Can Thought Experiments Prove Anything About the Will?

George Ainslie
Veterans Affairs Medical Center, Coatesville PA, USA
University of Cape Town, South Africa
George.Ainslie@va.gov

This material is the result of work supported with resources and the use of facilities at the Department of Veterans Affairs Medical Center, Coatesville, PA, USA. The opinions expressed are not those of the Department of Veterans Affairs of the US Government.

Published in
D. Spurrett, D. Ross, H. Kincaid and L. Stephens, Eds.,
Distributed Cognition and the Will:
Individual Volition and Social Context. MIT, 2007.

Abstract

Hyperbolic discounting predicts the strategic interaction of simple motivated processes to form more complex processes, ultimately the ego functions. Once the complex processes become internally recursive they elude study by controlled experiment. I advocate the use of a tool from the philosophy of mind, thought experiment, to test hypotheses about these complex processes. As an example, common assumptions about the nature of will are challenged by the findings of four thought experiments: Monterosso's problem, Kavka's problem, the conundrum of free will, and Newcomb's problem.

Text

Research in decision-making has always followed one of two basic strategies: Start with the complex and work toward the simple, or start with the simple and work toward the complex. Since real life choices are the complex part of the picture, most researchers have chosen the former strategy, particularly since the "cognitive revolution" made subjective processes a legitimate subject of scientific inquiry (Baars, 1986; Lachman et.al., 1979). These *top down* theorists have documented many subtle phenomena in detail, but have not even tried to find the basic mechanisms of these phenomena. Many of these theorists are fundamentally opposed to the attempt to reduce higher mental processes to simpler elements (Miller, 2003; Baars, 1986, e.g. quoting Maltzman, p. 109 and Mandler, p. 258). Moving in the opposite direction, the atomists *par excellence* have been the behaviorists. These *bottom up* theorists have isolated atoms of choice, but even slightly complex concatenations of these atoms have had a mechanical feel, and the combining principles that were observed in the laboratory often stretched credibility when applied to the important life incentives—the idea that emotions come from conditioned sensory experiences, for instance, or that human behavior is a straightforward response to environmental contingencies (e.g. Skinner, 1953). Recently the neuroeconomists have joined the bottom-up approach—You can't get much more elementary than single cortical neurons. They have begun to repeat the work of the behaviorists while monitoring CNS activity (Glimcher, 2005). However, complex interactions are still beyond the capacity of this research.

In recent years findings from parametric behavioral experiments with both humans and nonhuman animals have suggested how the simplest motivated process may combine into higher functions. The logic of this compounding process predicts the growth of even the ego functions from the interaction of elementary motives. Undoubtedly some of this growth is constrained by inborn predispositions, but the existence of these predispositions becomes an open question rather than the assumption that more holistic theories have needed. However, as this model predicts higher functions, the method of controlled experiment becomes inadequate to test many of its predictions, for a reason that I will go into presently. In this paper I will elaborate on an earlier suggestion of mine (Ainslie, 2001, pp. 125-139), that a tool of philosophy of mind—the thought experiment—can be a

sound way of testing the validity of theories of higher mental functions. The most developed example of both modeling and testing is the will.

Three Components of Will

The will has always been hard to study, to the point that some authors have claimed that it does not exist. Ryle famously called will "the ghost in the machine (1949/1984)," and a recent book by Wegner called it an illusion, at least in the form of which people are conscious (2002). Part of the problem is that the term "will" gets applied to at least three somewhat independent functions: the initiation of movement (which corresponds to the Cartesian connection of thought and action-- the function that Ryle found unnecessary), the ownership of actions, which gives you the sense that they come from your true self (the one that Wegner shows to be a psychological construction), and the maintenance of resolutions against shortsighted impulses.

The first two can be studied by experimental procedures and by observing their pathology in "nature's experiments." An intention to move can be tracked through electrical activity in cortical neurons, starting from a point before it reaches the stage of being an intention (as in Iacoboni's mirror neurons 1999, and Libet's early movement potentials, 1999). An act of intending can even be isolated experientially by the amputation of a limb, which often leaves the sense of initiating movement without the subsequent stages of movement. Ownership-- the integration of choice with the larger self-- can be studied in cases where it is absent: in splits of consciousness or activity below the threshold of consciousness. Splits remove the reporting self's "emotion" of agency by physically (split brain, alien hand) or motivationally (dissociation and probably hypnosis) blocking this part-self's awareness of the other will functions; however, even without the feeling of agency both of the other components of the will, initiation of movement and the steadfastness of resolutions may be preserved. Subthreshold phenomena include mannerisms, which can be shaped even in sleep, small drifts of activity that can be summed into ouija-like phenomena, and the insensible induction of one choice over another, e.g. by means of powerful transcranial magnetic fields (Brasil-Neto et.al., 1992; Wegner, 2002, reviews many of these phenomena).

The third function of will, the maintenance of resolution—*willpower* hereafter-- has been harder to study, and yet it is arguably the most important of the three. Pathologies of initiation are rare—the "locked-in" syndrome is the most dramatic example—and the pathologies of ownership just mentioned are uncommon. By contrast, as modern society progressively widens our freedom of choice, pathologies of willpower have become the most common preventable cause of death and disability in young and middle-aged people (Robins & Regier, 1992). These are not limited to failures of willpower, seen in addictions to alcohol, recreational drugs, cigarettes, food, gambling, credit card abuse, and many less obvious forms of preferring smaller, sooner (SS hereafter) satisfactions to substantially larger but later (LL) ones; in addition to these failures there are the overly narrow, rigid wills seen in obsessive compulsive personality disorder, anorexia nervosa, and many character pathologies both named and unnamed.

Data from neuroanatomy and neurophysiology have told us what parts of the brain are essential for willpower. Ablation studies dating back to the now famous Phineas Gage (lobotomy by tamping bar) have shown that without the functioning of the lateral prefrontal cortex people cannot carry out intentions over time, becoming notably impulsive (Burgess, 2006). A recent fMRI study by McClure and his collaborators showed that when student subjects chose LL rewards over SS ones their lateral prefrontal cortices and parts of their parietal cortices were more active than when they chose the other way (2004). Monterosso and his collaborators have shown that when deprived smokers are told to avoid puffing on a device that delivers cigarette smoke, their dorsal anterior cingulate gyri and supplementary motor areas are active (2006). However, such studies show only where, not how or why, willpower works. It is certainly foreseeable that greatly increased resolution in space and time will let brain imaging answer those questions, too; but this may not happen soon.

According to the standard of normality that is either explicit or implicit throughout the behavioral sciences, rational choice theory (RCT), willpower should not be necessary at all. Choices are assumed to have inertia, so that, once made, they will be steady over time in the absence of new information (Hollis & Sugden, 1993). In economics and behavioral psychology normal individuals have been explicitly held to discount future outcomes in an exponential curve; in other fields exponential discounting is implied by RCT's property of consistency, since all shapes other than the exponential sometimes predict reversals of preference as a function of time. Given exponential discounting, the role of a self or ego is merely to obtain the individual's greatest advantage by seeking as much information and freedom of action as possible. In RCT the ego is a hierarch that coordinates obedient subordinate processes; will in the sense of willpower is superfluous, and impulsive choices must be explained by some separate motivational principle.

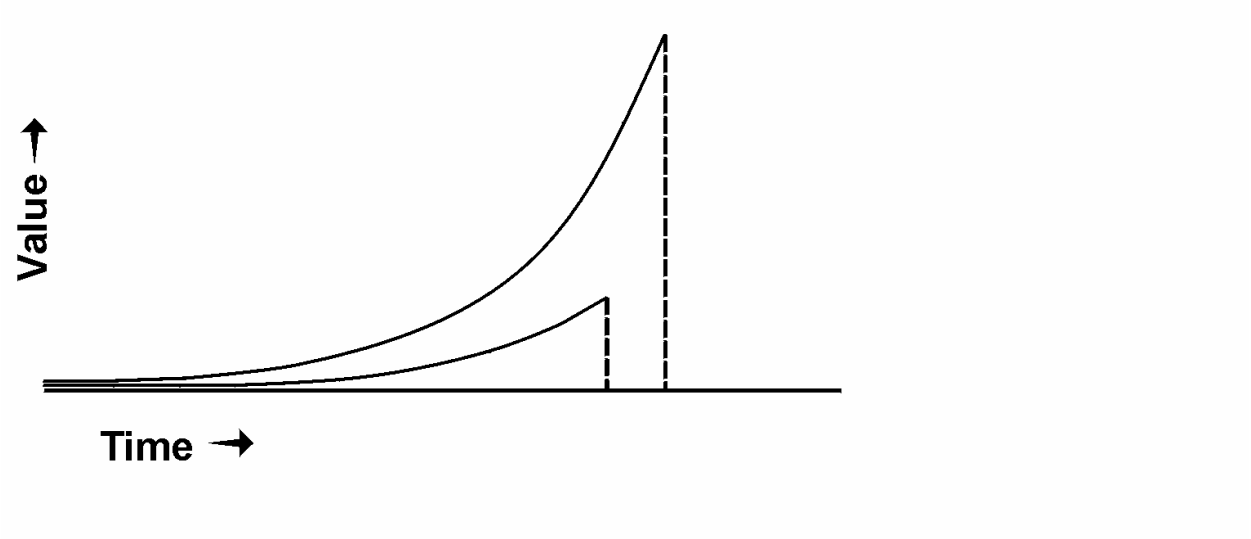
Basic Properties of Hyperbolic Discounting

However, parametric experiments on the devaluation (discounting) of prospective events in animal and human subjects have repeatedly found that an exponential shape does not describe spontaneous choice as well as a hyperbolic shape (inverse proportionality of value to delay-- reviews in Green & Myerson, 2004; Kirby, 1997; Figure 1). Three implications of hyperbolic discounting—preference reversal toward SS rewards as a function of time (*impulsiveness*), early choice of committing devices to forestall impulsiveness, and decreased impulsiveness when choices are made in whole series rather than singly (v.s.)—have also been found experimentally (reviewed in Ainslie, 1992, pp. 125-142 and 2001, pp. 73-78).

Such findings suggest an alternative to the hierarchical model of the self: Behavioral tendencies are selected and shaped by reward in a marketplace of all options that are substitutable for one another (developed at length in Ainslie, 1992, pp. 144-227 and 2001, pp. 73-104). Temporary preferences for SS options create conflict among successive motivational states, so that a currently dominant option must include means of forestalling any incompatible options that are likely to become dominant in the future. Neither better information nor greater freedom of action necessarily serves the person's

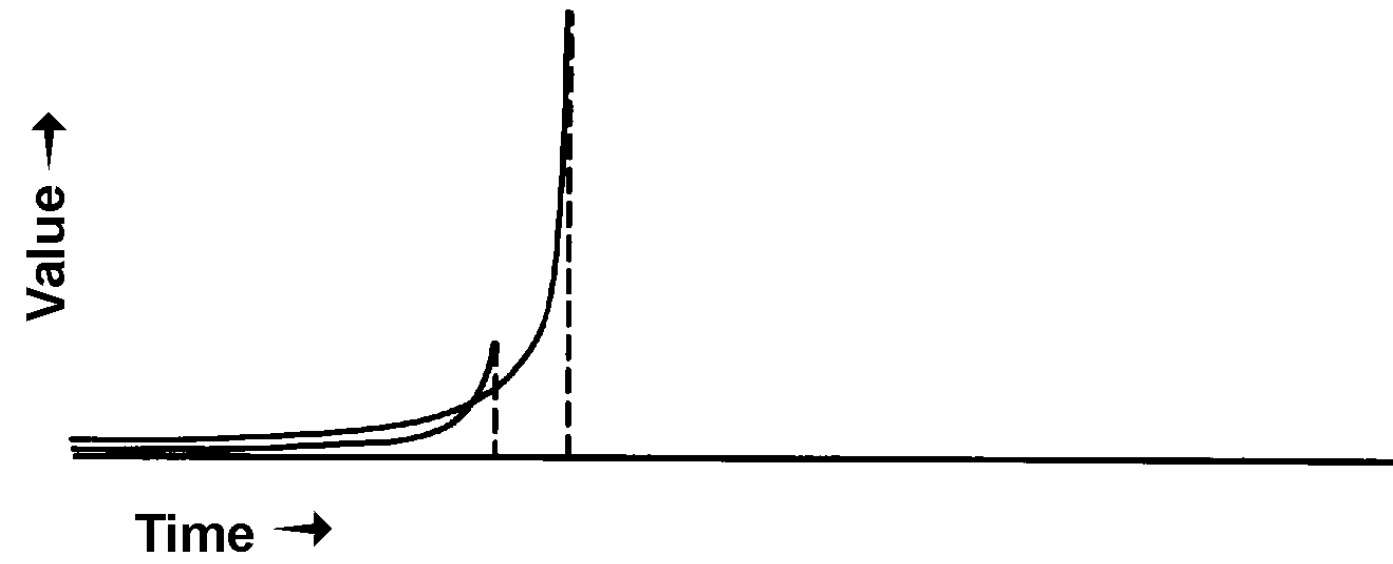
longest range interest. The key ego function becomes prediction and forestalling temporary preferences for SS rewards. This function need not involve a special organ, but rather can be expected to be learned by trial and error whenever an individual has adequate foresight. It may entail finding external incentives or constraints, diverting attention from sources of SS reward, or cultivating emotions with some motivational momentum; but the only method that is both powerful and widely applicable is willpower.

Figure 1A



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

Figure 1B



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

The finding that discounting the future tends to follow a hyperbolic curve puts even more demands on a theory of willpower. The will is needed not just for heroic cases, but as a basic means of knitting a person's intentions together from one moment to the next. Far from being a ghost in the machine, will may be the most fundamental ego function. Several theories of willpower have been proposed, or at least sketched out: that it is an organ-like faculty that gets stronger with practice but is exhausted by overuse (Baumeister & Heatherton, 1996; Kuhl, 1994); that it is a skill involving selectively not re-examining decisions once made (Bratman, 1999; McClennen, 1990); that it is an appreciation of molar patterns of choice that itself adds incentive not to spoil these patterns (Rachlin, 1995); or that it comes from the (usually tacit) perception of a limited-warfare relationship among successive motivational states, which makes current choices test cases for future choices in similar situations. Mine is the last of these, and it has the advantage of having been suggested by the shape of the hyperbolic curve itself. I outline its rationale here, but a rationale is not a proof.

Intertemporal Bargaining

Limited warfare occurs where agents share some goals but dispute others (Schelling, 1960, pp. 60-90): Warring countries share an interest in avoiding nuclear war; or a couple wants to save money but each member wants to exempt pet projects. Such situations set up a repeated game of prisoner's dilemma. Peace or at least a truce is achieved by establishing clear criteria for what will constitute defection in this game, in

such a way that each agent's long range interest is served better by continual cooperation than by *ad lib* defection. My hypothesis is that hyperbolic discount curves are continually putting us into intertemporal prisoner's dilemmas—cooperate with future selves for the long run vs. splurge for the present moment. The most powerful solution is simply to recognize this state of affairs, so that our current decision becomes a test case for how we can expect to decide similar choices generally. With such a perception our expected reward from having a consistent intention is staked on cooperating with our future selves, and is sharply reduced if we defect to an impulsive alternative. Although people conceive the mechanics of this contingency variously, under the rubrics of morality, principle, personal intention, and even divine help, we uniformly experience resolve when we have an adequate stake in a particular plan, and guilt or at least foreboding when a lapse causes loss of part of this stake. That is, the kind of guilt that arises from a failure of resolve represents your accurate perception that you have reduced the credibility of your promises in similar situations and perhaps generally, making intertemporal cooperation harder. The threat of this reduced credibility, I have argued, is the basis of willpower.

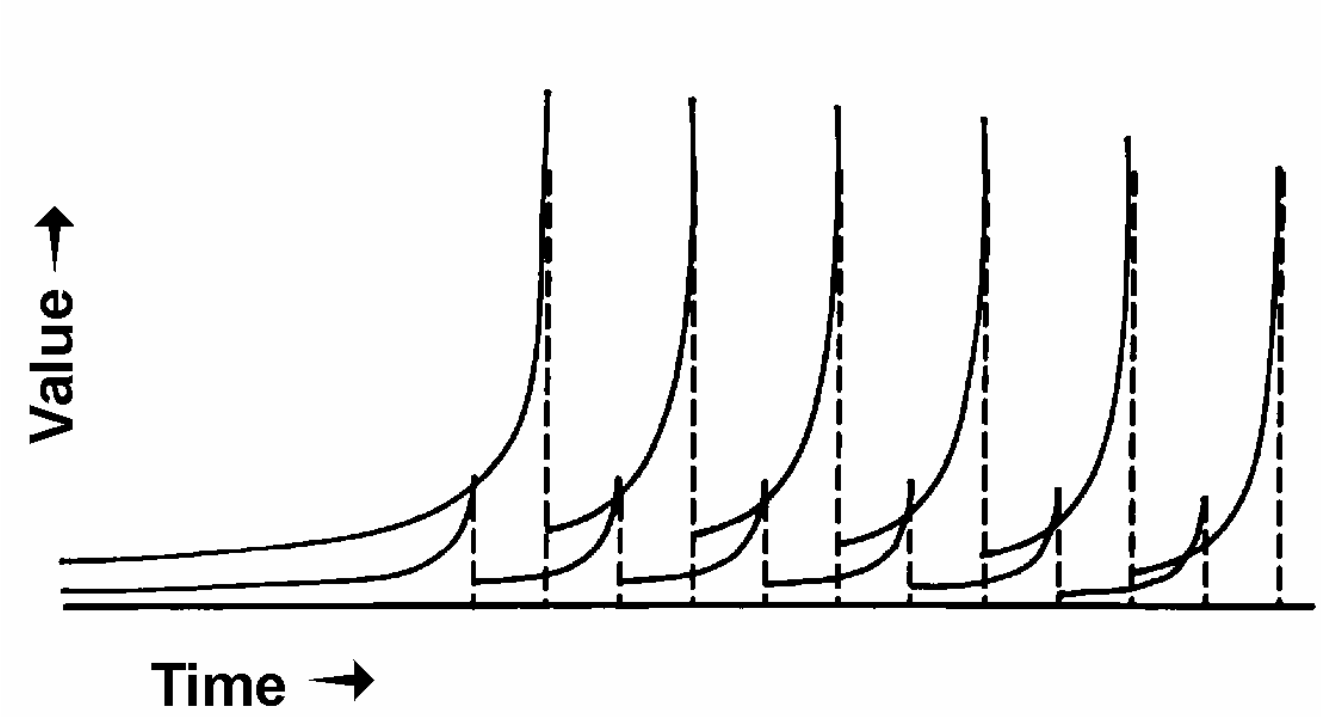
Empirically, this hypothesis has two components:

1. Making decisions between SS and LL rewards in a whole bundle increases the incentive to choose LL rewards; and
2. Perceiving current choices as test cases that predict a series of similar future choices forms these choices into such a bundle.

The Effect of Bundling Rewards

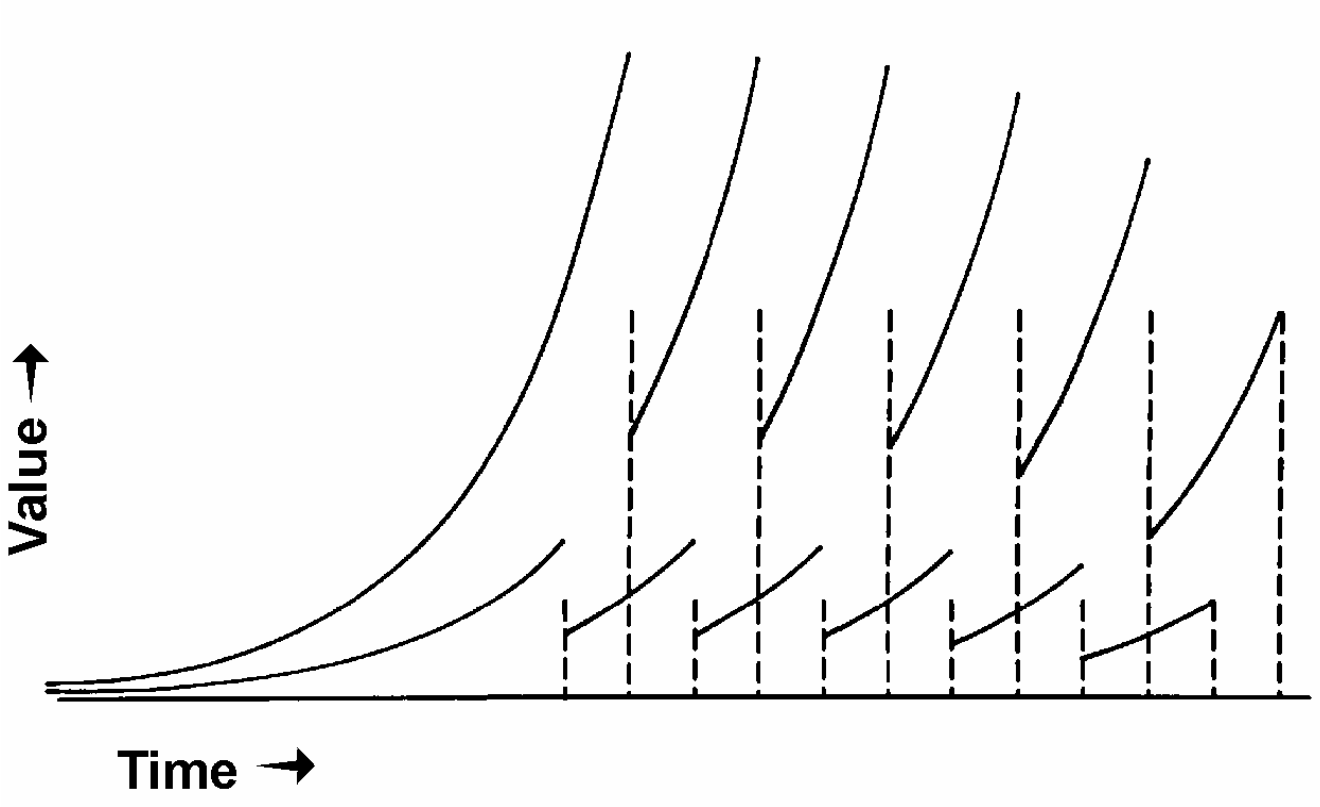
Both controlled experiments and uncontrolled observations have provided good evidence for the first hypothesis. Studies with nonhuman animals show that the hyperbolically discounted effects of each reward in a series simply add (analyzed in Mazur, 1997). Thus we should be able to estimate the effect of a series of rewards at a given moment by simply adding the heights of the curves from each reward at that moment. Because hyperbolic curves decay relatively slowly 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 (Figure 2A). Thus a bundle of LL rewards may be consistently worth more than a bundle of SS ones even where the discounted value of the most imminent smaller reward greatly exceeds the discounted value of its LL alternative.

Figure 2A



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

Figure 2B



Summed *exponential* curves from the same series as in figure 2A. 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.)

Experiments in both humans and rats have confirmed this effect. Kirby and Guastello reported that students who faced five weekly choices of a SS amount of money immediately or a LL amount one week later picked the LL amounts substantially more if they had to choose for all five weeks at once than if they chose individually each week (2001). The authors reported an even greater effect for SS vs. LL amounts of pizza. Ainslie and Monterosso reported that rats made more LL choices when they chose for a bundle three trials all at once than when they chose between the same SS vs. LL contingencies on each separate trial (2003). The effect of such bundling of choices is predicted by hyperbolic but not exponential curves: Exponentially discounted prospects do not change their relative values however many are summed together (Figure 2B); 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.

These experimental findings confirm a great deal of lore on willpower. From classical Greek times philosophers have claimed to see patterns of will and its failure, and these

reports have been consistent. The recommendation that has recurred most regularly through the ages has been to decide according to principle, that is, to decide in categories containing a number of choices rather than just the choice at hand (see Ainslie, 2001, pp. 117-120). Aristotle said that incontinence (*akrasia*) is the result of choosing according to "particulars" instead of "universals" (*Nichomachean Ethics* 1147a24-28); Kant said that the highest kind of decision-making involves making all choices as if they defined universal rules (the "categorical imperative," 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" (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.

Bundling via Recursive Self-Prediction

It is harder to find evidence about the second component of my hypothesis. This describes an internally recursive process, in which changes of current choice change expected future choice, and changes of expected future choice change current choice. There is no way to manipulate this process by varying external incentives; indeed that is one of the defining properties of the will, which the subject must experience as "free." Controlled experiments can only nibble at the edges of this process. For instance, the subjects in the Kirby & Guastello experiment who had to choose LL vs. SS rewards weekly for five weeks had a greater tendency to choose LL rewards if it was suggested to them that their current choice might be an indication of what they would choose in the future. This is suggestive, but hardly proof that the will depends on self-prediction.

Analogs using inter*personal* repeated prisoner's dilemmas can test whether the predicted dynamics of limited warfare actually occur. For instance, false feedback that a partner has defected changes a pattern of cooperation more and for longer than false feedback that a partner has cooperated changes a pattern of defection, an asymmetry that has been described for willpower (Monterosso *et.al.* 2002). Even the multiple person, single play games that most closely model intertemporal bargaining can be set up, as in this demonstration:

> Picture a lecture audience. I announce that I'll go along every row, starting at the front, and give each member a chance to say "cooperate" or "defect." Each time someone says "cooperate" I'll award a dime to her and to everyone else in the audience. Each time someone says "defect" I'll award a dollar only to her. And I ask that they play this game solely to maximize their individual total income, without worrying about friendship, politeness, the common good, etc. I say that I will stop at an unpredictable point after at least twenty players have played, at which time each member can collect her earnings. Like successive motivational states within a person, each successive player has a direct interest in the behavior of each subsequent player; and she'll guess their future choices somewhat by noticing the choices already made. Realizing that her move will be the most salient of these choices right after she's made it, she has an incentive to forego a sure dollar, but only if she thinks that this choice will be both necessary and

> sufficient to make later players do likewise. If previous players have been choosing dollars she's unlikely to estimate that her single cooperation will be enough to reverse the trend. However, if past choices have mostly been dimes, she has reason to worry that her defection might stop a trend that both she and subsequent players have an incentive to support.
>
> Knowing the other audience members' thoughts and characters-- whether they're greedy, or devious, for instance-- won't help a person choose, as long as she believes them to be playing to maximize their gains. This is so because the main determinant of their choices will be the pattern of previous members' play at the moment of these choices. Retaliation for a defection won't occur punitively-- a current player has no reason to reward or punish a player who won't play again-- but what amounts to retaliation will happen through the effect of this defection on subsequent players' estimations of their prospects and their consequent choices. So each player's choice of whether to cooperate or not is still strategic (Ainslie, 2001, p. 93).

This exercise has provided a serviceable illustration of how an intertemporal prisoner's dilemma works. However, the pattern of choices both in lecture audiences and roomfuls of volunteer subjects (residents of substance abuse rehabilitation programs) have been greatly affected by the participants' awareness of the social situation, as they have often said afterwards. The emotional incentives are apt to be greater than the monetary ones, and the precedent set by someone's move affects not only the moves of subsequent players but also her own self-image, e.g. as "a cooperative person (see Ainslie, 2005a)." Further research along this line is possible, but it promises to be both noisy and expensive.

Despite this suggestive evidence, a critic could still reasonably say that the second component of my hypothesis, recursive self-observation, has not been demonstrated. It might seem that with a process that is so much a part of daily life we could just ask people how it works. People can report using willpower among other ways to avoid impulses, and the pattern of reported endorsement of personal rules grossly correlates with expectable traits—positively with compulsive traits and negatively with reports endorsing external control devices (Ainslie, 1987). However, intensive interviews with a variety of subjects have shown that people notice disappointingly little about the properties of their own willpower. For instance, if subjects who are trying to break a bad habit are asked whether backsliding today will make it harder to avoid it tomorrow, they are almost as likely to say no as yes, which even to an intuitive layman standing outside the problem seems unlikely to be true (Ainslie, unpublished data). The common knowledge that must exist about how willpower works does not seem to be easily available to casual introspection. It is ironic that a process that is so familiar could be so elusive. There should be a way to marshal our everyday knowledge to test such a basic hypothesis, at least to rule it out or to rule out some of its alternatives.

Using Common Experience: Lessons from the Sociability Question

Most psychology before the positivist era drew its conclusions from common experience, just as the essayist or novelist did. There was little debate about will, but argument on

the basis of experience can be seen in many other topics, such as the mechanism of social feeling. An exchange between Alexander Bain and William James illustrates the failure of this method to get to the roots of theoretical disagreements: Bain argues that the value of human company is based on differential pleasure:

> Why should a more lively feeling grow up towards a fellow-being than towards a perennial fountain? It must be that there is a source of pleasure in the companionship of other sentient creatures, over and above the help afforded by them in obtaining the necessaries of life. To account for this, I can suggest nothing but the primary and independent pleasure of the animal embrace. For this pleasure every creature is disposed to pay something, even when it is only fraternal." (*Emotions and Will*, quoted in James, 1890, p. 551, note)

At which William James snorts,

> Prof. Bain does not explain why a satin cushion kept at about 98 degrees F. would not on the whole give us the pleasure in question more cheaply than our friends and babies do…The youth who feels ecstasy [when] the silken palm… of his idol touches him, would hardly feel it were he not hit hard by Cupid in advance. (1890, p. 552, note)

And again,

> With the manifestations of instinct and emotional expression, [pleasures and pains] have absolutely nothing to do. Who smiles for the pleasure of smiling, or frowns for the pleasure of the frown? Who blushes to escape the discomfort of not blushing?" (*ibid*, p. 550)

It is the authors' assumptions that limit their solutions. Bain assumes that all incentives must be like the rewards and punishments that come from concrete stimuli; hence he "can suggest nothing but" animal embrace as the hard currency that backs the otherwise flimsy experience of companionship. James assumes that only voluntary behaviors can be shaped by reward, thus ruling out any motivational basis for (involuntary) smiling, frowning, or blushing. These assumptions may have been useful at the time to protect the debate from unmanageable degrees of freedom, but they concealed possible solutions (such as the dependence of involuntary behaviors like blushing on intrinsic reward, which I advocate in Ainslie, 2001, pp. 48-70).

The positivist reaction to these experiential arguments took the form of behaviorism, which rendered many assumptions irrelevant but wound up limited by assumptions of its own—particularly the discipline of ignoring internal experience, which perversely evolved into the assumption that internal experience is not a factor in choice. Academic debate requires conventions of evidence just as organized sports must define boundaries for fair balls; but formalizing the scope of inquiry necessarily limits what it can find.

With the "cognitive revolution" researchers again began exploring common experience, but, naturally enough, no experiment has been devised to demonstrate that human companionship has a non-instrumental value, that is, a value beyond what it has as a means to other ends. You can't put experimental subjects into enough of a social vacuum for that. And yet the fact is obvious. Frustrated by the persistent convention in RCT that social transactions have to have instrumental bases, one author proposed a structured

introspection. Eric Posner asked us to imagine the case of the "pampered prisoner: A man lives in a giant building, all alone..." It is full of every resource except people. Posner argues that he lacks "autonomy:"

> An economist might say that the Pampered Prisoner is in an enviable position, but this claim is not plausible. We do not envy the Pampered Prisoner, and the reason is that much of our sense of accomplishment and well-being comes from our considered approval or rejection of the values to which others expect us to conform, and from our consistent action with these judgments despite the pressures put on us by others (Posner, 2000, pp. 207-208).

Faced with a counterintuitive assertion that can't be subjected to controlled experiment, Posner has tried to clarify intuition. He does not appeal to anyone's actual experience but to an experience that can be built from familiar components, with one element expunged—the instrumental need for other people. I won't argue whether or not this experiment is adequate to its purpose. I describe it to introduce the method of structuring intuitions to make them into shared observations. Shortly this illustration will also illustrate a hazard of that method.

Thought Experiments

What Posner proposed was a thought experiment, an exercise of imagination designed to let the reader examine her intuitive or previously unexamined knowledge in a framework that tests a hypothesis—in this case, whether instrumental purposes adequately explain the value of all social transactions. As in this example, thought experiments often suggest a counterfactual condition that removes one kind of doubt or variability in a real life choice situation, and ask how this condition affects what seems to be the best choice. The intuition sought does not constitute a theory, but a finding that can test a theory, just as the finding of a controlled experiment can.

Interestingly, thought experiments have been used almost exclusively in two disparate fields, physics and philosophy. It was Einstein's mentor, Ernst Mach, who first called them *gedankenexperimente* (Sorensen, 1992, p. 51), but they date back at least to Galileo. The broadest use of the term includes ways of visualizing logical truths, so that the proofs of plane geometry might count; but this usage is too inclusive to be meaningful. Most thought experiments in physics involve logical deductions from physical observations, albeit everyday ones. In this category would be Leibniz' argument against a famous dichotomy—Descartes' theory that a lighter body striking a heavier one recoils with equal speed, but that a heavier body striking a lighter one moves together with it. Leibniz imagined the continuous variation of the two masses from slightly unequal in one direction to slightly unequal in the other, so that at one point, in Descartes' theory, the striking ball would change suddenly from movement along with the other one to a movement just as fast in the other direction. Such a change would be logically possible, but unlike anything anyone had observed (*ibid*, p. 10).

Thus what I am calling thought experiments do not tap pure reason, but extend the reach of induction and hence the set of hypotheses that you can reject with a given degree of confidence. As Mach said, they permit you to draw from a storehouse of unarticulated experience (*ibid*, p. 4). They may sometimes be unnecessary for this—the laws of motion were soon subjected to physical experiments. Thought experiments in physics have been used mostly as heuristics, stepping-stones to quantified, controlled observations. However, in studying a mental process that is both inaccessible to direct observation and internally recursive, such verification is not available. In philosophy and, I will argue, in behavioral science, thought experiments may take the analysis of everyday experience beyond what controlled experimentation can provide, at least before brain imaging techniques become a great deal more sophisticated.

The simplest cases are hardly experiments at all. For instance, where someone has concluded from the common experience of internal dialogue that "we think in words," the crucial thought experiment might be just to recall occasions where you grasped a concept but "couldn't find the words." Thus when Hauser argued that "public languages are our languages of thought," Abbott proposed a "moron objection:" "Why is it so difficult to put my thoughts into English prose if they already are in English prose?" (Hauser & Abbott, 1995, p. 51). This test does not involve *a priori* logic. We can imagine mechanisms of thought that have only words as their units and that relate these units only by conventional syntax. However, it should not be necessary to survey a group of subjects about whether they have ever had the experience of not being able to find the right word. The moron objection seems to be a basic illustration of how to confront an intuitive hypothesis with an intuitive counterexample. It does not rule out the possibility that we sometimes think in words, nor does it support a particular alternative, such as "we think in a kind of machine language," but it clears the way for such alternatives.

Many of the experiments that have been conducted in social psychology since the cognitive revolution have actually been surveys of thought experiments. The violations of conventional rationality described by Kahneman, Tversky, and their school were demonstrated by what were essentially thought experiments, in which the authors asked a number of subjects to repeat what must originally have been the experimenters' personal intuitions (e.g. Kahneman & Tversky, 1984). For instance, most subjects said that if they lost money equivalent to the cost of a theater ticket they would still buy the ticket, but if they lost the ticket itself they would not buy another; or they would travel a certain distance to save five dollars on a calculator, but not to save five dollars on the combination of the calculator and a more expensive jacket. These framing effects became known through self-reports; the role of the experimenters was only to find questions that would elicit them—doubtless by their own introspection—ask these questions, and tabulate the data.

What makes it necessary to pose questions like these to a number of people, rather than just putting them before the reader? I would suggest that this surveying process is worth the effort in three circumstances:

1. where a substantial number of subjects might not report it, or where the reader might not be expected to believe that most subjects would report it. Ultimately

thought experiments, like logical arguments, convince insofar as the reader replicates them. However, reader and experimenter might share the elicited introspection but think it was weak enough or counterintuitive enough that other people might not share it. I have had hundreds of people say that they prefer $100 now to $200 in three years, but $200 in nine years over $100 in six years (or similar reversals of preference), even though this is the same choice seen at a greater distance; but some actual subjects have disagreed, and many reverse preferences at different values (see Ainslie & Haendel, 1983). Although the change in preference is widespread and might have been persuasive if presented as a thought experiment, it has taken actual experiments to be sure of it

2. where quantification would be useful. Parametric presentation of the amount-versus-delay question has been an important tool in establishing the specific shape of the spontaneous discount curve (Green & Myerson, 2004).
3. where the anticipated answers are illogical or otherwise noticeably "wrong" in some sense. Under those circumstances a reader might or might not be able to put aside her awareness of the "right" answer. The job of the experimenter is then to present the questions to naïve subjects, in such a way that the answer demanded by conventional assumptions is not apparent to them. This may not be difficult; I have always been surprised that subjects are willing to report changes of preference as a function of time.

By these criteria it would not be necessary to survey dozens of students on Abbott's moron objection. The finding of that thought experiment is just that people sometimes have a thought for which they can't find the words; it seems safe to assume that this experience is universal.

Thought Experiments on Willpower

There have been few thought experiments on self-control, let alone on willpower specifically. Some of those that have been proposed have just been heuristics rather than tests of hypotheses. They provide illustrations: "This is imaginable," rather than what is the only decisive outcome of testing a null hypothesis, "This cannot be true". An example is Quinn's self-torturer model of addiction (1993), which has been used to illustrate the possibility that addiction can be caused in exponential discounters by an intransitivity of choice between single and summed rewards (Andreou, 2005): Suppose that you are in a lifelong experiment, in which an implanted device can give you continuous electric shocks ranging from imperceptible (level 0) to excruciating (level 1000). Every week you can try out different levels, but are then given the choice only of whether or not to increase the shock by a single level, which is imperceptible but permanent, and collect ten thousand dollars each time you do. Quinn argues that you will prefer to increase the shock imperceptibly at each level, even though you would not accept level 1000 even for a thousand times ten thousand dollars. Thus he has created an analog of an addict, resembling an overeater, say, who is hungry enough (or emotionally needy enough) to gain an imperceptible 100gm every week, but would never choose to gain a massive 100 kg in twenty years. This model illustrates a mechanism of addiction, just as my lecture audience example illustrates a mechanism of intertemporal bargaining,

but it is not evidence either for the existence of this mechanism or against the existence of alternatives.[1]

By contrast with heuristic experiments like this, I am suggesting thought experiments that can rule out null hypotheses which are alternative to a target hypothesis, viz. the recursive self-prediction model of willpower. To do so each experiment must show that the null hypothesis cannot handle some intuition elicited by the experiment. As with heuristic thought experiments, and laboratory experiments for that matter, the design of the hypothesis-testing thought experiment must overcome two common shortcomings: It must accurately sample the process to be studied, and it must not confound its findings with interpretations of these findings.

Daniel Dennett has shown that many thought experiments intended to rule out the possibility of consciousness, autonomy, and similarly subtle qualities in a strictly deterministic world persuade only by tacitly assuming a lack of subtlety in the world's design (e.g. 1984). For instance, he comments on Ayer's attempt to demonstrate with an imaginary world that feelings of responsibility and blame would not withstand awareness that it was "possible to implant desires and beliefs and traits of character in human beings, [so] that it could be deduced… how any person who had been treated in this way would most probably behave in a given situation." Dennett complains that "those are circumstances in which… the actual causation of the imagined beliefs and desires has been rendered drastically cruder," that the imagined conditioning process has to be "powerful enough to override [the person's] delicate, information-gathering, information-assessing, normal ways of coming to have beliefs and desires (*ibid,* pp. 33-34)." In other words, the intuitions elicited by this thought experiment are not about people but about imaginary creatures. The result of such a sampling error is not a shared intuition but a shared convention of science fiction.

Posner's pampered prisoner example illustrates the hazards of claiming the authority of an intuitive finding for an interpretation of the finding. He has ruled out a need for practical cooperation, since the prisoner's environment has "every resource." But there is no comparison that demonstrates autonomy to be the missing element, as opposed, say, to intimacy or integrity or another Eriksonian quality, or another quality entirely. A diehard conditioning theorist might even assert Bain's idea (and Watson's, 1924) that we value people only for their association with sensory experience, and that we can't currently imagine how much we'd like being a pampered prisoner once we'd gotten used to it. The *finding* of the experiment is just our empathic sense that the pampered prisoner would be lonely. Explanation of this finding requires interpretation-- which, of course, is also true of the most positivistic controlled experiment.

---

[1] In the terms of the example, you know the contingencies in advance, so it is not an illustration of insidious onset—Herrnstein & Prelec's "primrose path" (1992). I think it would take a survey to see whether many knowing subjects would prefer 1/1000 of a fortune to avoidance of 1/1000 of agony if they preferred avoidance of the total agony to the total fortune. I personally doubt if intransitivity alone is an adequate explanation of preference for the weekly addictive increments, without either ignorance of the contingencies or the disproportionate weighting of imminent reward predicted by hyperbolic discounting.

In Dennett's view the unjustified simplification of intuition-based thought experiments, "intuition pumps," allows them "to entrain a family of imaginative reflections in the reader that ultimately yields not a formal conclusion but a dictate of 'intuition'" (1984, p. 12). Thus he may mean that all procedures intended to elucidate intuition are misleading, are pumps. But it is also possible that the problem is not in the fundamental unreliability of intuition, but in the lack of rigor that he himself has uncovered in identifying assumptions, in both their design and interpretation. Assumptions in design prevent these thought experiments from presenting a true sample of the phenomenon under study. Assumptions in interpretation confound what may have been a valuable intuition with additional, untested elements. If thought experiments are ever to serve as means of empirical observation, their degree of deviation from the experiences that presumably formed the reader's intuitions should be limited and clear, the null hypothesis should be specific, and the interpretation of the finding separated from the finding itself. I have selected and interpreted the following thought experiments with these principles in mind.

One more introductory idea: The most useful thought experiments test not just hypotheses but assumptions underlying hypotheses. The greatest obstacles to scientific progress have not been questions badly answered, but questions not asked. Assumptions have always constricted the asking process—for instance, that something solid must hold the planets up and must be both spherical to prevent snags and transparent to show the stars; that there is no way for blood to move between arteries and veins; or that organic matter has a fundamentally different quality from inorganic. These assumptions have plenty of equivalents in modern behavioral science. Elsewhere I have criticized the assumptions (among others) that choices have inertia—i.e. stay constant until acted upon by new information; that aversiveness is the opposite of rewardingness; that involuntary behaviors can't be based on motivation; and that "the" will is an inscrutable organ governed by a higher choice-making executive (reviewed in Ainslie, 2001 and 2005b). This last assumption is especially impenetrable and has elements of all three of the historical assumptions I just mentioned: that the will must be held together by something more solid than sheer incentive, that it can't be dissected and therefore must remain a black box, and that it has a nature which, if not immaterial, is at least outside the chain of causation as otherwise understood. The will is the last seat of vitalism. Thought experiments that can help penetrate this black box will be especially valuable.

We can now return to the second component of the intertemporal bargaining hypothesis, which challenges all the elements of this assumption about willpower. The source of willpower is not a force more powerful than motivation; willpower can be broken into steps, and it does not stand outside ordinary causality. Willpower is simply the perception of current choices as test cases for expectations of future choices. As I have said, controlled experimentation has been only suggestive; but four thought experiments—or three and a conundrum—call into question assumptions that contradict this hypothesis: Monterosso's problem, Kavka's problem, the free will conundrum, and Newcomb's problem.

Monterosso's problem. Monterosso's problem is the simplest:

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

Here imagination adds a single element to real life: certainty about the person's future smoking. The null hypothesis is that a present choice has no necessary motivational relationship with future choices. The finding is the strong conclusion that a person with certainty about her future smoking has no incentive not to smoke today, whichever the direction of the certainty—which is contrary to the null hypothesis. I interpret the finding to mean that self-control on the current day of a self-control plan seems necessary only to influence future choices. That is, if we imagine that our current choice will have no influence on future choices, we get a sense of release-- and from what, I would ask, if not from the felt need to avoid setting a bad example for yourself?

Kavka's problem. Unlike Monterosso's problem, Kavka's problem has disquieting implications: A person is offered a large sum of money just to intend to drink an overwhelmingly disgusting but harmless toxin. Once she has sincerely intended it, as verified by a hypothetical brain scan, she's free to collect the money and not actually drink the toxin (Kavka, 1983). Philosophical discussion has revolved around whether the person has any incentive to actually drink the toxin once she has the money—a majority of audience members initially say that she does not-- 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. Having had this pointed out, actual audiences tend to get raucous and search for ways that intention is still possible, usually not involving the need for a reassessment of whether to drink the toxin.

In the spirit of reducing unreal elements in thought experiments, it is desirable to replace the brain scan, even though such a capability is probably not far off. I have re-cast the problem in entirely realistic terms: Say you're a mediocre movie actor, and a director casts you, with some misgivings, to play a pipsqueak who gets sent down a terrifying toboggan run. You don't have to go down the run yourself-- the director is perfectly

happy to have one of the stunt men do it-- but you have to play a scene right beforehand in which you're frightened out of your wits. You realize that you can't fake the necessary emotion, but also that you are genuinely terrified of the toboggan run. The role is your big break, but if you can't do it convincingly the director will fire you.

Under these circumstances, you think it's worth signing up to do the run yourself in order to ace the preceding scene. But if, after playing this scene, you find out you can still chicken out of the toboggan run, is it rational to do so? And if so, won't your realization of that spoil your acting in the previous scene? Much the same discussion ensues as with the toxin, except that someone sometimes comes up with an answer like: If the anticipation scene had to be shot over, and you had chickened out of the run the first time, it would be hard for you to believe any intention to go through with it next time. This answer is a partial solution, but what if you knew that the scene was a wrap?

The null hypothesis of this thought experiment is that volition affects choices but is not affected by them in turn, so that your future ability to will difficult behaviors is not affected by any information about this ability that a current choice may convey to you. The findings are not the answers to Kavka's questions, which are usually garbled. Rather the two findings are (1) our lack of comfort with either answer, the sense that we may be doing the wrong thing by reneging, but that we can't put our finger on why—and (2) we can't find a clear reason why intention is possible at all in such a situation. I interpret this finding to show that there is a conceptual piece missing in the common theory of how people intend difficult behaviors. The null hypothesis is wrong. It is not possible to intend to drink or toboggan if you expect to renege, and conversely. Beyond this finding, I argue that hyperbolic discounting makes it possible to commit yourself, more or less, not to renege. You do this by putting up a pledge of sufficient value; and the only pledge available to put up irrevocably in this situation is the credibility of your pledges in difficult situations in the future. This kind of pledge is recursive: The more you believe that you will keep it the more you can keep it and the more you will subsequently believe; the less you believe you will keep it the less you can keep it, and so on. The current pledge need not put all future pledges at risk, but if you intend it to include only choices involving toboggan runs you will probably expect it to be inadequate from the start, and have to throw in more collateral, as it were, such as the credibility of your intentions to face fears in general, if you are to play that scene with conviction.

This description makes the process sound deliberate, but for most people it might be better described thus: If you notice that the toboggan choice is similar to other choices where you face major fears, you can expect to believe your intentions less in these cases if you intend and then renege in the present case. You might regard toboggan runs as a special case, of course, perhaps because of the uniqueness of the movie situation, but a future self might or might not exclude this example from her estimate of the credibility of a future intention.

Someone who objects to this interpretation needs to propose an alternative mechanism for how a person can expect not to renege in this kind of situation. Someone who needs experimental confirmation of the original finding will need to find naïve subjects: Once I

have suggested the above logic to an audience they soon accept it as the solution, and stop showing the tell-tale difficulties about how there can be intention when it is possible to renege. I take this development, too, to be confirmatory, but it is no longer a thought experiment.

Free will. The conundrum of free will must have the largest N of any puzzle on record, even if we count only publications on it; but it is essentially a thought experiment without a counterfactual story. The paradox exists in real life. In brief, we feel that our choices are not determined by the incentives we face unless we "let" them be so determined, and that the basis for letting them or not is imponderable, even to ourselves. On the other hand, random indeterminacy, such as the kind that has been supplied by atomic physics (Landé, 1961) doesn't feel like the basis for our choices either (Garson, 1995). Our experience of free will seems to contradict our belief in universal causality. The null hypothesis, if I may characterize this conundrum as having one, is that choices must either be caused by antecedent factors in linear fashion or not be strictly caused at all. As in Kavka's problem, the finding here is not any of the solutions that people try out, but the discomfort that the question causes. We depend both on our belief in physical causality and on our belief in the autonomy of our wills, and do not feel that we should have to choose between these beliefs.

I interpret this finding, too, as showing a missing piece in our concept of will. It would be impossible to try out all possible pieces to add, but the one suggested by hyperbolic discounting and its implied intertemporal bargaining seems to fit well with both sides. It preserves strict causal determinism, but it also makes individual choices imponderable, for the same reason that they can't be dissected by controlled experiments: They are recursive—internally fed back-- and sensitively dependent on small shifts in the bookkeeping of expectation.

The concepts of chaos theory have been suggested before as the root of experienced freedom of the will (Garson, 1995), but they have been rejected for the same reason as random indeterminacy: "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, chaotic 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)." I hypothesize that the location of recursiveness within the will itself supplies the missing sense of ownership. Although we can only guess at our future choices, the fact that these guesses change the incentives that govern those choices creates the compound contingencies that feel like will. By our vigilance about those choices we are actively participating in the choice process, all the more so because of our genuine suspense as to the outcome.

Newcomb's problem. Discussions of free will have often involved Newcomb's problem, one of the most irritating thought experiments to people who mistrust them (e.g. Rachlin, 2005), perhaps because it postulates a being who knows what you will choose before you do, something harder than usual to picture. Nevertheless it has had great

power to provoke debate, which suggests that it evokes some meaningful part of our choice-making process.

> A being in whose power to predict your choices correctly you have great confidence is going to predict your choice in the following situation. There are two boxes, B1 and B2. Box B1 contains $1,000; box B2 contains either $1,000,000 ($M) or nothing. You have a choice between two actions: (1) taking what is in both boxes; (2) taking only what is in the second box. Furthermore, you know, and the being knows you know, and so on, that if the being predicts you will take what is in both boxes, he does not put the $M in the second box; if the being predicts you will take only what is in the second box he does put the $M in the second box. First the being makes his prediction; then he puts the $M in the second box or not, according to his prediction; then you make your choice. Since the money is already in place before you make your choice, orthodox decision theory says that you should choose (1)-- both boxes-- and get $1000 more than if you chose (2), whatever the being had decided. The problem is that you believe the being to have anticipated this and to have left nothing in B2; since you feel perfectly able to choose B2 alone, and can believe the being to have anticipated this as well, you may increase your present expectation by choosing B2 (Nozick, 1993, p. 41).

The null hypothesis here is that we will maximize the reward that is literally at stake within the stated problem, and thus unambivalently prefer the two box option. The finding is our strong hunch that we should pick the single box, even though we can't affect the contents and two boxes will always contain $1000 more than one box. Shafir and Tversky took the trouble to present this problem to college student subjects, and, not surprisingly, elicited the same hunch (1992). These authors interpreted the finding as evidence of their subjects' magical thinking—their belief that what could objectively be only self-diagnostic thinking became causal thinking, in that by presenting the appearance of being a one-boxer they would cause the $M to be in that box.

It could be that we all have a magical streak, of course, but it seems more likely that we assimilate this unfamiliar situation to some familiar one. The Newcomb problem is an isolated choice with no history and no future and foreign to our experience. If this strange situation is to have any emotional meaning for us at all, we have to look for a similar framework in our experience. I offer an alternative to the magical thinking hypothesis: that we are intuitively familiar with a kind of *diagnostic* thinking that is also *causal,* and apply it to the Newcomb presentation

Everyone is familiar with problems that have almost exactly these terms, differing only in their repetitiveness and in the personal salience of the payoffs. Make the first box contain your aggregated expectation of reward for resisting your greatest temptation over time, the second your reward for giving in to it only this time. The being in whose predictiveness you have great confidence is yourself. 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 one- or 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 box B2 (empty for the two-boxer, full for the one boxer). 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, which are choice tendencies with momentum: Insofar as you have habitually chosen two boxes you *are* a two-boxer, a trait that governs your choices unless some factor happens to balance one of them closely. If you know you are a two-boxer you will be *unable* to motivate yourself to choose only B2—If you know you're a drunkard, there's no point in trying to have one sober day-- thus confirming the judgment of the predicting being. If you try to be a one-boxer and succeed, then you must *be* a one-boxer, also confirming the being's judgment. In the isolated example of the thought experiment, being a one-boxer is easy. In real life it takes the effort of resisting temptation. There are hard things you can do to overcome a two-boxer trait, and tempting things that will wreck a one-boxer trait, but Newcomb's problem elicits the motives contingent on recursive self-diagnosis without this ambiguous middle ground.

Newcomb intended his problem as a model of the Calvinist theory of sin and salvation, whose anti-impulsive effects have also been said to arise from magical thinking—seeking to believe you are saved when your behavior will have no effect on the actual fact (Weber, 1904/1958,p. 115). However, I have argued against the magical thinking interpretation here, too:

> Under such a belief system doing good is sometimes held to be a superstitious behavior, in that it is purely diagnostic, so that emitting it for the sake of seeing oneself emit it is fooling oneself (eg. Quattrone & Tversky, 1986). The several authors who have pointed this out do not consider an important possibility: Doing good for its diagnostic value may not invalidate this diagnostic value. That is, if one can do good for any reason it may be valid evidence of being among the saved; such a situation would not contradict predestination, but only provide another mechanism through which destiny might act. The expectation of salvation might be self-confirming, and the ability to maintain such an expectation might be the defining trait of the saved. This is much the same thing as saying that a higher power will grant sobriety to some alcoholics, and that acknowledgement of one's helplessness against alcohol is a sign that a person may be one of those who will receive this favor. Such a shift in a person's concept of causality is not casuistry. It marks the formation of a wider personal rule, one which apparently deters the hedging that weakens personal rules whenever they are recognized as such (Ainslie, 1992, pp. 203-204).

As with the prisoner's dilemma game itself, what is realistic play in the repeated game is not realistic in the one-shot version; but we rarely encounter the one-shot version. Our instinct is to follow the strategy of repeated play.

Conclusion

These four thought experiments produce little specific information about choice in the face of temptation—only one finding each, two in the case of Kavka's problem. Their value is that these findings are anomalous for both RCT, in which choice simply tracks expected reward, and for the conventional picture of will as an organ that can send orders to lower processes without having to recruit incentive from among these processes. Neither of these are *theories* of will; in fact our culture's assumptions about will are antithetical to theories of it: Based only on the difficulty of observing it, we have assumed that it has an organ like other life functions do, that this organ can't be subdivided into component processes, and that it is outside of the chain of causality that governs everything else. If will is to be a subject for behavioral science we must have a theory that is not disturbed by these four thought experiments; and to have such a theory we cannot be bound by these three assumptions.

I have proposed intertemporal bargaining theory, an expectable outgrowth of the now well-established phenomenon of hyperbolic discounting, as a basis for both the strength and freedom of the will. This model was not dictated by these thought experiments, but these experiments have revealed intuitive problems with theories based on the currently dominant assumptions, problems that do not arise with the intertemporal bargaining model. Given the intrinsic difficulty of studying intertemporal bargaining by controlled experiment, the very compatibility of this model with these findings is informative. As this bottom-up approach to motivation comes to address higher level mental processes, means of accessing common intuitions about them in a way that can test compatibility will be increasingly important. Suitably focused thought experiments look like a promising addition to the conventional methods of behavioral science.

References

Ainslie, George (1987) Self-reported tactics of impulse control. *The International Journal of the Addictions* 22(2), 167-179.

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

Ainslie, George (2001) Breakdown of Will. New York, Cambridge U.

Ainslie, George (2005a) You can't give permission to be a bastard: Empathy and self-signalling as uncontrollable independent variables in bargaining games. *Behavioral and Brain Sciences* 28(4).

Ainslie, George (2005b) Précis of *Breakdown of Will*. *Behavioral and Brain Sciences* 28(5).

Ainslie, George and Haendel, V. (1983) The motives of the will.in E. Gottheil, K. Druley, T. Skodola, H. Waxman (eds.),Etiology Aspects of Alcohol and Drug Abuse, Springfield,Ill.: Charles C. Thomas, pp. 119-140.

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

Andreou, Chrisoula (2005) Going from bad (or not so bad) to worse: On harmful addictions and habits. *American Philosophical Quarterly* 42, 323-331.

Baars, Bernard J. (1986) *The Cognitive Revolution in Psychology* Guilford.

Baumeister, Roy F. and Heatherton, Todd (1996) Self-regula-tion failure: An overview. *Psychological Inquiry* 7, 1-15.

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

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

Dennett, D.C. (1984) *Elbow Room: The Varieties of Free Will Worth Wanting*. Cambridge, Mass.: MIT.

Garson, James W. (1995) Chaos and free will. *Philosophical Psychology* 8, 365-374.

Glimcher (2005) Neural mechanisms of temporal discounting in humans. Paper presented at the *Third Annual Meeting of the Society for Neuroeconomics*, Kiowah Island, SC, September 16.

Green, Leonard and Myerson, Joel (2004) A discounting framework for choice with delayed and probabilistic rewards. *Psychological Bulletin* 130, 769-792.

Hauser, Larry and Abbott, Barbara (1995) Natural language and thought: Point and counterpoint. *Behavior and Philosophy* 23, 41-55.

Herrnstein, Richard J. and Prelec, Drazen (1992) A theory of addiction. In George Loewenstein and Jon Elster, eds.,*Choice Over Time*. Sage, 331-360.

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

Hollis, Martin and Sugden, Martin (1993) Rationality in action. *Mind* 102, 1-35.

Iacoboni, M., Woods, R. P., Brass, M., Bekkering, H., Mazziotta, J. C. & Rizzolatti, G. (1999) Cortical mechanisms of imitation. *Science* 286, 2526-2528.

James, W. (1890) Principles of Psychology, New York: Holt.

Kahneman, D. & Tversky, A. (1984) Choices, values, and frames. *American Psychologist* 39, 431-350.

Kane, R. (1989) Two kinds of incompatibilism. *Philosophy and Phenomenological Research* 50, 220-254.

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

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

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

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

Kuhl, Julius (1994) Motivation and volition. In G. d'Ydewalle, Bertelson, and Eelen, Eds., *International Perspectives on Psychological Science* vol.2. Hillsdale, NJ Erlbaum, 311-340.

Lachman, R., Lachman, J. L., and Butterfield, E. C. (1979) *Cognitive Psychology and Information Processing: An Introduction*. Erlbaum.

Landé, Alfred (1961) The case for indeterminism. in Hook, Sidney, Ed. Determinism and Freedom in the Age of ModernScience. New York, Collier.

Libet, Benjamin (1999) Do we have free will? *Journal of Consciousness Studies* 6, 47-57 (nos. 8-9 bound as *The Volitional Brain: Towards a Neuroscience of Free Will* Benjamin Libet, Anthony Freeman, and Keith Sutherland, Eds. Thorverton, UK: Imprint Academic).

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

McClennen, Edward F. (1990) *Rationality and Dynamic Choice*.New York: Cambridge.

McClure, Samuel M., Laibson, David I., Loewenstein, George, and Cohen, Jonathan D. (2004) The grasshopper and the ant: Separate neural systems value immediate and delayed monetary rewards. *Science* 306, 503-507.

Miller, William R. (2003) Comments on Ainslie and Monterosso. In Rudy Vuchinich and Nick Heather, Eds., *Choice, Behavioural Economics, and Addiction.* Pergamon, pp. 62-66.

Monterosso, John and Ainslie, George (1999) Beyond Discounting: Possible experimental models of impulse control. *Psychopharmacology* 146, 339-347.

Monterosso, John Robert, Ainslie, George, Toppi Mullen, Pamela, and Gault, Barbara (2002) The fragility of cooperation: A false feedback study of a sequential iterated prisoner's dilemma. *Journal of Economic Psychology* 23:4, 437-448.

??Monterosso, Mann, Ward, Ainslie, Xu, Brody, Engel, Cohen, London, (2006) Neural Activation During Smoking Self-Control: fMRI Assay [CPDD]

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

Posner, Eric A. (2000) *Law and Social Norms*. Harvard U.

Quinn, Warren (1993) The puzzle of the self-torturer. *Morality and Action*

Quattrone, G. and Tversky, A. (1986) Self-deception and the voters illusion. In J. Elster, (ed.)., *The Multiple Self.* Cambridge, U.K.: Cambridge University Press.

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

Rachlin, Howard (2005) Problems with internalization. *Behavioral and Brain Sciences* 28(5).

Robins, L.N. and Regier, D.A. (1990) *Psychiatric Disorders in America*. New York: Free Press.

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

Sappington, A.A. (1990) Recent psychological approaches tothe free will versus determinism issue. *Psychological Bulletin* 108, 19-29.

Schelling, Thomas C. (1960) *The Strategy of Conflict*. Cambridge, Mass: Harvard University Press.

Shafir, E. and A. Tversky, (1992). "Thinking through uncertainty: Nonconsequential reasoning and choice." *Cognitive Psychology*, 24, 449-474.

Skinner, B.F. (1953) Science and Human Behavior, New York:Free Press.

Sorensen, Roy A. (1992) *Thought Experiments*. New York,Oxford.

Sully, J. (1884) *Outlines of Psychology.* N.Y.: Appleton.

Watson, J. B. (1924) *Behaviorism.* NY: The Peoples Institute Publishing Co..

Weber, M. (1904/1958) *The Protestant Ethic and the Spirit of Capitalism.* New York: Charles Scribners Sons.

Wegner, Daniel M. (2002) *The Illusion of Conscious Will*. MIT.