, Z. & Perner, J. (1999) A theory of implicit and explicit knowledge. *Behavioral and Brain Sciences* 22:735–808. [KES]
Donahoe, J. W., Burgos, J. E. & Palmer, D. C. (1993) A selectionist approach to reinforcement. *Journal of the Experimental Analysis of Behavior* 60:17–40. [aGA]
Donahoe, J. W., Palmer, D. C. & 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. [aGA]
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*, pp. 43–108. Cambridge University Press. [aGA]
Evans, J. St. B. T. (2003) In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Sciences* 7:454–59. [KES]
Evans, J. St. B. T. & Over, D. E. (2004) *If*. Oxford University Press. [KES]
Fodor, J. A. (1983) *The modularity of mind*. Bradford Books. [FS]
Forzano, L. B. & Logue, A. L. (1992) Predictors of adult humans' self-control and impulsiveness for food reinforcers. *Appetite* 19:33–47. [aGA]
Fox, C. R. & Tversky, A. (1991) Ambiguity aversion and comparative ignorance. *The Quarterly Journal of Economics* 110(3):585–603. [HA-C]
Frank, R. H. (1988) *Passions within reason*. Norton. [aGA]
Frijda, N. H. (1988) The laws of emotion. *American Psychologist* 43:349–58. [TC]
Gerall, A. A. & Obrist, P. A. (1962) Classical conditioning of the pupillary dilation response of normal and curarized cats. *Journal of Experimental Psychology* 50:261–63. [aGA]
Gergen, K. J. (1985) The social constructionist movement in modern psychology. *American Psychologist* 40:266–75. [aGA]
Gibbon, J. (1977) Scalar expectancy theory and Weber's law in animal timing. *Psychological Review* 84:279–325. [rGA]
Glenberg, A. M. (1997) What memory is for. *Behavioral and Brain Sciences* 20:1–55. [KES]
Gollwitzer, P. M. (in press) If–then plans and the intentional control of thoughts, feelings, and actions. In: *Disorders of volition*, ed. N. Sebanz & W. Prinz. MIT Press. [rGA]
Gosselin, P., Kirouac, G. & 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. [aGA]
Grace, R. C. (1994) A contextual model of concurrent chains choice. *Journal of the Experimental Analysis of Behavior* 61:113–29. [aGA]

Granda, A. M. & Hammack, J. T. (1961) Operant behavior during sleep. *Science* 133:1485–86. [aGA]

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

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

Green, L., Fry, A. & Myerson, J. (1994b) Discounting of delayed rewards: A life-span comparison. *Psychological Science* 5:33–36. [arGA]

Green, L. & Myerson, J. (1993) Alternative frameworks for the analysis of self control. *Behavior and Philosophy* 21:37–47. [LG]

Green, L., Myerson, J. & Macaux, E. W. (in press) Temporal discounting when the choice is between two delayed rewards. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. [rGA, LG]

Green, L., Myerson, J. & McFadden, E. (1997) Rate of temporal discounting decreases with amount of reward. *Memory and Cognition* 25:715–23. [LG]

Harland, R. (1987) *Superstructuralism: The philosophy of structuralism and post-structuralism*. Methuen. [aGA]

Harrison, G. W., Lau, M. I. & Rutström, E. E. (2005) Dynamic consistency in Denmark: A field experiment. Working Paper 5–02, Department of Economics, College of Business Administration, University of Central Florida, January 2005. [rGA, GWH]

Harrison, G. W., Lau, M. I. & Williams, M. B. (2002) Estimating individual discount rates for Denmark: A field experiment. *American Economic Review* 92(5):1606–17. [GWH]

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

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

Hearst, E. & Jenkins, H. M. (1974) *Sign-tracking: The stimulus-reinforcer relation and directed action*. The Psychonomic Society. [FS]

Heath, C., Larrick, R. P. & Wu, G. (1999) Goals as reference points. *Cognitive Psychology* 38:79–109. [XTW]

Herrnstein, R. J. (1997) *The matching law: Papers in psychology and economics*, ed. H. Rachlin & D. I. Laibson. Sage. [rGA]

Herrnstein, R. J. & Prelec, D. (1992) A theory of addiction. In: *Choice over time*, ed. G. F. Loewenstein & J. Elster, pp. 331–60. Sage. [aGA]

Heyman, G. M. (1996) Resolving the contradictions of addiction. *Behavioral and Brain Sciences* 19:561–610. [aGA]

Hilgard, E. R. & Hilgard, J. R. (1994) *Hypnosis in the relief of pain*, revised edition. Brunner/Mazel. [aGA]

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

Holcomb, J. H. & Nelson, P. S. (1992) Another experimental look at individual time preference. *Rationality and Society* 4(2):199–220. [GWH]

Hollerman, J. R., Tremblay, L. & Schultz, W. (1998) Influence of reward expectation on behavior-related neuronal activity in primate striatum. *Journal of Neurophysiology* 80:947–63. [aGA]

Hutchinson, J. M. C. & Gigerenzer, G. (2005) Simple heuristics and rules of thumb; where psychologists and behavioral biologists might meet. *Behavioural Processes* 69:97–124. [FS]

Iacoboni, M., Woods, R. P., Brass, M., Bekkering, H., Mazziotta, J. C. & Rizzolatti, G. (1999) Cortical mechanisms of imitation. *Science* 286:2526–28. [rGA]

Joyce, J. M. (1999) *The foundations of causal decision theory*. Cambridge University Press. [HA-C]

Kacelnik, A. & Bateson, M. (1997) Risk-sensitivity: Crossroads for theories of decision making. *Trends in Cognitive Science* 1:304–309. [XTW]

Kahneman, D. (1999) Objective happiness. In: *Well-being: The foundations of hedonic psychology*, ed. D. Kahneman, E. Diener & N. Schwartz, pp. 3–25. Sage. [rGA]

(2003) A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist* 58:697–720. [TC]

Kahneman, D. & Frederick, S. (2002) Representativeness revisited: Attribute substitution in intuitive judgment. In: *Heuristics and biases: The psychology of intuitive judgment*, ed. T. Gilovich, D. Griffin & D. Kahneman, pp. 49–81. Cambridge University Press. [KES]

Kahneman, D. & Tversky, A. (1979) Prospect theory. *Econometrica* 47:263–92. [XTW]

Kane, R. (1989) Two kinds of incompatibilism. *Philosophy and Phenomenological Research* 50:220–54. [rGA]

Kant, I. (1793/1960) *Religion within the limits of reason alone*, trans. T. Green & H. Hucken, pp. 15–49. Harper and Row. [aGA]

Karmiloff-Smith, A. (1992) *Beyond modularity: A developmental perspective on cognitive science*. MIT Press. [KES]

Kavka, G. (1983) The toxin puzzle. *Analysis* 43:33–36. [aGA]

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* 66:175–78. [aGA]

Kirby, K. N. (1997) Bidding on the future: Evidence against normative discounting of delayed rewards. *Journal of Experimental Psychology: General* 126:54–70. [aGA, LG]

Kirby, K. N. & Guastello, B. (2001) Making choices in anticipation of similar future choices can increase self-control. *Journal of Experimental Psychology: Applied* 7:154–64. [arGA]

Kirby, K. N. & Marakovic, N. N. (1995) Modeling myopic decisions: Evidence for hyperbolic delay-discounting within subjects and amounts. *Organizational Behavior and Human Decision Processes* 64:22–30. [aGA, LG]

Klein, B. & Leffler, K. B. (1981) The role of market forces in assuring contractual performance. *Journal of Political Economy* 89:615–40. [aGA]

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. [aGA]

Laibson, D. (1997) Golden eggs and hyperbolic discounting. *Quarterly Journal of Economics* 62:443–79. [arGA]

Landman, J. (1993) *Regret: The persistence of the possible*. Oxford University Press. [TC]

Larrick, R. P. & Boles, T. L. (1995) Avoiding regret in decisions with feedback: A negotiation example. *Organizational Behavior and Human Decision Processes* 63:87–97. [TC]

Levi, I. (1975) Newcomb's many problems. *Theory and Decision* 6:161–75. [HA-C]

(1986) The paradoxes of Allais and Ellsberg. *Economics and Philosophy* 2:23–56. [HA-C]

Licklider, J. C. R. (1959) On psychophysiological models. In: *Sensory communication*, ed. W. A. Rosenbluth. MIT Press. [aGA]

Loewenstein, G. (1996) Out of control: Visceral influences on behavior. *Organizational Behavior and Human Decision Processes* 35:272–92. [aGA]

Logue, A. W. (1988) Research on self-control: An integrating framework. *Behavioral and Brain Sciences* 11:665–78. [FS]

Loomes, G. & Sugden, R. (1982) Regret theory: An alternative theory of rational choice under uncertainty. *Economic Journal* 92:805–24. [TC]

Lopes, L. L. (1987) Between hope and fear: The psychology of risk. *Advances in Experimental Social Psychology* 20:255–95. [XTW]

Lotka, A. (1957) *Elements of mathematical biology*. Dover. [rGA]

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

Machina, M. J. (1989) Dynamic consistency and non-expected utility models of choice under uncertainty. *Journal of Economic Literature* 27:1622–68. [GWH]

Malekzadeh, A. R. & Nahavandi, A. (1987) Merger mania: Who wins? Who loses? *Journal of Business Strategy* 8:76–79. [aGA]

Marks, I. & Tobena, A. (1990) Learning and unlearning fear: A clinical and evolutionary perspective. *Neuroscience and Biobehavioral Reviews* 14:365–84. [aGA]

Mazur, J. E. (1986) Choice between single and multiple delayed reinforcers. *Journal of the Experimental Analysis of Behavior* 46:67–77. [rGA]

(1987) An adjusting procedure for studying delayed reinforcement. In: *Quantitative analyses of behavior V: The effect of delay and of intervening events on reinforcement value*, ed. M. L. Commons, J. E. Mazur, J. A. Nevin & H. Rachlin. Erlbaum. [aGA]

(1997) Choice, delay, probability, and conditioned reinforcement. *Animal Learning and Behavior* 25:131–47. [aGA]

McClennan, E. F. (1990) *Rationality and dynamic choice*. Cambridge University Press. [aGA, GWH]

McClure, S. M., Laibson, D. I., Loewenstein, G. & Cohen, J. D. (2004) The grasshopper and the ant: Separate neural systems value immediate and delayed monetary rewards. *Science* 306:503–507. [rGA, DR]

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

McGaw, C. (1966) *Acting is believing: A basic method*. Holt, Rinehart & Winston. [aGA]

Meek, C. & Glymour, C. (1994) Conditioning and intervening. *British Journal for the Philosophy of Science* 45:1001–21. [HA-C]

Mellers, B. A., Schwartz, A. & Ritov, I. (1999) Emotion-based choice. *Journal of Experimental Psychology: General* 128:332–45. [TC]

Melzack, R. & Casey, K. L. (1970) The affective dimension of pain. In: *Feelings and emotions*, ed. M. B. Arnold, pp. 55–68. Academic Press. [aGA]

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

Millar, A. & Navarick, D. J. (1984) Self-control and choice in humans: Effects of video game playing as a positive reinforcer. *Learning and Motivation* 15:203–18. [aGA]

Miller, N. (1969) Learning of visceral and glandular responses. *Science* 163:434–45. [aGA]

Mischel, H. N. & Mischel, W. (1983) The development of children's knowledge of self-control strategies. *Child Development* 54:603–19. [aGA]

Montague, P. R. & Berns, G. S. (2002) Neural economics and the biological substrates of valuation. *Neuron* 36:265–84. [DR]

Montegut, M. J., Bridgeman, B. & Sykes, J. (1997) High refresh rate and oculomotor adaptation facilitate reading from video displays. *Spatial Vision* 10:305–22. [BB]

Monterosso, J. R., Ainslie, G., Toppi Mullen, P. & Gault, B. (2002) The fragility of cooperation: A false feedback study of a sequential iterated prisoner's dilemma. *Journal of Economic Psychology* 23:437–48. [rGA]

Myerson, J. & Green, L. (1995) Discounting of delayed rewards: Models of individual choice. *Journal of the Experimental Analysis of Behavior* 64:263–76. [aGA, LG]

Myerson, J., Green, L., Hanson, J. S., Holt, D. D. & Estle, S. J. (2003) Discounting delayed and probabilistic rewards: Processes and traits. *Journal of Economic Psychology* 24:619–35. [LG]

Navarick, D. J. (1982) Negative reinforcement and choice in humans. *Learning and Motivation* 13:361–77. [aGA]

Nemiah, J. C. (1977) Alexithymia: Theoretical considerations. *Psychotherapy and Psychosomatics* 28:199–206. [aGA]

Nichols, S. & Stich, S. P. (2003) *Mindreading: An integrated account of pretence, self-awareness, and understanding other minds*. Oxford University Press. [KES]

Nozick, R. (1969) Newcomb's problem and two principles of choice. *Essays in honor of Carl G. Hempel*, ed. N. Rescher, pp. 107–33. Reidel. [HA-C]

(1993) *The nature of rationality*. Princeton University Press. [aGA]

Ostaszewski, P., Green, L. & Myerson, J. (1998) Effects of inflation on the subjective value of delayed and probabilistic rewards. *PsychonomicBulletin and Review* 5:324–33. [DR]

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. [aGA]

(1993) Beyond hedonism: Motives for inhibiting good moods and for maintaining bad moods. In: *Handbook of mental control*, ed. D. M. Wegner & F. W. Pennebaker. Prentice-Hall. [aGA]

Peijnenburg, J. (2004) Nemen gedane zaken geen keer? (Is what is done done?). *General Dutch Journal of Philosophy (Algemeen Nederlands Tijdschrift voor Wijsbegeerte)* 96:114–25 (in Dutch). [JP]

(forthcoming) Shaping your own life. *Metaphilosophy*. [JP]

Perner, J. (1991) *Understanding the representational mind*. MIT Press. [KES]

Piliavin, J. A., Callero, P. L. & Evans, D. E. (1982) Addiction to altruism? Opponent-process theory and habitual blood donation. *Journal of Personality and Social Psychology* 43:1200–13. [aGA]

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

Rachlin, H. (1995) Self-control: Beyond commitment. *Behavioral and Brain Sciences* 18:109–59. [aGA, LG]

(2000) *The science of self-control*. Harvard University Press. [LG, HR]

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

Raineri, A. & Rachlin, H. (1993) The effect of temporal constraints on the value of money and other commodities. *Journal of Behavioral Decision Making* 6:77–94. [LG]

Ramsay, R. W. (1997) Behavioural approaches to bereavement. In: *The best of behavior research and therapy*, ed. S. Rachman & H. J. Eysenck. Pergamon. [aGA]

Reb, J. (2005) *The role of regret aversion in decision making*. Doctoral dissertation, University of Arizona. [TC]

Reb, J. & Connolly, T. (2005) Determinants of regret intensity: A comparison of justifiability and normalcy accounts. Working paper, University of Arizona. [TC]

Rescorla, R. A. (1988) Pavlovian conditioning: It's not what you think it is. *American Psychologist* 43:151–60. [arGA]

Rhue, J. W. & Lynn, S. J. (1987) Fantasy proneness: The ability to hallucinate "as real as real." *British Journal of Experimental and Clinical Hypnosis* 4:173–80. [aGA]

Richard, R., van der Pligt, J. & de Vries, N. K. (1996) Anticipated regret and time perspective: Changing sexual risk-taking behavior. *Journal of Behavioral Decision making* 9:185–99. [TC]

Roberts, R. D. (1991) Myopic discounting: Empirical evidence – Comment. In: *Handbook of behavioral economics, vol. 2B*, ed. S. Kaish & B. Gilad. pp. 342–45. JAI Press. [GWH]

Ross, D. (2005) *Economic theory and cognitive science: Microexplanation*. MIT Press. [DR]

Rubenstein, A. (2003) "Economics and psychology"? The case of hyperbolic discounting. *International Economic Review* 44:1207–16. [rGA, HA-C]

Ryle, G. (1949/1984) *The concept of mind*. University of Chicago Press. [aGA]

Sappington, A. A. (1990) Recent psychological approaches to the free will versus determinism issue. *Psychological Bulletin* 108:19–29. [rGA]

Sartre, J.-P. (1939/1948) *The emotions: Sketch of a theory*, trans. B. Frechtman. Philosophical Library. [aGA]

Schelling, T. C. (1960) *The strategy of conflict*. Harvard University Press. [aGA]

Scitovsky, T. (1976) *The joyless economy: An inquiry into human satisfaction and consumer dissatisfaction*. Oxford University Press. [aGA]

Sen, A. K. (1977) Rational fools: A critique of the behavioral foundations of economic theory. *Philosophy and Public Affairs* 6:317–44. [aGA]

(2002) *Rationality and freedom*. Harvard University Press. [HA-C]

Shafir, E. & Tversky, A. (1992) Thinking through uncertainty: Nonconsequential reasoning and choice. *Cognitive Psychology* 24:449–74. [rGA]

Shizgal, P. & Conover, K. (1996) On the neural computation of utility. *Current Directions in Psychological Science* 5:37–43. [aGA]

Siegal, S. (1983) Classical conditioning, drug tolerance, and drug dependence. In: *Research advances in alcohol and drug problems, vol. 1*, ed. R. Smart, F. Glaser, Y. Israel, H. Kalant, R. Popham & W. Schmidt. Plenum. [aGA]

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

Simonson, I. (1992) The influence of anticipating regret and responsibility on purchase decisions. *Journal of Consumer Research* 19:105–18. [TC]

Simpson, C. A. & Vuchinich, R. E. (2000) Reliability of a measure of temporal discounting. *The Psychological Record* 50:3–16. [LG]

Skinner, B. F. (1953) *Science and human behavior*. Macmillan. [FS]

Sloman, S. A. (1996) The empirical case for two systems of reasoning. *Psychological Bulletin* 119:3–22. [KES]

Smith, A. (1759/1976) *The theory of moral sentiments*. Oxford University Press. [aGA]

Smith, A. (1776/1976) *An inquiry into the nature and causes of the wealth of nations*. Oxford University Press. [aGA]

Smith, V. (1992) Game theory and experimental economics: Beginnings and early influences. In: *Toward a history of game theory*, ed. E. R. Weintraub. Duke University Press. [rGA]

Solnick, J., Kannenberg, C., Eckerman, D. & Waller, M. (1980) An experimental analysis of impulsivity and impulse control in humans. *Learning and Motivation* 2:61–77. *Review*, 2:217–25. [aGA]

Stanovich, K. E. (1999) *Who is rational? Studies of individual differences in reasoning*. Erlbaum. [KES]

(2001) Reductionism in the study of intelligence. *Trends in Cognitive Sciences* 5:91–92. [KES]

(2004) *The robot's rebellion: Finding meaning in the age of Darwin*. University of Chicago Press. [KES]

Stephens, D. W. & Anderson, D. (2001) The adaptive value of preference for immediacy: When shortsighted rules have farsighted consequences. *Behavioral Ecology* 12:330–39. [FS]

Stephens, D. W. & Krebs, J. R. (1986) *Foraging theory*. Princeton University Press. [XTW]

Strasberg, L. (1988) *A dream of passion: The development of the method*. Dutton. [aGA]

Strotz, R. H. (1955) Myopia and inconsistency in dynamic utility maximization. *Review of Economic Studies* 23:165–80. [HA-C]

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

Sunstein, C. R. (1995) Problems with rules. *California Law Review* 83:953–1030. [aGA]

Tanaka, S. C., Doya, K., Okada, G., Ueda, K. Okamoto, Y. & Yamawaki, S. (2004) Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. *Nature Neuroscience* 7(8):887–93. [rGA, DR]

Tomkins, S. S. (1978) Script theory: Differential magnification of affects. *Nebraska Symposium on Motivation* 26:201–36. [aGA]

Tversky, A. (1972) Elimination by aspects: A theory of choice. *Psychological Review* 79:281–99. [LG]

Tversky, A. & Kahneman, D. (1981) The framing of decisions and the psychology of choice. *Science* 211:453–58. [XTW]

Vermaij, G. (2004) *Nature: An economic history*. Princeton University Press. [DR]

Vuchinich, R. E. & Simpson, C. A. (1998) Hyperbolic temporal discounting in social drinkers and problem drinkers. *Experimental and Clinical Psychopharmacology* 6:292–305. [aGA]

Wegner, D. M. (1994) Ironic processes of mental control. *Psychological Review* 101:34–52. [aGA]

Winston, G. C. & Woodbury, R. G. (1991) Myopic discounting: Empirical evidence. In: *Handbook of behavioral economics, vol. 2B*, ed. S. Kaish & B. Gilad, pp. 325–42. JAI Press. [GWH]

Wogar, M. A., Bradshaw, C. M. & Szabadi, E. (1993) Effect of lesions of the ascending 5-hydroxytryptaminergic pathways on choice between delayed reinforcers. *Psychopharmacology* 111:239–43. [rGA]

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