'Free Will' as Recursive Self-Prediction: Does a Deterministic Mechanism Reduce Responsibility?

George Ainslie 11,09

To appear in Poland, Jeffrey and Graham, George, *Addiction and Responsibility*. MIT

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.

Advances in brain imaging have revealed more and more about the physical basis of motivated behavior, making the age-old dispute about free will and moral responsibility increasingly salient. Science seems to be delineating chains of causality for feelings, choices, and even beliefs; but if all mental life is strictly caused by prior events, and those by still earlier events in a chain extending back before birth, how can individuals be held responsible for their actions? Most people feel that they originate their actions (Nadelhoffer *et.al.*, 2005) and will readily give opinions about whether particular circumstances make an action blameworthy or not (Monterosso *et.al.,* 2005); but when philosophers take the chain of causality explicitly into account they generally distance themselves from these direct introspections, holding for instance that blame is just a way of assuaging instinctive resentment (Strawson, 1962/2003) or a threat to manipulate people's motives (Dennett, 1984, pp. 131-172). I come to this subject with a behavioral science rather than a philosophy background, but the free will dispute looks to this outsider like something that recent empirical findings might resolve.

The dispute is age old because of a fundamental conundrum. We insist on the truth of each of two propositions, the compatibility of which is far from obvious, perhaps absurd:

1. that all events are fully caused by pre-existing factors, and
2. that a person's choices are not always caused by pre-existing factors.

David Hume was already testifying to the historical scope of the conundrum a quarter-millenium ago:

> To proceed in this reconciling project with regard to the question of liberty and necessity-- the most contentious question of metaphysics, the most contentious science-- it will not require many words to prove, that all mankind have ever agreed in the doctrine of liberty as well as in that of necessity… (1748/1962, p. 104)

Hume thought the dispute was "purely verbal," and sought to clear it up with a flourish, but over the years the failure of countless re-wordings to produce a widely accepted reconciliation has revealed it to be substantive. There is something we have not understood, either in the operation of physical causality, or in the nature of human will.

After first looking briefly at the chain of causality, I develop here a proposal that the will is not only deterministic but mechanistic—the outgrowth of a specifiable interaction of simpler processes. However, I argue that this mechanistic will fits compatibilists' requirements for being free, which are close to the intuitive conception of freedom. This conception is not just the illusion of freedom, but the accurate

introspection of a discrete natural process. Finally, I argue that fitting this mechanism to the intuitive understanding of responsibility, in the sense of blameworthiness, requires an additional step-- a specific implication of the free will mechanism that reverses the conventional explanation.

A misunderstanding of causality?

The use of the word "indeterminacy" in atomic physics has led some authors to ascribe the conundrum to a false certainty about physical causality (e.g. Landé, 1961). From the discoveries in quantum mechanics in the 1920s, these authors have hoped for a way around strict determinism. However, it was soon pointed out that the random movement of particles translates into highly predictable averages at observable volumes—it matters little to the flow of a river that some water molecules are always moving upstream, for instance-- and is unlikely to produce a sense of efficacy at either micro or macro levels (see Smart, 1961). In reply there were suggestions, mostly by physical scientists, that made the brain into an analog of Schroedinger's famous (but hypothetical) Geiger counter, which could kill a cat or not on the basis of a single particle emission. Physiologist John Eccles proposed that a non-determined, spiritual self could interact with the brain by its effect on a single, strategically placed neuron (1994), and physicist Henry Stapp suggested that chaotic systems in the brain could be sensitively dependent on a few calcium ions determining neurotransmissions (1998). Such models fail to predict a sense of responsibility, as opposed to happenstance, unless the particles are endowed with mystical properties that supply the real responsibility, properties that would have to be miniaturized to an astounding degree. How many angels can dance on a subatomic particle?

Despite this problem, indeterminacy at the subatomic level continues to be put forward as a mechanism. When physicist Gerard t'Hooft recently constructed a model in which subatomic activity is deterministic, another physicist, Antoine Suarez, retorted

> "If t'Hooft is really correct then the work for which he is famed was not carried out as a result of his free will. Rather, he was destined to do it from the beginning of time… Maybe his Nobel prize should rightfully have been presented to the big bang instead" (Merali, 2007).

Physicists seem fascinated by this approach, and continue to look for the flaw in our understanding of physical causality. However, a failure to recognize atomic happenstance seems unlikely to have concealed a freedom of will that could be the basis for moral responsibility.

In early concepts that freedom was easy to accept. Hume, calling it "liberty," said it was "a power of acting or not according to the determinations of the will" (1748/1962, p. 104). It was just the ability to do what you wanted, leaving alone the question of whether what you wanted was subject to a chain of causality. By the time of William James that question had become unavoidable, and the answers had hardened. "The juice has ages ago been pressed out of the free-will controversy" (1884/1967). To keep moral responsibility, his main concern, from being crushed by determinism, he felt the need to postulate some wiggle room: "Indeterminism… says that the parts have a certain amount

of loose play on one another, so that the laying down of one of them does not necessarily determine what the others shall be" (*ibid.*). But attempts to specify what that loose play might consist of led inexorably to Eccles' microscopic indeterminism. An examination of our idea of human will seems more promising.

A misunderstanding of will?

There has always been antipathy to the notion of strict determinism because it seems to imply a loss of our humanity—of the subtle balance we experience in facing close choices, of our pride in feats of self-control, and our outrage at people who do harm through losses of self-control. Even stronger has been antipathy to one implication of determinism, that we are assemblages of component parts that function and combine according to material laws. The idea that our choices occur entirely through physical mechanisms dates back at least to Hobbes (1651/1996, ch. 6, pp. 37-46), and has caused repeated scandals through concrete interpretations of what it implies. In the Enlightenment it was carried furthest by la Mettrie, who declared that man was a machine, and that since the machine was driven by pleasure and pain, rational behavior consisted of finding ways to intensify pleasure (1748/1999), research that was undertaken by the Marquis deSade among others. At the time the threat to received religion seemed even greater than the risks of dissipation, and eighteenth century orthodox opinion lumped the new physiocratic philosophy together with pornography (Darnton, 1995, pp. 3-21). In the next century social Darwinists seized upon evolution to equate rationality with biologically driven dog-eat-dog competitiveness, which became synonymous with social irresponsibility. In the 1920s the behaviorism popularized by John B. Watson asserted that choice consisted of habits conditioned to external stimuli (1924, pp. 207-210), an opinion that attracted both admiration and revulsion before it was shown to be a misinterpretation of the experimental evidence (e.g. Rescorla, 1988). Suspicions about determinism are still founded especially on the fear that it is incompatible with morality.

Determinists have felt a need to defend their view largely by explaining how it can preserve personal responsibility, or, in the case of "hard" determinism, the *illusion* of personal responsibility (Smilansky, 2002; Strawson, 1962/2003). There have been many formulations of the properties needed for a subjectively free will. These properties seem to boil down to three: unpredictability, initiative, and moral responsibility (summarized in Haji, 2002). A brief look at discussions of these properties in philosophical literature reveals some useful suggestions, especially in the topic of initiative, but no basis for a synthesis.

Unpredictability. To be experienced as free a choice cannot be a foregone conclusion even for someone who knows all the preexisting incentives that bear on it, and even if it is the person herself who is trying to predict the choice. This condition is often expressed as, "I could have done otherwise." William James famously pointed out to his lecture audience that he could walk home either by Oxford Street or Divinity Avenue, and either choice was possible; thus looking back on whichever choice he made, he would not be able to say that the other choice was impossible (1884/1967). The possibility of alternatives continues to be put forward as a test for freedom of choice (*e.g.* Broad,

1962); but as a subjective experience, possibility of alternatives means no more than unpredictability of choice—"*for all I know* I can take/could have taken Oxford Street." James was using an example of near indifference, but a similar test for freedom can be applied to great motivational struggles: "For all I know, my effort to resist the urge to get drunk will succeed/might have succeeded.." Tomis Kapitan is unusual in characterizing unpredictability not as "could have done otherwise" but in the person's inability to consciously add up her motives (1986). This suggestion points us toward a valuable insight, as we will see.

Unpredictability by itself should feel the same as indeterminacy—as Dennett has pointed out, the pseudorandom numbers generated by a computer cannot be distinguished from the (?truly) random numbers generated by radium emissions (1984, pp. 144-152)—but this is not enough to create what feels like free will. Many authors have pointed out that simple unpredictability, even to the person herself, could not make choices seem free. Indeed, one argument against subatomic indeterminacy as a mechanism for freedom has been that the resulting choices would not feel *made* but just encountered (Strawson, 1986). The same problem would attach to any other model based entirely on unpredictability. However, any explanation of free will needs to account for the *introspective opacity* that keeps you from reading your impending choices from your perceived incentives. The quality of unpredictability is necessary but not sufficient for the experience of freedom.

Initiative. In a free choice you add something beyond identifying and aggregating the existing incentives. Some authors have proposed that the addition comes from the assignment of weights to motives (Nozick, 1981, pp. 294-309) or "a second-order capacity to reflect critically upon one's first-order preferences" (Dworkin, 1988, p. 108). Aside from the obvious problem of causal regression—what are your motives to reflect upon your motives?-- this mechanism seems to involve a setting up of motives in advance, not part of initiation in the moment. As Richard Holton has put it, "action is experienced as something that the agent instigates, rather than something that just happens to the agent as the result of the state that they were antecedently in" (Holton, 2009). However, it is unclear where such instigation would come from. What keeps a person from being simply a throughput that translates discerned incentives of the first, second, or nth degrees into actions?

One answer to this question is implicit in Cartesian dualism, which, although no longer viable as a theory, has a suggestive variant in Kant's proposal that the human faculty he called reason could cause phenomena in the physical world without itself being caused by them (1781/1968, pp. 369-384). The part of Kant's proposal that is of continuing interest is that he saw reason as structured by moral law, so its role in initiating actions would be to inhibit urges that violated a categorical imperative. For Kant, the important function of (and theoretical need for) this reasoned initiation would be what we would call self-control. Recent authors have also pointed out the close relationship of freedom and strength of will (Dennett, 1984, pp. 51-73; Holton, 2009;

Mele, 1995, pp. 32-143).[1] The relationship of freedom and strength will be important to our discussion presently.

James Garson has proposed a mechanism for the introspective opacity that makes choices unpredictable from a knowledge of their motives. He suggests that this opacity is the result of a chaotic choice-making process, in the technical sense—that it is modulated by ongoing feedback (1995). Certainly we are aware of cybernetic systems in ourselves, which let us follow a moving target with a finger or keep ourselves from getting too hot or cold. But these, too, feel like assigned tasks, not the sites of true initiative. Garson was not thinking of negative feedback, as in a thermostat—or exactly positive feedback, as in an atomic reaction, either—but offsetting feedback, in which one output could knock the next output onto a radically different course, an idea he got from chaos theory. In chaos final outcomes are *sensitively dependent* on small differences in input. However, this suggestion alone does not explain a sense of initiative. Why would choosing chaotically not feel like being swept about by the weather, or even overtaken by epileptic fits (Kane, 1989, p. 231)? As Sappington put it,

> 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 (1990).

Furthermore, conventional physiology and psychology do not suggest a process that would behave in this way—plenty of homeostatic processes, but nothing chaotic. However, this suggestion, too, supplies a root from which an adequate explanation can grow.

Moral responsibility. Compatibilist analyses of unpredictability and initiative have tried to spot these processes in the introspection of moments of choice—"What does free choice feel like?" Analyses of moral responsibility, by contrast, start with the necessity of preserving it, and look for justifications. These discussions are largely based on examples that distinguish intuitively between behaviors for which we would assign responsibility and behaviors we would excuse (Haji, 2002). Those authors who have tried to connect this approach with causality have generally concluded that blame is just another psychological phenomenon that arises from our natures. Peter Strawson, for instance, holds that moral "reactive attitudes" are ingrained and could not be given up on the basis of philosophical argument (1962/2003). Paul Russell does make a proposal like Kant's by pointing out that such attitudes can be modified, like any emotion, by the dictates of reason (Russell, 1992); however, it is not clear why such modification is caused any differently than the attitudes themselves.

Similarly, an empirical search for brain locations of moral reasoning has led to the conclusion that a sense of free will is merely one of the experiences our brains are wired to have. After a thorough and generally insightful examination of the problem in the light of modern neuroscience, psychologists Joshua Greene and Jonathan Cohen conclude that a rather despairing form of compatibilism is inevitable (2004): Hard determinism is true and "every decision is a thoroughly mechanical process" (p. 1781), but there are also "mechanisms that underlie our sense of free will" (*ibid*). Hardwired human "folk psychology" makes us ascribe responsibility to some agents, in such a way that "seeing something as an uncaused causer is a *necessary but not sufficient* condition for seeing

something as a moral agent" (p. 1782; their italics). As a result, "the problem of free will and determinism will never find an intuitively satisfying solution because it arises out of a conflict between two distinct cognitive subsystems that speak different cognitive 'languages' and that may ultimately be incapable of negotiation" (p. 1783). We are left with a sorry dualism, not between spirit and flesh but between two differently programmed areas of brain, an obedient motivation-follower and a somewhat deluded introspector.

I propose that these three properties—unpredictability, initiative, and responsibility--have remained elusive because our understanding of human will has been inadequate:

-- that there is good reason within strict determinism that people cannot predict their own choices with certainty;
-- that there is a strictly determined mechanism by which people take genuine initiative in their choices; and
-- that ascription of personal responsibility for actions does not require denial of (or other inattention to) the strict determination of those actions.

My method is to abandon cultural assumptions of the self as a unitary governor, and of the will as this self's organ of selection. Rather, I present evidence that the self is a population of partially conflicting interests, and the will is a property that emerges from these conflicts. A sense of responsibility comes in turn from personal experience with this emergence.

Re-casting the choice-maker: Hyperbolic discounting

Freedom of choice has always been considered at single moments, without regard to how the person's expectations for her own choices at future moments bear upon her current choice. This has not seemed to be an important limitation, since people are conventionally assumed to expect that they will keep following their current choice in the absence of new information about its likely consequences. However, a great deal of research over the last forty years has shown that all reward-seeking organisms devalue delayed prospects according to a function that often leads them to change preferences as time elapses, without getting any new information. Humans are the only species hampered by this phenomenon, since, with minor exceptions (deWaal, 2007, pp. 184-187; Henderson, 2009), we are the only one that makes future plans, as opposed to just obeying the promptings of current instincts that have evolved to protect future welfare–hoarding, dam building, migrating, and so on. To some extent we can learn to compensate for this *dynamic inconsistency*-- *akrasia* is the popular term in philosophy--but the most powerful kind of compensation requires our choosing to be a different operation than has been conventionally assumed. I have developed this argument elsewhere (Ainslie, 2001, pp. 27-104; 2005), but will present the essentials.

Future prospects are usually thought of as devalued, or *discounted*, according to the only function that would make preference stable over time, the exponential curve. Exponential curves describe a prospect's loss of value by a constant percentage of the value it has at a given time, for every unit of time that it is delayed:

$$\text{Present value} = \text{Value}_0 \times \delta^{\text{Delay}}$$

where Value_0 = value if immediate and δ = (1 – discount rate).

However, controlled experiments with both humans choosing spontaneously and nonhumans have shown that delayed options are discounted in a fairly simple inverse proportion of their expected delay. The formula was given its most-cited form by Mazur (1987):

$$\text{Present value} = \text{Value}_0 / [1 + (k \times \text{Delay})]$$

where Value_0 = value if immediate and k is degree of impatience. This function predicts that for some cases where smaller rewards precede larger alternatives subjects will prefer the larger, later (LL) reward when both are distant, but change to preferring the smaller, sooner (SS) reward as time elapses. Inverse proportionality also describes how tall a building appears on your retina as you walk toward it, so a shorter, closer building may loom larger than a taller, more distant one as you get close to it. We are used to ignoring this effect without a second thought when it occurs on our retinas. However, if it occurs in the centers where we evaluate prospective rewards, it may have a direct influence on our motives, of the kind Homer described for the Sirens on Ulysses. We cannot reason away distortions arising from the hyperbolic function, but must deal with them strategically.

It might seem that an evaluation process that regularly led to preference reversals as a function of time would have been selected against in evolution. However, this process is only one aspect of a general perceptual organization, described by the Weber-Fechner law (Gibbon, 1977), in which changes in any psychophysical quantity are perceived in proportion to the baseline quantity—that is, hyperbolically. It is problematic only in humans—animals that do not plan are best off trying to gratify their instinctive urges as quickly as possible—and the few hundred thousand years of human evolution may not have been enough to change such basic apparatus. Evolution may be seen instead in compensatory processes such as the larger prefrontal cortices, which seem to be crucial for the process of self-control, in *Homo sapiens* than in *Homo heidelbergensis* (DuBreuil, 2009).

Whatever the evolutionary rationale, hyperbolic discount functions have been well demonstrated. 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 accurately as a hyperbolic shape (reviews in Green & Myerson, 2004; Kirby, 1997; Mazur 2001). Three implications of hyperbolic discounting have also been found experimentally—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 (reviewed in Ainslie, 1992, pp. 125-142 and 2001, pp. 73-78).

These findings suggest a model for the choice-making process: Mental processes are learned to the extent that they are rewarded. Hyperbolic discount curves predict that mental processes based on incompatible rewards available at different delays do not simply win or lose acceptance, but interact over time. Processes that are congenial to each other cohere into the same process. Contradictory processes treat each other as strategic enemies. Ineffective ones cease to compete at all. Thus hyperbolically discounted reward creates what is in effect a population of reward-seeking processes that group themselves loosely into *interests* on the basis of common goals, just as economic interests arise in market economies (compare the "constituencies" that vie to elect governments in a self that has been analogized to a democracy—Humphrey & Dennett, 1989/2002). The choice-making self will have many of the properties of an economic marketplace, with a scarce resource—access to the individual's limited channel of behavior—bid for with a common currency—reward.[2] The competition of interests creates regularities within the internal marketplace, including support for those farsighted processes that can forestall or foster future behaviors whose rewards are not yet highly valued.

The finding that making choices in series increases patience is a key to how an autonomous self, one which initiates choices and is morally responsible for them, can arise from the interaction of elementary mechanisms. Standing alone, hyperbolic discounting depicts the opposite of such responsibility: A person at different moments is in a state of *limited warfare* with herself at other moments, sharing with them some long range goals but also motivated to shift resources away from these goals for current gratification. The self of one moment is helpless against what future selves may momentarily prefer, and can influence their choices only by literal commitment (think of Ulysses) or negotiation. But with no overriding government to appeal to, what does the present self have with which to negotiate?

We can get a hint from the advice that philosophers and psychologists have given about the will over the centuries. Beginning with Aristotle they have discerned several attributes, most notably a basis in choosing according to principle rather than according to the particulars of the current circumstance. The power of abstract principle to reduce actual impulsiveness is problematic for an exponential discounting model, which depicts people as naturally consistent to begin with; but it is predicted by the hyperbolic discount function, given only two conditions: that the cumulative discounted value of a series of expected rewards is roughly additive (figure 1), and that a person's expectation of getting the whole series can be made contingent on her current choice without physical commitment. The additivity condition has been verified experimentally (Mazur, 2001; Kirby, 2006). So has its implication that subjects will show greater preference for LL over SS rewards when choosing a whole series at once than they do when choosing singly. This increase in patience has been found in students choosing between amounts of money, and of pizza (Kirby & Guastello, 2001) and in rats choosing amounts of sugar water (Ainslie & Monterosso, 2003). The replication of this finding in animals shows that the increase in patience comes from the properties of the elementary discount function, rather than depending on cultural suggestion or on other effects seen only in humans (e.g. the "magnitude effect-- Green *et.al.*, 2004).

Figure

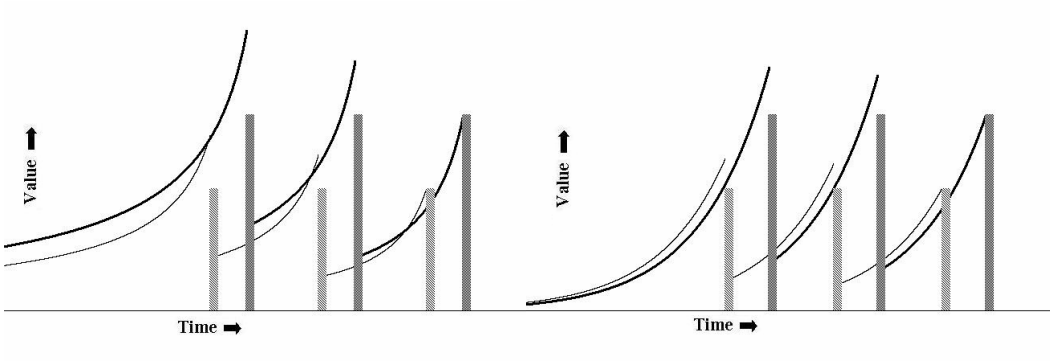

Figure 1a Figure 1b

Figure Captions

1A. The effect of bundling 3 pairs of larger, later (LL) rewards and smaller, sooner (SS) alternatives. Each *hyperbolic* curve shows the cumulated expected value of all similar rewards still to come. At the beginning of the series there is no temporary preference for the SS rewards. The curves from the last pair show what such a temporary preference would be if there were no bundle.

1B. Absence of a bundling effect with cumulative *exponential* curves, from the same series of alternative reward pairs as in figure 1A. Cumulation still increases the values of the prospective series over those of a single pair (*e.g.* the last pair), but the values of LL and SS rewards keep the same proportionality to each other.

The second condition— that mere self-prediction can create binding commitments to bundle series of choices together—does not lend itself to direct experimental test, since the subject's awareness of being in an experiment would itself create an exception to what she would require herself to do in everyday life. However, the dependence of large expectations on current test cases is a common intuition. The cost to a dieter of eating a piece of chocolate is clearly not a detectable gain in weight, but the loss of her expectation that she will stick to her diet. It is as if she were playing a variety of *repeated prisoner's dilemma* with her future selves, with a personal rule such as her diet serving as the criterion for which moves are cooperations (serving the common interest in slimness) and which moves are defections (abandoning the common interest for momentary pleasure). Uncontrolled observations of several kinds support this intuition: The lore on willpower mentions the disproportionate effect of a single lapse in reducing willpower (*e.g.* Bain, 1859/1886, p. 440), and disproportionate damage done by single defections

has been observed in interpersonal prisoner's dilemmas (Monterosso *et.al.*, 2002). Furthermore, when Kirby and Guastello added a condition to the repeated choice experiment just mentioned, in which they suggested to freely choosing subjects that their current choice predicted future ones, the subjects increased their preference for LL alternatives when choosing between single pairs (2001). Even more significant is the finding that when smokers and nonsmokers are run in a design similar to that of Kirby and Guastello, the smokers' initial preferences are much less patient than the nonsmokers', but they increase their patience in both the forced bundling and the suggested bundling conditions (Hofmeyr *et.al.*, under review). The nonsmokers do not further increase their patience. It seems that they are already avoiding *akrasia*, but the smokers are open to improvement from strategic methods.

Perhaps the best way to test the longstanding cultural intuition about choosing according to principle is to sharpen it by a device popular in the philosophy of mind, the thought experiment. I have argued that a small number of selected thought experiments yield a valid rejection of the null hypothesis—the hypothesis that volition affects choices but is not affected by them in turn (Ainslie, 2001, pp. 126-129, and 2007). Most illustrative is Gregory Kavka's problem (1983), here re-described so as not to rely on his magical brain scanner: You are 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 do not have to go down the run yourself-- the director is happy to have a stunt man do it-- but you have to play a big scene beforehand in which you are frightened out of your wits at the prospect. You realize that you cannot fake the necessary emotion, but also that you are genuinely terrified of the toboggan run. The role is your big break, but if you cannot do it convincingly the director will fire you. Under these circumstances, you think it is worth signing up to do the run yourself in order to ace the preceding fright scene. But if after playing this scene you can still chicken out of the toboggan run, is it rational to do so? And if you realize in advance that you will find this rational, will not this realization undermine your intention and thus spoil your acting in the fright scene?

Conventional utility theory says that it would be rational to chicken out, as do members of the lecture audiences who have been given this problem. Neither can then say how *intending to* go down the toboggan run would be possible when aware of this rationality. I interpret this finding as showing that there is a conceptual piece missing in the commonsense theory of how people intend difficult behaviors. The null hypothesis is wrong. It is not possible to intend to toboggan if you expect to renege, and it is rational not to renege so as to preserve the credibility of your intentions for future challenges.

How can you commit yourself not to renege? The relatively high tails of hyperbolic discount curves make it possible (see figure 1). 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 major fears in general, if you are to play that scene with conviction. You expect to follow through with the toboggan run to the extent you care that the credibility of future resolutions is at stake. Whether or not you actually go, the substantive impact of perceiving your choices in bundles is clear: Your present choice affects and is affected by the choices you expect to make in the future. This *recursive self-prediction* fulfills the second condition—binding commitment from self-prediction alone-- of our hypothesis about self-control without a central organ of will.

Now we can extend our model of choice-making. Maintenance and change of choice will be governed by intertemporal bargaining, the activity in which reward-seeking processes that share some goals (e.g. long term sobriety) but not others (the pleasure of having some drinks now) maximize their individual expected rewards, discounted hyperbolically to the current moment. This limited warfare relationship is familiar in interpersonal situations (Schelling, 1960, pp. 21-80), where it often gives rise to "self-enforcing contracts" (Klein & Leffler, 1981) such as nations' avoidance of using a nuclear weapon lest nuclear warfare become general. In interpersonal bargaining, stability is achieved in the absence of an overarching government by the parties' recognition of repeated prisoner's dilemma incentives. In intertemporal bargaining personal rules arise through a similar recognition by an individual in successive motivational states, with the difference that in a future state she is not motivated to retaliate, as it were, against herself in the past states where she has defected. In the intertemporal case, her risk of a loss of confidence during future states in the success of her personal rule, and her consequent defection in favor of short term interests during those states, will present the same threat as the risk of actual retaliation. These contingencies can create a will without an organ, serving a self without a seat, just as the "will" of nations not to use nuclear weapons seems to be guided by an invisible hand.

It might seem incredible that intertemporal bargaining was not described in so many words a long time ago. However, at the interpersonal level negotiations ranging in importance from ordinary courtesy to whether wars will be escalated have long had the form of repeated prisoner's dilemmas, but the formal game was described only in 1950 (Poundstone, 1992), despite the fact that interpersonal prisoner's dilemma contingencies made sense in terms of conventional utility theory. Without the limited warfare relationship among successive selves that hyperbolic discount curves predict, there would be no reason to suppose that intertemporal prisoner's dilemmas would arise in the first place. As with interpersonal prisoner's dilemmas, intertemporal ones are apt to be perceived intuitively, without deliberation. If you notice that the toboggan choice is similar to other choices where you face major fears, you have the sense that you will lose something of larger importance if you intend and then renege. At various degrees of awareness, you evaluate current choices partly as test cases predicting bundles of those similar future choices, bundles that hyperbolic discounting predicts will increase your patience.

Diets and other resolutions are examples of consciously constructed personal rules, with clearly defined conditions as to what kinds of choice are members of the relevant bundle, and criteria for which choices are cooperations and which are defections. However, once you have discovered that your current choice gives you predictive information about your future choices, even choices that are not governed by actual resolutions are apt to be influenced by this information to a greater or lesser extent. This influence will be largely nameless, or be hidden in seemingly disparate processes with names like force of habit, being true to yourself, following your intuition, or even responding to external necessity. After repeated experiences with resolutions, an individual with foresight who notices the predictiveness of present choices should develop by experience alone processes that look very much like a will. She will not usually need explicit resolutions, much less a faculty supplied *ex machina* by an intrinsically unitary self.

At this point someone is apt to object that strength of will often feels more like the direction of attention, for instance avoidance of reconsidering a prior resolution (Bratman, 1999, pp. 58-90; McClennen, 1990, pp. 200-218). I agree that avoiding reconsideration is one tactic of impulse control, and is often the "effort of will" of which people are most aware. However, it is a limited tactic, apt to be unstable over time, "holding your breath." You can avoid considering a potential reward for only so long, especially when the activity of considering that reward offers some pleasure in its own right. The systematic direction of attention itself requires willpower, though perhaps less than would resistance to the index temptation itself. Willpower in my sense is William James' kind, in which "both alternatives are steadily held in view, and in the very act of murdering the vanquished possibility the chooser realizes how much in that instant he is making himself lose" (1890, p. 534). The amorous teenager who is advised to avoid intercourse by avoiding sexual thoughts or play needs to make some use of will to employ this tactic; if she wants to have sex play and still not have intercourse then will is her sole weapon, and accordingly must be stronger. Diversion of attention and the related control of emotion are ancillary tactics that are distinct from willpower itself (Ainslie, 1992, pp. 133-142).[3]

The experience of free will: Unpredictability and initiative

With intertemporal bargaining will can grow from the bottom up, through the selection by elementary motivations of increasingly sophisticated processes. In many depictions from Descartes onward the will has the appearance of a canoeist steering through rapids—using skill and foresight to ride forces much stronger than itself, but still something made of different stuff, a spirit, a homunculus. The intertemporal bargaining process can generate the canoeist from the stuff of the rapids, different in skill and foresight but subject to the same motivational forces, and in fact developed by those forces. It is when the canoeist learns to include her own future tendencies as part of the currents she must anticipate that a pattern recognizable as a self develops. There are many implications of this learning, but here I will focus just on the way that recursive self-prediction permits the leap from current to canoeist, that is, from strict causality to the experience of free will.

When the incentives for alternatives are closely balanced, small changes in the prospects for future cooperation swing the decision between cooperation and defection. In that case an expectation about the direction of the present choice will be a major factor in estimating future outcomes. But this estimate in turn affects the probability that the present choice will be in that direction. Such a recursive decision process is not tautological, but continuously fed back like the output of a transistor to its own input. Where the person's predictions about her propensity to make the choice in question are at all in doubt, this feedback process may play a bigger role in her decision than any pre-existing incentive, external or internal. For instance, a dieter faces a tempting food, guesses that she will be able to resist it, experiences this guess as an increase in the likelihood that she will reap the benefits of her diet, and thus has more prospective reward to stake against the temptation. Then she notices a possible excuse. Her guess that she will try the excuse and not get away with it—that is, that she will subsequently judge her choice to have been a lapse-- will reduce her expectation of a successful diet and thereby her stake against lapses. This fall may be so great as to make the expected values of eating this tempting food vs. trying to diet about equal, until some other consideration tips her self-prediction one way or the other.

Sometimes, of course, the alternative choices may not seem to be closely balanced at all. When recursive self-prediction puts at stake a major element of self-esteem a remarkable degree of leverage can result (discussed by Bodner & Prelec as "self-signaling"—2001). David Premack described the example of a father who put off picking up his children in the rain to get a pack of cigarettes, and, when he noticed what this meant about his character, gave up smoking on the spot (Premack, 1970, p. 115; quoted in Miller, 2003, p. 63). The sensitive dependence of strongly motivated decisions on interpretation of such small observations, or of thoughts without any new observation at all, lends itself to theories of an overarching ego (which was Miller's purpose in quoting this example); but it can be fully derived from intertemporal bargaining. The power of symbolic acts that so impressed Freud and his followers needs no more explanation than their salience to aggregate expectations of prospective reward. Self-signaling is not subtle conceptually, but it lets choice elude any prediction based only on the contingencies of reward, and insulates the person's decision from coercive contingencies such as addictive cravings and looming toboggan runs. Thus it can be argued to generate the experience of free will (also in Ainslie, 2001, pp. 129-134). Furthermore, such an explanation allows us to characterize free choices better than by saying that they are too close to predict. After all, many behaviors are quite predictable in practice and are still experienced as free. What becomes crucial is the person's belief that a given choice depends on this self-prediction process, in whatever way she has come to represent this process to herself.

I am thus proposing that freedom of will comes from the same chaotic mechanism that generates strength of will. I hypothesize that the sensitive dependence of choices on the perception of bundles of reward supplies the experience of unpredictability, and that the siting of recursiveness within the process of will itself supplies the missing sense of initiative. Although we can only guess at our future choices, the fact that these guesses change the incentives that govern those choices creates, I would argue the "self-forming

actions" that libertarian Robert Kane locates at the root of free will (1989). 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.

Returning to the pair of conflicting propositions that I listed at the beginning of this chapter, identification of an unsuspected assumption makes this conundrum solvable. We have commonly assumed the first proposition to mean: All events are always fully caused *in linear fashion* by pre-existing factors. This proposition is false; we have not understood something about the nature of human will. Choices by self-aware humans are subject to recursive self-prediction. The truth of the second proposition, that a person's choices are not always caused by pre-existing factors, depends on what we take "pre-existing" to mean. Choices are still completely predetermined; but I have argued that the process of recursive self-prediction removes the sting from determinism by providing both unpredictability and the experience of initiation. The ultimate causes pre-exist, but they have by no means completed their activity when they have entered the person's motivational process. Their dynamic interaction during intertemporal bargaining is what initiates choices. To demand more of initiation, such as being a true first cause, is to add a layer of cosmological fantasy to a perfectly adequate interpretation of the subjective event.

The experience of free will: Responsibility

So far we have a proposal about the unpredictability of choices and people's sense of initiating them. Up to this point it is still not clear whether people can be held morally responsible for them, given that this proposal preserves a line of strict causality from before birth. The question of responsibility is what gives the free will controversy its urgency. To analyze it we will need to move beyond the nature of human will to the nature of blame. Given determinism, the only justification for blame has seemed to be as a threat to provide an incentive for good behavior (*e.g.* Dennett, 1984, pp. 131-172). However, it has been objected that blame as a tool of deterrence does not capture the common understanding of the term, and might not even be practical once people saw it as cynical manipulation (Railton, 1984; Smilansky, 2000; Strawson, 1962/2003). We would prefer a model in which blame reflects the perception of an intrinsic deservingness in the person's conduct. However, that would seem to require that the conduct is freely chosen, that is, not predetermined.

There is a way out of this dilemma, too. The literature on responsibility is concerned with social blame; self-blame, if mentioned at all, is a subsidiary process. If your choices spring entirely from causes that existed before you were born, the common argument goes, you cannot help them. Society cannot hold you responsible. You cannot hold yourself responsible, either, but that implication has seemed relatively unimportant. Given conventional theories of motivation the subordination of self-blame makes sense, but I will argue that it is backwards. The way to understand social blame is to understand self-blame.

People are said to learn self-blame from their parents—first the learning of their rules and then an "internalization" of those rules by a process that is still a matter of debate, often said to be something like classical conditioning. Initially your motivation is to win praise or escape blame, but after you have internalized the rules your motivation is to achieve pride or escape guilt, with guilt said perhaps to be the conditioned expectation of blame that now occurs whether or not it is realistic. In this view internalization is somewhat magical, unrealistic, perhaps the product of social deception or self-deception.

Hyperbolic discounting, and the recursive process of deliberate choice that can be derived from it, suggest a different motivational picture. Parental authority is still the original source of behavioral control, but obedience serves two purposes: to make your choices fit your parents' wishes, and to protect you from temporary preferences that you yourself would later regret (including damage to whatever quantum of other people's welfare you value for its own sake). To the extent that you become aware of the latter function you have two prospects at stake in obeying rules—your expectation of avoiding external blame and your expectation of containing impulses that would be harmful in their own right. As you become able to escape the scrutiny of parents and others, the second kind of stake becomes a separate incentive that has to stand on its own if the "internalization" is to endure. In the intertemporal bargaining game between your present and future selves, you risk a loss in addition both to the larger, later reward that is literally available and to your reputation in society, but your sense of this additional risk may not have a name—an unaccountable reverence for some received wisdom, perhaps, or just a nagging intuition. When you catch yourself violating your diet, what do you call the cost? Functionally it is the credibility of your intentions, regarding this diet and to some extent other diets and perhaps even more general kinds of self-control, but you are apt to call it just guilt or chagrin or self-reproach. If you have found it helpful to maintain your original sense of being watched by your parents, you may have cultivated a sense of ongoing presence in the form of an ancestor or god or saint, or even a living other, who somehow knows what you did and has become hurt or aloof.[4] In any case the loss is genuine—resolutions without external sanctions are self-enforcing contracts that are maintained by the value of your reputation with yourself, and you have injured that reputation. It does not matter whether you have retained—internalized—the rules you observed others using, or conceived new ones. You have a practical motivational basis for self-blame.

Your awareness that the loss of credibility your lapse caused was foreordained would not mitigate it. What would mitigate it would be an interpretation that removes your act from having been a test case for your self-control. A list of permissible excuses is intuitively clear: You did not know what you were doing, you did not realize your diet forbade it, you had an overriding justification, you could not have done otherwise, and so on. When you say you "could not have done otherwise" you do not mean that your act was predetermined, but that it was constrained-- not subject to motivation, or, more controversially, subject to overwhelming motivation. The bottom line is whether or not your act tells you something about what you can expect from yourself in the future. Intertemporal bargaining is a practical tool for self-control, and an awareness of determinism would not make it cease to function. A Laplacean demon might know

whether or not your self-control is about to suffer a setback, but you do not know it; getting an estimate about it was one of the expected outcomes of your current choice, and thus one of the incentives for this choice. Your choice was based upon your imperfect self-prediction; and this is true whether or not there was a demon that knew for sure and whether or not you believed that such a demon could exist.

People do not all wield this tool well. Some are too ready to accept excuses for themselves, and hence suffer from a reduced expectation of actually doing what they intend. Some stretch the obvious criteria in the other direction and blame themselves for outcomes that were outside their control, such as failing a test that was beyond their abilities. Self-blame is not immune to wishful thinking, the rationale in the case of excessive blame being a hope that perhaps the unfortunate phenomenon really is subject to your intentions—"If failing this test is my fault, it means I can still believe I am smart enough to have passed it." However, there are intrinsic constraints on wishful thinking, and however much it distorts our ascription of responsibility, it does not change the nature of the process. Self-blame is a primitive, an intrinsic contingency of recursive self-prediction, and it operates without regard to the question of ultimate causality. If our choice gives us bad news about ourselves we unavoidably "kick ourselves" for it, that is, pay a hedonic cost over and above the cost of the choice itself. This is the *cogito* of personal responsibility. In a theory that holds the person to be intrinsically unitary this extra cost would be puzzling and possibly superfluous, but it makes perfect sense for a population of successive selves in a limited warfare relationship with one another.

Social blame. A crude case could be made for interpreting self-blame in a pure deterrence model: Long range interests would threaten the current short range interest with guilt, and impose guilt as a sanction if the threat failed. However, guilt arising from intertemporal bargaining is not literally an action but the perception of a loss that has already happened. Certainly the notion of someone taking revenge on a past self would be odd. The threat that faces a person in bargaining with her future selves is not retaliation but a prediction of poorer reward in the future as those selves simply try to maximize it. She does not *choose* to blame herself, but suffers from her awareness of blameworthiness. It might then make sense to ask whether our sense of social blameworthiness is modeled on our internal experience. Might social blame be vicarious self-blame, rather than self-blame be internalized social blame?

My proposal is that people perceive social blame as an empathic extension of their personal processes of self-blame: "I blame her because in her shoes I would blame myself." If we understand social blame as our perception of another's conduct through the filter of our own intertemporal bargaining situation, it appears not mainly as a contingent incentive bearing on the control of that person's behavior, but an event in its own right, a loss that has already happened. It would be hard to say what the loss consists of—of trust, of some sort of credit, perhaps of the extent of possible empathy—but my point is that this kind of blame is not imposed but discerned, and discerned by analogy with our personal responsibility for our own lapses.

The foregoing analysis suggests that the characterization of self-blame as an internalization of social blame is backwards. Admittedly blame has a major role as a social deterrent. The criminal justice system is only the procedural extreme of the informal accounts kept among small groups and families. But the view of blame simply as social manipulation is too Machiavellian. In practice the assignment of social blame appears to be quick and sure, a process more of "affective intuition than deliberative reasoning." (Greene & Haidt, 2002, p. 517). Despite academic theories about when a person is blameworthy, people tend to base judgments on a sense of deservingness. For instance, theoretical tests for legal insanity vary widely from state to state in America, from merely being impelled by a mental disease to not even knowing right from wrong, but the rate at which juries accept this defense varies little from state to state (Cirincione *et.al.*, 1995) or within a state when the charge to the jury is changed (McGreevy *et.al.*, 1991). Furthermore, however much people may endorse a need to deter wrongdoers, punishing someone just "so as to make an example of her" is regarded as unfair.

The difference between this proposal and the deterrence model is best illustrated by comparing the intertemporal prisoner's dilemma with an interpersonal one. Given a recent history of cooperation, perception of a partner's defection in an interpersonal prisoner's dilemma causes a loss of trust in the partner, which gives you an incentive to punish her. Perception of your own defection in an intertemporal prisoner's dilemma causes a loss of trust in your future selves, and this loss of trust *is* the punishment. In the interpersonal case the punishment is an action, deliberate and tactical. In the intertemporal case the punishment is an inescapable perception of lowered prospects. The idea of a social loss as the root of interpersonal blame accords better with the intuition of deservingness or a "debt to society" than do pure deterrence theories. In both personal and social cases the loss of trust can be repaired by credible evidence, either subsequent cooperations or a side-transaction involving penance (in religion, "atonement"), in which you assure your partners, or your future selves, that there will be no defections without a commensurate cost. In many situations the empathic extension model and the pure deterrence model will make the same predictions, but empathic extension also describes the sense of deservingness.

Our increasing recognition of how basic is our tendency to put ourselves into others' situations (Iacoboni & Dapretto, 2006) accords with this hypothesis. Given the truth of such an extension, the legitimacy of social blame is not threatened by determinism. Why shrink from blaming someone else for transgressions when you would blame yourself for them? However, this legitimacy is limited by the extent to which another person's self-control functions resemble our own. Clearly we would not want to see justice administered by someone on the far end of the autistic spectrum, who cannot vicariously model another's motives,[5] or by a sociopath, who is not moved by them. Both, if intelligent, could follow a contingency plan for deterrence, but they could not assess the quality of deservingness that goes beyond deterrence. We are suspicious of even ordinary folks' empathic abilities, and look for ways to correct for differences of experience and condition. The legal principle of trying the accused by a jury of peers is one example.

What to do with the drunken sailor. Still, different juries intuit differently, and have often appeared from the outside to have miscarried justice. Judgments by individuals are even more variable. Thus we search for objective criteria that could divide the blameworthy from the merely hapless, without our having to imagine ourselves in various different shoes. Unfortunately there seem to be no distinct lines in nature that match our personal tests for blame with any better than fair regularity. The most popular has been whether the behavior was caused (or, more weakly, influenced) by a disease. Alcoholism and other addictions have been shown to have a large hereditary component (Goodwin, 1986). To the extent that an inborn appetite for a substance varies among people, those at the high end of the scale might be said to have a disease. And yet addicts have often been observed to resist "irresistible" cravings (Heyman, 2009), and have changed their behavior when given even relatively small structured payments to do so (Stitzer & Petry, 2006). To have an itch is a disease. To scratch it is a choice. Almost everyone is especially tempted by one kind of impulse or another. But is an itch sometimes so intense or persistent that a person should not be held responsible for scratching it?

Cause by a disease has always been a gold-standard excuse. Subjects who are told of hypothetical misbehaviors report that they excuse the perpetrators much more often if the antecedents were physiological (e.g., low levels of a particular neurotransmitter) as opposed to experiential (e.g., severe parental abuse--Monterosso *et.al,* 2005). Subjects also feel freer to misbehave after hearing arguments for determinism (Holton, 2009; Vohs & Schooler, 2008), suggesting that they irrationally interpret determinism, which they fail to differentiate from fatalism, as a sort of universal disease. In the addictions the identification of changes in addicts' brains pushes the argument toward a disease model. However, as science is increasingly able to identify physical proximate causes for behaviors, it is becoming apparent that such causes are universal. A physical change does not necessarily imply that the person is coerced by a motive that used to be considered resistible. Often it is the motive that induces the physical change; decisions themselves can be seen happening in the brain by functional magnetic resonance imaging (fMRI—Daw *et.al.*, 2005; Glimcher *et.al.*, 2007). Identification of a physical basis for motives has always been crude as a test for responsibility, but by chance the level of our ability to observe them used to apportion our sympathy to an extent that felt about right. The recent rise in our observational ability has spoiled this test, leading to what Dennett calls "creeping exculpation" (1984, pp. 156-169). We can no longer pardon every behavior of which we can see physical roots. Nor does there seem to be any other independent indicator of blameworthiness.

What, then, should we make of the cases where our best understanding of an addict's state of mind leads to the conclusion that she is *unable* to resist temptation—that failing an epiphany she would not be able to recruit enough motivation for a sustained period of abstinence? The unusually great rewardingness for a particular person of a particular modality—drinking, gambling, buying-- might reasonably be called a disease, but the resourcelessness that follows her repeated defections in intertemporal bargaining is more like a budgetary crisis. When the addict cannot find enough credibility to stake against her temptations to consume, we might say that she was no longer responsible for

her choices-- but because of bankruptcy, not sickness. There is no natural test for whether such bankruptcy "exists" or not, nor even a test for when we should recognize it. Such a recognition would necessarily be culture-bound and would resist theoretical benchmarks, just as attempts to define legal insanity have done. And whereas the financially bankrupt are not able to discover the funds they need by a radical restructuring of their books, sudden regenerations of will sometimes happen as addicts reframe their choices (Heyman, 2009, pp. 44-64, and Premack's smoking example, above). The determination of a motivational bankruptcy in place of a disease is not apt to make much practical difference, and will not resolve the issue of blame any more than financial bankruptcies do. It will merely let us understand addiction's psychological damage within a marketplace model of decision-making.

A mechanism for free choice

Descartes was struck by how the developing laws of physics, particularly mechanics, applied to the human body, but the motivational process seemed to be a world apart. His solution was that physical activity is mechanistic but that mental activity is independent: "I have a clear and distinct idea of myself as a thinking, non-extended thing, and a clear and distinct idea of body as an extended and non-thinking thing" (quoted in Klein, 1970, p. 346). He may have had additional reasons for the distinction, but it looks on the surface as if he was responding to a simple difference in observational standpoint: He could see his body move in the world of things, but could observe his "self" only introspectively. Thus he could conceive his arm as a system of levers, but had no picture to which he could analogize his self. We still have not agreed upon a picture of the self, though I am suggesting one here, and thus we keep circling the intellectually forbidden dualism, as if hoping to find a permissible rationale for it.

Eccles' subatomic indeterminist model (1994) could be anchored easily in Descartes' pineal gland. Compatibilist models avoid specifying this kind of nexus, but still they shift awkwardly between the external viewpoint of objective science and the introspective description of moral responsibility (Haji, 2002). Incompatibilists scoff at these efforts, but they have accounted for the robust introspective experience of freedom only as an illusion (Smilansky, 2000), if perhaps an illusion anchored firmly in our evolved cognitive apparatus (Greene & Cohen, 2004). Strictly speaking, a self united by intertemporal bargaining is never an unmoved mover; and yet it is an emergent phenomenon. The experience of free will is not an empty perception, or illusion. If we are dominated by any illusion it is one that we also inherited from Descartes, that in contrast to the body, "the mind is entirely indivisible" (quoted in Klein, 1970, p. 346). A person who is a population of partially conflicting interests in a limited warfare relationship is continually engaged in negotiation. Her choice is based only on incentives, but these incentives include the effect she expects to have on her own future motivational states, an expectation that arises and shifts freshly as she tries out ways of framing her choice. Her assessment of these incentives, knowing that the resulting conclusions will be fed back as further incentives, or even knowing that they might be so fed back, should create both genuine surprise and an accurate sense of personal initiative. It will not create the sense of being an unmoved mover, whatever that would feel like.

In discussing moral responsibility most compatibilists shift the frame of reference from the chain of causality to the robustness of the concept in personal experience. This shift is necessary, but the resulting discussions arrive at either blame as social manipulation, consistent with strict determinism, or blame as something deserved, an intuition that goes beyond the manipulation hypothesis but is seemingly inconsistent with determinism. However, this intuition can be better grounded by a rationale for deservingness in self-blame, a phenomenon that has heretofore seemed secondary. An understanding of intertemporal bargaining reveals self-blame to be a loss of self-trust rather than a nonsensical retaliation against a former self. This loss is unaffected by the question of whether it is strictly determined by a chain of prior causes. Perceiving social blame as an empathic extension of self-blame likewise makes determinism cease to be relevant to its assessment.. The intertemporal bargaining hypothesis provides a deterministic model that permits, to paraphrase Daniel Dennett, all the aspects of free will worth having.

References

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

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 (2005) Précis of *Breakdown of Will*. *Behavioral and Brain Sciences* 28(5), 635-673.

Ainslie, George (2007) Can thought experiments prove anything about the will? In D. Spurrett, D. Ross, H. Kincaid and L. Stephens, Eds., *Distributed Cognition and the Will: Individual Volition and Social Context*. MIT.

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.

Bain, A. (1859/1886) The Emotions and the Will, New York: Appleton.

Benson, Paul (1994) Free agency and self-worth. *Journal of Philosophy* 91, 650-658.

Bodner, Ronit & Prelec, Drazen (2001) The diagnostic value of actions in a self-signaling model (in Isabelle Brocas & Juan D. Carillo, *Collected Essays in Psychology and Economics*, Oxford.

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

Broad, C.D. (1962) Determinism, indeterminism and libertarianism. in S. Morgenbesser and J. Walsh (eds.),Free Will, Englewood Cliffs, N.J.: Prentice-Hall.

Cirincione, Carmen, Steadman, Henry J., and McGreevy, Margaret A. (1995) Rates of insanity acquittals and the factors associated with successful insanity pleas. *Bulletin of the American Academy of Psychiatry and Law* 23, 399-409.

Darnton, Robert (1995) *Forbidden Best-Sellers of Pre-Revolutionary France.* Norton.

Daw, Nathaniel D., Niv, Yael, and Dayan, Peter (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience* 8, 1704-1711.

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

DeWaal, Frans (2007) *Chimpanzee Politics.* Johns Hopkins U.

DuBreuil, Benoit, (2009) Paleolithic public goods games: The evolution of brain and cooperation in Mid-Pleistocene hominins.
http://african.cyberlogic.net/bdubreuil/pdf/PPGG.pdf.

Dworkin, Gerald (1988) *The Theory and Practice of Autonomy.* Cambridge U.

Eccles, John (1994) *How the Self Controls its Brain.* Berlin, Springer.

Frith, Uta and de Vignemont, Frederique (2005) Egocentrism, allocentrism, and Asperger syndrome. *Consciousness and Cognition* 14, 719-738.

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

Gibbon, John. (1977) Scalar expectancy theory and Webers law in animal timing. *Psychological Review* 84, 279-325.

Glimcher, Paul William, Kable, Joseph, and Louie, Kenway (2007) Neuroeconomic studies of impulsivity: Now or just as soon as possible? *American Economic Review* 97, 1-6.
Goodwin, Donald W. (1986) Heredity and alcoholism. *Annals of Behavioral Medicine* 8, 3-6.

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

Green, Leonard, Myerson, Joel, Holt, Daniel D., Slevin, John R., and Estle, Sara J. (2004) Discounting of delayed food rewards in pigeons and rats: Is there a magnitude effect? *Journal of the Experimental Analysis of Behavior* 81, 39-50.

Greene, J. and Cohen, J. (2004) For the law, neuroscience changes nothing and everything. *Philosophical Transactions of the Royal Society London B*, 359, 1775-1785.

Greene, J. and Haidt, J. (2002) How (and where) does moral judgment work? *Trends in cognitive Science* 6, 517-523.

Haji, Ishtiyaque (2002) Compatibilist views of freedom and responsibility. In Kane, R., ed. *The Oxford Handbook of Free Will*. Oxford, pp. 202-228.

Henderson, Mark (2009) Chimpanzee's plan to attack zoo visitors shows evidence of premeditated thought. *The Times (London)*, March 10.
http://www.timesonline.co.uk/tol/news/science/article5877764.ece

Heyman, Gene M. (2009) *Addiction: A Disorder of Choice.* Harvard U.

Hobbes, Thomas (1651/1996) *Leviathan.* Cambridge, UK, Cambridge.

Hofmeyr, André, Ainslie, George, Charlton, Richard and Ross, Don (under review) The relationship between smoking and reward bundling in a group of South African university students. *Addiction.*

Holton, Richard (2009) Determinism, self-efficacy, and the phenomenology of free will. *Inquiry,* 52, 412-428.

Hume, David (1748/1962) An inquiry concerning human understanding. In A. Flew (ed.), *Hume on Human Nature and Understanding*. Collier, pp. 21-163.

Humphrey, Nicholas and Dennett, Daniel C. (1989/2002) Speaking for our selves: An assessment of multiple personality disorder. In Humphrey, Nicholas, ed., *The Mind Made Flesh: Essays from the Frontiers of Psychology and Evolution*. Oxford, pp. 19-47.

Iacoboni, M. and Dapretto, M. (2006) The mirror neuron system and the consequences of its dysfunction. *Nature Reviews Neuroscience*

James, W. (1884/1967) The dilemma of determinism. in J. McDermott (ed.), The Writings of William James, Chicago: University of Chicago.

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

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.

Kapitan, Tomis (1986) Deliberation and the presumption of open alternatives. *The Philosophical Quarterly* 36, 230-251.

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.
La Mettrie, Julien (1748/1999) *Man a Machine*. Open Court.

Kirby, Kris N. (2006) The present values of delayed rewards are approximately additive. *Behavioural Processes* 72, 273-282.

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.

Klein, D. B. (1970) *A History of Scientific Psychology.* Basic.

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

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

Mazur, J.E. (1987) An adjusting procedure for studying delayedreinforcement. in M.L. Commons, J.E. Mazur, J.A. Nevin, andH. Rachlin, (eds.), Quantitative Analyses of Behavior V: The Effect of Delay and of Intervening Events on Reinforcement Value, Hillsdale, N.J.: Erlbaum.

Mazur, James E. (2001) Hyperbolic value addition and general models of animal choice. *Psychological Review* 108, 96-112.

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

McGreevy, M.A., Steadman, H.J. and Callahan, L.A. (1991) The negligible effects of California 1982 reform of the insaanity defense test. *American Journal of Psychiatry* 148, 744-750. In PennLaw.

Mele, Alfred R. (1995) *Autonomous Agents: From Self-Control to Autonomy*. New York: Oxford.

Merali, Zeeya (2007) Free will: Is our understanding wrong? *New Scientist* 2615, 10-11.

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 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, John, Royzman, Edward B., and Schwartz, Barry (2005) Explaining away responsibility: Effects of scientific explanation on perceived culpability. *Ethics and Behavior* 15, 139-158.

Nadelhoffer, Thomas, Morris, Stephen G., Nahmias, Eddy A. and Turner, Jason (2005) Surveying freedom: Folk intuitions about free will and moral responsibility. *Philosophical Psychology* 18, 561-584.

Nozick, Robert (1981) *Philosophical Explanations.* Belknap.

Poundstone, William (1992) *Prisoner's Dilemma: John von Neumann, Game Theory, and the Puzzle of the Bomb.* Doubleday, 1992.

Premack, David (1970) Mechanisms of self-control. In W. A. Hunt, ed., *Learning Mechanisms in Smoking* Chicago, Aldine, pp. 107-123.

Railton, Peter (1984) Alienation, consequentialism, and the demands of morality. *Philosophy and Public Affairs* 13, 134-171.

Rescorla, Robert A. (1988) Pavlovian conditioning: Its not what you think it is. *American Psychologist* 43, 151-160.

Russell, Paul (1992) Strawson's way of naturalizing responsibility. *Ethics* 102, 287-302.

Sappington, A.A. (1990) Recent psychological approaches tothe free will versus determinism issue. *Psychological Bulletin* 108, 19-29.
Smart, J.J.C. (1961) *Philosophy and Scientific Realism.* New York, Humanities Press.

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

Smilansky, Saul (2000) *Free Will and Illusion.* Oxford, Clarendon.

Smilansky, Saul (2002) Free will, fundamental dualism, and the centrality of illusion. In Kane, R., ed. *The Oxford Handbook of Free Will*. Oxford, pp. 489-505.

Stapp, Henry P. (1998) Pragmatic approach to consciousness. In K. H. Pribram Thieme, ed., *Brain and Values*, Erlbaum, pp. 237-248.

Stitzer, Maxine and Petry, Nancy (2006) Contingency management for treatment of substance abuse. *Annual Review of Clinical Psychology* 2, 411-434.

Strawson, Peter (1962/2003) Freedom and resentment. In Watson, Gary, ed., *Free Will,* 2d Edition. Oxford, pp. 72-93.

Strawson, Galen (1986) *Freedom and Belief.* Oxford.

Vohs, Kathleen D. and Schooler, Jonathan W. (2008) The value of believing in free will: Encouraging a belief in determinism increases cheating. *Psychological Science* 19, 49-54.

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

Wolf, Susan (1990) *Freedom within Reason.* Oxford.

Notes

[1] Some have pushed this idea even further, perhaps unnecessarily far, by denying that "content neutral" mechanisms are enough to make the will free without an origin in the person's core values (*e.g.* Benson, 1994; Wolf, 1990).

[2] Of course what literally governs the bidding in the internal marketplace is the *prospect* of reward, which does not itself occur until after the choice is made; but choosing according to prospects is also the process in literal markets.

[3] Richard Holton gives a good example of two sequential stages of temptation in a passage from Ignatius Loyola: "One sins venially when the… thought of committing a mortal sin comes and one gives ear to it, dwelling on it a little or taking some sensual enjoyment from it…" (2009, p. 421). Presumably failing to avoid the thought is sinful because it raises the amount of willpower needed to avoid the deed.

[4] It might seem that unwillingness to hurt the other was itself the motive for self-control, but we are talking about the case where the other's knowledge of your behavior—and possibly the very existence of the other—is fantasied. I have hypothesized that the factor promoting this hypothetical case above mere imagination is the verifiability of its effect on intertemporal bargaining (Ainslie, 1975): If you believe that Saint X will help you avoid a temptation, this will be a self-confirming prophesy, as will be the effect of lapsing and disappointing her. To the extent that the imagined other is a way of understanding a larger class of incentives at stake in individual choices, she becomes not just a fantasy or a memory but a substantive factor in choice-making.

[5] The actual deficit in autistic spectrum disorders is probably more complex, perhaps "a disconnection between a strong naïve egocentric stance and a highly abstract allocentric stance" (Frith & deVignemont, 2005).