An empirical theory of the will which philosophy ignores at its peril. Review article of George Ainslie: *Breakdown of will.* Cambridge: Cambridge Univ. Pr., 2001.

How does a reviewer do justice to a book as rich, innovative and difficult as Ainslie's *Breakdown of will*?[1] In this review my strategy will be to address readers in a situation similar to my own when I first encountered the book: not *au fait* with many of the theories which form its backdrop (behaviourism, utility theory, game theory), and not inclined to see these theories as interesting and worth pursuing. In my own case this lack of interest was because these theories did not seem to illuminate the aspects of human life that I found most interesting – the stuff of novels and of psychoanalysis; everything that makes daily life so complex, unpredictable and, sometimes, unbearable (even where external conditions do not seem to warrant it). The self-induced suffering that novels and psychoanalysis deal with has always been important, but in the First World it becomes even more so as wealth, the rule of law, science and technology make more and more of the suffering coming from outside avoidable. The agent standard utility theory seems to presuppose is nothing to write a novel about, and neither is the human organism as described by classical behaviourism.

To convince sceptics that *Breakdown of will* (henceforth: *Breakdown*) is important enough to justify reading a review of it, let me thus first situate it by means of a thumbnail sketch. George Ainslie is not only a psychiatrist who has worked extensively with addicts and war veterans, but also a distinguished experimenter and theorist in behavioural psychology. He is also a long-standing participant in Jon Elster's international interdisciplinary seminar on irrationality. At the heart of his theory stands a single move, a twist to standard utility theory, which introduces disorder into the heart of human desire. This elegant twist modifies our picture of human decision-making fundamentally: *the preferences of real humans fluctuate for no better reason than the passage of time, in contrast to those of the economic agent presupposed by standard utility theory*. This shows itself most simply in the reversals of preference that occur regarding paired options of sooner smaller versus larger later rewards: sooner sources of reward tend to be overvalued as they become imminent, so that sources of larger later reward are sacrificed to them. To account for this, Ainslie modifies our understanding of the fundamental nature of the way in which agents *discount the future*, which is to say: value future goods less than the same goods in the present. Building on a substantial body of empirical research, Ainslie replaces the exponential discount function generally assumed by utility theory with a hyperbolic function. A picture of humans as having stable preferences is accordingly replaced by a picture in which their preferences are intrinsically unstable. Having modified this one assumption of standard utility theory, Ainslie tries to think through its

[1] In this article I refer to pages in Ainslie's book simply by numbers in brackets, so as to avoid countless repetitions of (Ainslie 2001: *nn*). In my exposition I sometimes quote passages from two other publications by Ainslie (esp. Ainslie 1992; Ainslie 2005), where I think they are more suited to my purposes than the corresponding passages in *Breakdown of will.*

implications. Utility theory and behaviourism then suddenly give a picture of human beings as recognisably pig-headed, complex, internally fractured and conflictual, and crucially: given to self-defeating behaviour – because prone to reversals of preference.

In assuming exponential discounting, standard economic theory takes preference rankings to be stable over time (bar changes in budget or information)[2]. In contrast the hyperbolic discounting hypothesis predicts that preferences over two options will often reverse as – and for no better reason than the fact that – possible satisfactions for the sooner and smaller of two options draw closer in time. So if on Friday I prefer being sober next week Monday morning to getting drunk on Sunday evening, standard economic theory predicts I will also prefer sobriety come Sunday evening. But of course this is not how the 'temptation of drink' works. If our default discount function is hyperbolic rather than exponential, this fact in itself predicts that our preferences will often show such reversals.

Ainslie's title tells us that his book is about the will. How does willpower, a concept which standard utility theory and many philosophers (such as Ryle 2002) find redundant, fit into this picture? 'The will,' or 'willpower,' is a name for some of the strategies we use so as not yield to temptation (no news this to countless religious and moral traditions), or in Ainslie's language, so as to safeguard ourselves – our future selves, to be precise – against the damage wrought by preference reversals, which involve sacrificing longer-term interests to shorter-term interests. According to Ainslie, when a hyperbolic discounter bundles series of future choices into categories, this can stabilise preference, and such a bundling is exactly what willpower involves. The hyperbolic drinker, tempted to drink on Sunday night even if it means a highly undesirable disruption of her Monday, can change the terms of her choice significantly if she frames it as a choice between drinking on Sunday evenings *in general* and having productive, hangover-free Mondays *in general*. (Below we will see why).

Ainslie shows how brittle willpower is as an achievement, how easily it 'breaks down'. This is because those interests of ours that are served by maintaining the personal rules involved in willpower (such as 'no drinking on Sunday nights') are opposed by other interests that undermine these rules by constantly seeking exceptions to them. Given that these conflicting interests play themselves out in different time frames, Ainslie describes the will as an intertemporal bargaining situation between successive selves. These successive selves arise because my utility function – my hierarchy of preferences – is not stable over time. Accordingly the relationship I have to my future selves resembles my relationship to other people. Because I can expect my future selves not to share my current hierarchy of

[2] Though economists, particularly behavioural economists, should know post-Simon (1956) that the old notion of agents with perfect information is empirically incorrect, most economics textbooks and many, if not most, practising economists still assume the simplifying fiction of agents with perfect information. For an agent with perfect information, information only changes if the facts with which the information deals, change.

preferences I'll sometimes act strategically to forestall them from acting on *their* preferences.

If his account of impulsivity and willpower had been all Ainslie had to offer, his achievement would already have been impressive. But, contrary to most writers on the will, he sees willpower as a mixed blessing. The last section of the book is devoted to systematically unpacking the downsides of willpower – of a strong will, or of 'rationality' – downsides such as compulsivity and loss of appetite (the potential for 'reward').

Ainslie presents a scientifically highly sophisticated application of the (suitably modified) economic paradigm to psychological issues. He offers us a form of behaviourism that in the complexity and subtlety of the phenomena it predicts bears little resemblance to most of what I had associated with that much-maligned paradigm in the past. His form of utility theory is simultaneously a fundamental critique of the approach usually going under this name, and a critical/sympathetic reworking of many psychoanalytic ideas. (In Ainslie's hands behaviourism and psychoanalysis *mirabile dictu* cease being essentially incompatible; one of his articles (Ainslie 1989) was originally titled "Sending Skinner to find Freud", but the editor of the journal objected). Because of all this, *Breakdown* fundamentally dislodged many of my entrenched prejudices. It straddles all sorts of divides often treated as unbridgeable: the humanities versus the sciences, the continental versus the analytic, the romantic versus the classical, the postmodernist versus the modernist. While satisfying the standards of clarity, elegance, rigour, argument and evidence usually associated with science, 'classicism', 'modernism' and analytic philosophy, it gives a powerful account of the irrationality, complexity, multiplicity, unpredictability, instability, conflictuality and paradoxical self-referentiality of the individual psyche, features that take centre stage for humanists, romantics, existentialists and 'postmodernists'. To avoid attacking straw men, students of human thought and action who are inclined to reject or criticise behaviourism, utility theory or economics will have to take note of this book, in which these paradigms attain a remarkable power and subtlety for understanding the human psyche.

For me, discovering Ainslie was somewhat similar to discovering chaos theory years ago. I had never imagined that the unpredictability and complexity of weather phenomena (for instance) could be modelled by such simple equations. There is no inductive route leading from the complexity of weather phenomena to dynamic systems theory, and it was counterintuitive to think that these phenomena could be modelled by a few simple mathematical equations. Similarly, the everyday phenomena of impulsivity, temptation and willpower would never have led me to suspect that some simple mathematical formula could give a key to a plausible model of these phenomena, and make them a function of time. With the benefit of hindsight, however, the route that led to the theory in *Breakdown* makes absolute sense.

To philosophers who think that it is possible to give a conceptual clarification of the will without committing oneself to any empirical hypotheses, this book should be an eye-opener. It is exactly Ainslie's empirical hypothesis that leads to a clarification, and partly revision, of our notion of will. (As is often the case with a good theory, it also enables the author to give a coherent account of the history of previous thinking on the subject, as well as a searching critique of previous positions). However, describing *Breakdown* as a work just addressing one specific topic – the will – risks obscuring the fact that it offers a fundamentally new vision of personhood and human agency in general. (We could also say that it puts the will – back – in the centre of any theory of mind).

The foregoing trailer should allow readers to decide whether they wish to sit through the feature film, which will now be screened. The territory which we above covered rapidly from a bird's eye view, will now be traversed on foot, one (admittedly still large) step at a time.

*** *** *** *** ***

Ainslie develops his theory of hyperbolic discounting and the will on the basis of a sophisticated version of behaviourism to which utility theory is integral. This theory allows him to account consistently and parsimoniously for a wide range of phenomena. Before proceeding, let us outline the theoretical premises on which Ainslie's theory is based, most of which are argued for at some point in *Breakdown*. (While deviating from the metapsychology in Freud and other variants of psychoanalysis, Ainslie's theory shares with psychoanalysis the conviction that all behaviour, however distressing to the person involved, or otherwise bizarre, is motivated. As in Freud, Ainslie's notion of action/behaviour is much broader and more inclusive than the corresponding folk notion).

1. Economic analyses apply to *all* rewards – in Ainslie's hands such analyses aren't skewed in the direction of monetary or material rewards. In fact he helps us see what it is about money that makes it easier to be 'rational' about money than about other, subtler goods that are more resistant to quantification (and thus mental arithmetic), and concomitantly makes it easier to 'take account of' monetary rewards than of such subtler, and often larger, rewards.

2. All behaviour[3] is driven by expected reward rather than being stimulus-driven. This means that, in line with Dennett's intentional stance, to think in terms of motivation is to think teleologically: the crucial question is: what reward does this behaviour aim at? In line with this, Ainslie sees no need to separate classical and operant conditioning theoretically. "[R]eward and the selective process in UCSs [unconditioned stimuli] are identical" (Ainslie 2010: 229). "[T]he main

[3] This excludes responses of the reflex-arc variety.

reason for their theoretical separation has been the need to explain why subjects pay attention to negative stimuli or engage in negative emotions. If pain, anger, disgust etc. don't have to be imposed on you but rather can lure you, this need goes away" (Ainslie, personal communication, 13 December 2010). Ainslie does not take aboard the blank-slatism which is conventionally seen as part and parcel of behaviourism: the proximal mechanism whereby an organism follows instinct is that doing so is rewarding to the organism.

3. Behaviour is everything that can be modified by reward. Reward is everything that can serve to modify behaviour. The categories of 'reward' and 'behaviour' thus become highly abstract. *Behaviour* ceases to be limited to externally perceptible behaviour; Ainslie (1985:54) says that behaviourism erred where it promoted a methodological principle (scientists must necessarily limit their observations, measurements and experiments to observable behaviour) to an ontological principle (behaviour is coextensive with observable behaviour). Moreover, *reward* is no longer equated with 'pleasure' or 'the avoidance of pain,' as happened in Bentham and much of the utilitarian tradition he engendered.[4]

4. There is a single dimension of reward. When I am confronted with a cheating student, I am torn between various motives that all have considerable force for me: moral demands (students who cheat should not get off scot-free), feelings of sympathy (pity and liking for the student I will have to report), and various 'egoistic' considerations (if I report the cheating, the disciplinary ramifications of the case are going to take a lot of my time; if I don't report it I may get into trouble). In the end, however, I will act in one way or another; there will be an outcome to this clash of what some would argue to be incommensurable motives. Ainslie gives a sophisticated argument why different motives must in the end be commensurable: every motivational factor impinging on my behaviour must be expressed in a single currency if it is to be able to interact with, or be weighed against, other factors. This is not a disguised form of psychological egoism – the claim that in the end we always act for selfish reasons, never for altruistic ones. In terms of Revealed Preference Theory (a systematisation of utility theory): how I spend my time, money and other resources *reveals* my preferences – what value I attach to having consumer articles, doing something for the needy,

[4] Had he followed the same route as Ainslie, one of Freud's basic premises – that all behaviour is motivated – need not have been formulated as 'all behaviour aims at pleasure'; he would then also not later, in *Beyond the pleasure principle* (Freud 1975 (1920)) have had to formulate a special exception to this formula, leading him to the compulsion to repeat and the death drive. In fact the death drive is partly an attempt to account for the very phenomenon at the centre of Ainslie's concerns: the existence and tenacity of self-defeating behaviour.

spending time with my family, doing well in my field of work. And the very fact of appearing in a preference ranking means that ranked items are in that sense commensurable.

5. Whatever "promises the greatest discounted reward at a given moment gets to decide my move at that moment" (Ainslie 2005:637). "Utility theory says that the experience of reward is the fundamental selective factor for behaviors, so that you can't stand outside of that experience and choose dispassionately among rewards" (39).[5] This means that in thinking about rationality, morality, impulse control or the will, Ainslie assumes the principle that the subject is always bound to the constraint that no choice can decrease the discounted reward expected at the moment it is made. The experimental evidence for this is associated with Herrnstein's Matching Law: in foraging behaviour (and all behaviour can be seen as foraging for reward) the amount of time spent on two different activities each giving rewards at zero delay is proportional to – "matches" – the rewards associated with each activity.[6] When there is a difference in the expected delay of the rewards, the rewards are discounted hyperbolically for delay. This means that the discounted values of two otherwise identical rewards A and B differing only with respect to their delay, are inversely proportional to their delays: if the delay to B is twice the delay to A, the discounted value of B will be half the discounted value of A. (This quick and dirty account of how delay enters into the Matching Law, will be refined below (p. 15)).

Hyperbolic discounting and preference reversals. A central premise of standard utility theory is that people discount future rewards as a function of *delay* – the longer the delay, the more the value of a reward is discounted. Given the choice between $100 at an earlier date and $100 at a later date, we tend to opt for the earlier. So far, so good. Ainslie agrees. People discount the future. But according to what type of discount function? Standard economic theory takes the economic agent to be an exponential discounter. Why should one discount the future exponentially? *Exponential discounting is the only stationary discount function that gives stable preferences over time. "All other discount functions imply that preferences are dynamically inconsistent: preferences will sometimes switch with the passage of time*" (Laibson

[5] This formulation on its own sounds slightly more mechanical, less teleological, than the usual formulation which makes *expected* reward the crucial factor. Expected reward will often closely track reward experienced in the past; however, the two can also diverge widely – if they couldn't, most advertising would be pointless.

[6] Among academics in the humanities there is a popular belief that behaviourism has been refuted in psychology. In fact nobody has shown that the Matching Law, the core of state-of-the-art behaviourism, is fundamentally mistaken. The turn to cognitivism in psychology (Baars 1986) was essentially a turn *away* from questions of motivation (Baars 1986:109-110), not a turn *to* a different, improved theory of motivation. People still are Skinnerian creatures (Dennett 1981), even if they are *also* capable of much more sophisticated forms of learning and cognition (Dennett 1996).

2005:10; my italics).[7] Preference reversals mean I cannot consistently pursue any set of interests, including, crucially, my longest term interests. They thus lead to an inefficient use of time and other resources. (Exponential discounters will follow a straighter route than hyperbolic discounters to reach their goals). I have to devote current resources to undoing the consequences of my previous choices. Being subject to preference reversals also puts you at a serious competitive disadvantage in a market where other players have stable preferences – or even: preferences more stable than yours.

Of all possible discount functions, the economic agents described by standard economics use the one which gives preference stability over time. Economic agents are rational. However, real people are not rational, and it has been a major challenge to economic theory to account for these lapses from rationality, lapses from behaviour consistent with the default exponential discounting function it assumes.

The Matching Law. This is where Ainslie's central trick comes in: on the basis of research conducted by Herrnstein, himself and others, he concludes that the default discount function of real people is *not* exponential. This means that their preferences need not be stable over time. Ainslie thinks that the experimental results achieved are best fitted by a hyperbolic function. As such results are obtained not only with humans, but with *all* vertebrates (Ainslie 2005:649), hyperbolic discounting cannot just be an artefact of culture. Ainslie builds on Herrnstein's (1961) proposal, inspired by the consistency of animal findings, of his 'Matching Law', to account for all choices. According to the Matching Law the frequency with which rewards are chosen is directly proportional to their size and frequency, but inversely proportional to their delay. "The word 'matching' comes from his original experimental design, in which pigeons pecked to get food on two independent keys that paid off at different rates. He found that relative rates of pecking matched the amounts, frequencies and immediacies of reward" (207n14). Ainslie's whole model assumes this matching of behaviour and reward. All behaviour, including human behaviour, is taken to aim at harvesting reward: a behavioural preference for A over B means that the discounted expected reward for A should be assumed to be higher than that for B. If we ignore future discounting or assume exponential discounting, human irrationality would seem to belie such a match between behaviour and reward. However, it is exactly the substitution of hyperbolic discounting for exponential discounting that enables utility theory to account for irrationality, while holding on to such a match. This substitution changes the whole complexion of the way in which an agent's behaviour reveals her preferences; it now makes sense to see a clash between her de facto behaviour and her long-range interests. Where reversals of preference occur in an exponentially discounting economic agent, we must assume that there has been a change of information or budget on the agent's part; in a hyperbolic discounter, however, such reversals are ubiquitous, even when there are no changes in budget or information.

[7] Changes of preference because of changes of budget or because of new information – see note 2, above – are not what is at issue here.

Ainslie was the first to see that Herrnstein's Matching Law implied instability of preference, and thus had devastating consequences for the applicability of standard economic theory, which assumes intrinsically stable preferences, to human and animal behaviour.

If people's default discount function is in fact non-exponential, then rationality – stability of preference – becomes the exception needing to be accounted for. (Instead of being puzzled about lapses from rationality we are now puzzled about lapses from irrationality).

If standard utility theory thus assumes a discount function that is contradicted by recent empirical research on individual humans and other vertebrates, why has it seemed to function so well as a guide to sound policy? Ainslie's answer is simple: standard utility theory is not a *descriptive* (psychological) theory, but a *normative* (non-psychological) theory. It is a model of how economic agents *should* behave to maximize their utility in the long run – or as we say, 'behave rationally'. (We paved the way for this conclusion by posing the leading question "Why *should* one discount the future exponentially?").

A person with instability of preferences can prefer A to B and B to C, and then C to A. His preferences are then said to be cyclical or intransitive: A>B>C>A. If agents with cyclical preferences are offered the right deals at the right times, they can be stripped of all their assets. This is why exponential discounting is seen as optimal: it does not lead to preference-cycling or preference intransitivity, of which the following is an example: Every autumn the exponential discounter Edward Dagwood (ED) can sell the hyperbolic discounter Hilda Doolittle (HD) a winter coat at a premium, and every spring buy it back from her for a song. ED's (future discounted) preference ranking of dollars versus coats remains unchanged, whereas for HD it fluctuates as the annual occasion for needing the coat draws nearer and then passes.[8]

Everyday observation teaches us that people are often inconsistent, that they are impulsive and prone to yielding to 'temptation'. On the other hand, they are often also consistent, able to resist impulsivity and not yield to temptation. How on earth is utility theory supposed to model this pattern of not sticking to a pattern, discover method in this madness?

To account for this complex picture, a venerable tradition resorted to some sort of dualism, such as soul *versus* body, head *versus* heart, reason *versus* passions, or – in Freud's case – secondary process *versus* primary process. Fluctuations in behaviour would then express the vicissitudes of the interactions between whatever two principles we have chosen, and explain why we sometimes act 'impulsively' – 'succumb to temptation', in an older language – and sometimes act 'prudently' – are able to 'resist temptation'.

[8] The better ED is able to track HD's other preference reversals, the better he will be able to exploit them as well. (Because of the abstract nature of money, banks can even exploit the instability of HD's preferences without having to track their specific content).

In Freud the primary process, which characterises the unconscious (earlier model) or the id (topographic model), seeks immediate gratification, while the secondary process, which characterises the System Conscious/Preconscious (earlier model) or the ego (topographic model), is able to delay gratification for the sake of long-term satisfaction. This model constitutes a strange dualism of 'immediate' and 'long-term', rather than presenting time as a mathematical continuum. How exactly are we to picture the time scales involved? And what becomes of intermediate points on the time line?

In speaking of how we weigh present and future benefits against each other, utility theory, by contrast, treats time as a mathematical continuum, so that it is possible to specify any point on the continuum by a number.[9] So, given a fixed interest rate, my bank manager can tell me how much interest I will pay if I repay my loan tomorrow, in 63 days' time, or in 1,034 days' time. A single formula covers all these cases. It cannot be replaced by an 'immediate/long-term' dichotomy.

However standard utility theory would seem to have nothing to say about the phenomena of impulsivity, temptation and resisting temptation, because it presupposes exponential discounting, and thus stability of preferences.

*** *** *** *** ***

Enter Ainslie's modification of utility theory. Like standard utility theory, it treats time as a mathematical continuum[10], but unlike standard utility theory, it addresses the phenomena of impulsivity, temptation and resisting temptation, directly. Such phenomena, and thus the necessity of the will, are exactly what we should expect if *homo sapiens* is a hyperbolic discounter (that is: behaves in accordance with the Matching Law). To say that people are impulsive, is to say that they are subject to preference reversals. To find a strategy against impulsivity is to find a strategy for stabilising preferences. Willpower is such a strategy. Ainslie's hyperbolic discounting hypothesis tells us how both impulsivity *and* willpower work. Instead of needing two principles (where any explanation for why victory goes sometimes to the one, and sometimes to the other, is bound to be *ad hoc*), a single principle suffices.[11]

---

[9] Intermediate time values are theoretically well-defined. Even if there may be technical problems of application, in principle we know how to deal with marginal increments of time. The assumption of our being able to treat time as a mathematical continuum may break down at very small time scales (Dennett 1991: 139-170).

[10] "[H]yperbolic discounting seems to occur over all time ranges. Subjects choosing between hypothetical amounts of money at delays of years show it as much as those choosing between differences in food, or comfort, or direct brain stimulation, over periods of seconds" (44). This is a far cry from a *dichotomy* between 'immediate' and 'long term'.

[11] Ainslie's insistence on the existence of a single currency goes hand in hand with his anti-dualism: without a common currency, there is again room for some sort of dualism (or pluralism).
Science always has to try to unify things that are disparate; any dualism (or pluralism) of principles signals a scientific impasse, except if there is a determinate account of how the two (or more) principles relate to each other – which would mean that the dualism (or pluralism) is in effect subsumed under a monism. Freud's (1920—SE XVIII: 53) self-confessed propensity to dualism must thus necessarily be epistemically worrying:

Freud's model is an economic model, but by and large metaphorically so. It is not formulated in mathematical terms, and as we have just seen, time is not treated as a mathematical continuum. Previously, I never saw this as an objection. The clinician working with clients will not be able to measure preferences, and will not measure time intervals to understand what is going on. For a clinician using Ainslie's theory this will be no different. But in Ainslie's case the mathematics of the theory is exactly what makes the phenomena understandable, as well as making the principle on which the theory is based empirically testable. The basic mechanism can be established in experimental situations that make preferences and time intervals measurable. If they chime with the results of a large number of diverse experiments, we can assume that hyperbolic discounting also applies in everyday situations, where such measurements may be unavailable.

Now that Ainslie has shown that these complex, unstable phenomena *are* in principle amenable to an empirical, mathematical treatment which yields an elegant, highly plausible explanation for them, the absence of something similar in Freud counts against him in a way it didn't when there was no contender à la Ainslie. As long as the Ptolemaic system is the only astronomical game in town, its complexities are a surmountable objection (if any). But the moment a Keplerian or Copernican alternative comes along and reaches a particular level of articulation, this changes.[12]

To clarify how Ainslie's mathematically formulated hypothesis explains preference reversals – and beyond that, willpower as a strategy for stabilising preferences – we have to look at the technical heart of his argument. My account here does not aim at mathematical elegance, but at accessibility for those who, like myself, aren't used to thinking in mathematical or economic terms. (The maths is in any case simple).

Our views have from the very first been *dualistic*, and today they are even more definitely dualistic than before.

[12] It is clear that Ainslie's model, especially its fundamental explanans, hyperbolic discounting, could never have been inferred from, or even suspected on the basis of, qualitative data like those found in standard clinical experience. Only experimental situations in which time intervals and reward quantities are made strictly measurable and comparable could have generated data of the needed precision and type. (In line with this, the principle on which Ainslie's theory is based cannot be captured adequately or elegantly in non-mathematical terms). We can therefore safely assume that Freud did not formulate his psychological hypotheses or metapsychology after a considered rejection of an alternative approach à la Ainslie.

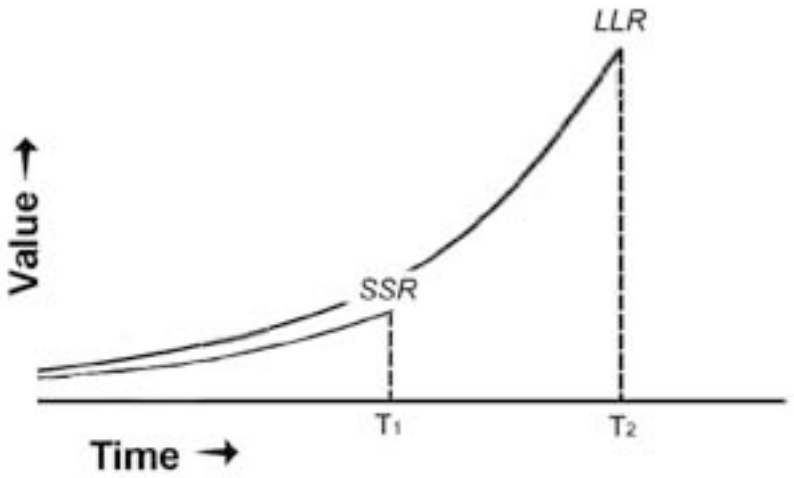
Like most theories in the social sciences, Freud's theory is path-dependent. Had he started from other presuppositions (e.g. the natural science, behavioural science and philosophy of science of the late XXth rather than the late XIXth Century) his theory – the bridge he built to link the clinical phenomena he encountered with what he, as a XIXth Century thinker, took to be the basic principles of science – would have taken a very different form. Even a genius is not capable of reading the laws governing behaviour from clinical material, unmediated by the prior beliefs with which she comes to this clinical material, and this becomes especially clear if these laws are necessarily mathematical in form.

These remarks are not meant to suggest that where Ainslie and Freud differ on any point, Ainslie is bound to be right, but only that there is a strong prima facie case that psychoanalytic metapsychology needs to be modified to accommodate the fact people discount the future hyperbolically, whatever the further merits or weaknesses of psychoanalysis.

In exponential discounting, preferences are stable. Given an exponential discount function, a fixed discount rate, and two options whose value (at zero delay) is fixed, and that are separated by a fixed time interval, the lines plotting the discounted value of the two options will never cross. This is true whatever the steepness of the discount function. I give two examples where the amounts and their timing are identical, but the steepness of the discount function differs, resulting in opposite preferences:

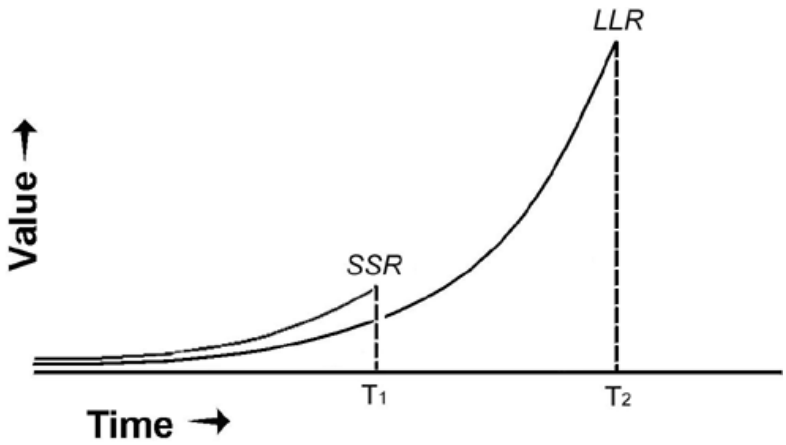
EXPONENTIAL DISCOUNTING

Figure 1A



Conventional (exponential) discount curves from an SSR (sooner smaller reward) and an LLR (larger later reward), where LLR is preferred.

Figure 1B



Conventional (exponential) discount curves from an SSR and an LLR, where SSR is preferred. [13]

In the second example the discount function is steeper than in the first one, leading to the smaller sooner option A being preferred over the larger later option B – but consistently so. A consistent exponential discounter will not regret having chosen the sooner smaller option A over the larger later option B; will not think in terms of yielding to temptation; and will not wish there had been a way it could have been resisted.

Any value generated by an exponential discounting function decreases or increases at a fixed proportion for each unit of time. If two options are both discounted at the same rate per time unit of delay, and I value the option A at 5% more than option B at any one time, then I will also value it at 5% more at any other time. My preference rankings will be stable; the lines plotting the value of each option

---

[13] Figure 1A describes a situation in which the LLR is preferred over the SSR, and Figure 1B a situation where the opposite is the case, despite the fact that the values of the SSR and LLR are the same, as is the delay between them. The reason the preference is different is because the future is discounted more steeply in the second case. This is of course not the only way in which such a shift could occur. Modifying the situation pictured in Figure 1A, the preference would shift to the SSR if

- the delay between the SSR and LLR were to become sufficiently large, or
- the difference between the undiscounted values of the SSR and LLR were to become sufficiently small, or
- the discounting rate of the exponential discounter were to become sufficiently steep (as in Figure 1B), or
- two or more of the above factors were to shift sufficiently in the indicated direction (even if none of them on its own was sufficient to change the preference to one of SSR over LLR).

Lest the reader feel exasperation that a point as elementary as this is belaboured here: it is my experience that many interpreters of Ainslie, even the brightest ones, tend to assume mistakenly that an exponential discounter will always prefer the LLR to the SSR, or that this follows from Ainslie's theory. Ainslie does however feed this misunderstanding with a slip of the pen.

- Under a diagram equivalent to our Figure 1A, Ainslie mistakenly says that when we are dealing with exponential curves, "[a]t every point at which the subject might evaluate earlier and later rewards, their values stay *proportional to their objective sizes*" (32; my emphases). This mistaken caption (corrected in Ainslie 2005) is then also quoted by Dennett (2003: 209). If this were true, it would necessarily imply that exponential discounters will always choose a LLR over a SSR. (For instance, prefer \$1010 in ten years' time to \$1000 tomorrow). But of course such proportionality does not obtain, not even in the case where the larger later reward is preferred by the exponential discounter – only where there is no difference in delay between two rewards do their discounted values stay proportional to their objective sizes. (However, this is trivial, as in this case it would also be true for a hyperbolic discounter). What is true is that for exponential discounters *the discounted values attached to any two rewards remain proportional to each other*. If I attach twice as much value to a later reward than to an earlier reward at one time, I will do the same at any other time (at which both rewards remain available).
- From a similar misreading, Wang & Simons (2005: 663-664) criticise, in the venerable pages of *BBS,* Ainslie's "assumption … that LL rewards are always superior to their SS alternatives". He generally assumes nothing of the sort, but his own lapse quoted above, as well as misunderstanding the generality of the diagram in Figure 1A, supports this misreading.

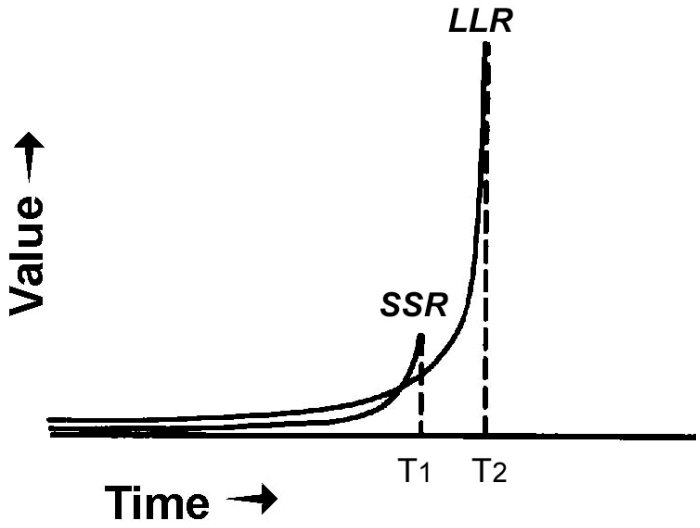
To avoid these endemic confusions, when explaining Ainslie's theory I always supplement Figure 1A above with Figure 1B, to underline that being an exponential discounter does not imply that LLRs are chosen over SSRs.

will never cross. The values remain proportional. Whatever the moment of choice, I will always choose A over B. This is in stark contrast to the preference reversals that can occur where discounting is hyperbolic.

From this flows another crucial feature of exponential discounting: the time factor that is relevant in choosing between two options is not the delay from the moment of choice to each of the two options, but the time interval between the first and the second option. As long as the two options are separated by an equal time interval, there will be no preference reversals. If an ED prefers $20 in six years' time to $10 in four years' time, then she also prefers $20 in two years' time to $10 immediately. For a hyperbolic discounter this need not be the case.

HYPERBOLIC DISCOUNTING

Figure 2



"Hyperbolic discount curves from a smaller-sooner (SS) and a larger-later (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". (Ainslie 2005:636)

For all discount functions except exponential ones, the values of two options occurring at different moments in time do not remain proportional, so that preference reversals can occur (but *need* not *necessarily* occur). Here we show how line crossing, and thus preference reversal, can occur in the case of a hyperbolic discount function. (Hyperbolic curves are more concave than exponential ones). For the exponential discounter the choice between the options A and B is simply a function of three factors: the value of A and B at consumption, the steepness of the discount curve, and the interval of time between the availability of the first and the second option. For the hyperbolic discounter a factor that was irrelevant for the exponential discounter becomes crucial: the delay from the moment of choice till the availability of the sooner option. Our graph shows how hyperbolic discounting can lead to preference reversal: as the moment of possible consumption of A becomes imminent, the discounted value of A becomes larger than the discounted value of B. A temporary

preference reversal occurs, in which A is preferred over B. An example would cover the same amounts mentioned above, so that the time-discounted preference for $20 (B) at T2 over $10 (A) at T1 is reversed for a short period as T1 becomes imminent. In retrospect the agent is liable to regret her choice for the sooner smaller option, as the opposite choice would have increased total reward; and the better she realises this, the more she will try to forestall similar choices in future.

Note that Ainslie's hypothesis is that our default discount function is hyperbolic, not exponential. It is not a hypothesis about how steep our discount function is; its steepness differs between individuals, and within individuals differs with situation and from one phase of life to the next.

On what sort of experimental findings does Ainslie base his claim that the default discount function of human beings is hyperbolic?[14]

An example: Subjects are offered a choice between one pizza now and two pizzas in a week's time. Then the subjects are offered the same choice at four weeks remove – that is, the choice between one pizza in four weeks' time, or two pizzas in five weeks' time. For exponential discounters, the two choices will be the same, and the same option will be chosen both times. A subject who discounts hyperbolically, on the other hand, will often prefer the sooner smaller option when it is available immediately (or at a short delay), while preferring the larger later option at a longer delay. The experimental findings confirm the hyperbolic discounting hypothesis.

Ainslie formulates the general principles exemplified in the pizza experiment as follows:

> The experiment used to test whether a subject's discount curves cross is simple: you offer subjects a choice between a small reward at delay D versus a larger reward of the same kind that will be available at that delay plus a constant lag, L. A subject gets the small reward at delay D from the moment she chooses or the larger reward at delay D + L. If she discounts the choices according to conventional utility theory, her curves will stay proportional to each other … But if she chooses the larger reward when D is long but switches to the smaller reward as D gets shorter, she's showing the temporary preference effect that implies a discount curve more bowed than an exponential one (31).

An immediate complication for the experimental investigation of such choices is that in everyday life the person – that is, the hyperbolic discounter – choosing a larger later reward at some remove is apt to realise that she is in danger of switching to the sooner smaller reward as it becomes imminent, instead of sticking to her choice for

[14] I initially formulated objections to and alternative explanations for these findings, for example: "subjects choose sooner smaller rewards because the larger later rewards are less certain than the sooner smaller ones". But I found that every objection that occurred to me had been anticipated at some point in the design of the experiments Ainslie refers to. Ainslie (2010) gives a systematic and, I think, judicious defence of his hyperbolic discounting model against two alternatives: conditioning (of which visceral learning is the latest avatar) and cognitive framing.

the larger later option. She will thus seek ways to forestall such a change. To avoid eliciting learnt strategies like this, the experimenter trying to gauge the subject's default discount function uses rewards that the subject experiences immediately on delivery, and that resist having mental arithmetic applied to them. Experiments of this nature do indeed indicate that subjects'

> basic discount curves cross and are thus more hyperbolic than exponential: People exposed to noxious noise and given a choice between shorter, earlier periods of relief and longer, more delayed periods choose the shorter periods when D is small and the longer periods when D is long. College students show the same pattern when choosing between periods of access to video games. Retarded adolescents show it in choosing between amounts of food. Certainly at the gut level, people's discount curves cross (33).

Ainslie (2005:636) thinks that Mazur's (1987) fine-tuning of the Matching Law covers the experimental findings (e.g. Grace 1994; Mazur 1997; Kirby & Marakovic 1995; Vuchinich & Simpson 1998; Green et al. 1994; Kirby 1997) adequately:

$$\text{Value} = \frac{\text{Value at no delay}}{[1 + (\text{Impatience factor} \times \text{Delay})]}$$

(The 'impatience factor' reflects how steeply the subject discounts the future (35; 208n15)).

Now, we noted before that people regularly come to regret their choice for sooner but disproportionally smaller rewards over larger later ones. On Monday (and later) I regret having to pay for Sunday night's revelry with feeling bad and being unproductive; after having chosen one pizza yesterday, I start thinking two pizzas in six days' time would have been a much better choice; immersed in noxious noise from which there is no escape I regret having chosen ten minutes' respite early on over half an hour's respite later on. People who realise (clearly or dimly) that their behaviour shows a pattern of preference reversals - 'falling for temptation' and regretting it afterwards – are motivated to find ways to stop decreasing available reward by sacrificing long-term interests to short-term ones. *This amounts to ways of committing themselves to the choices they make when the sooner smaller and the larger later options are both still further away in the future.* Without such *commitment* their choices will come to be empty – because invariably abandoned when crunch time comes – and they will seem to themselves the hapless playthings of their 'impulses'. Ainslie (75-85) distinguishes four basic strategies for commitment, of which the first three are strategies of *pre*commitment:

1. *Arranging external constraints.* Such a constraint can be physical - pills that make you feel nauseous if you drink alcohol, or Ulysses tying himself to the mast. More common is utilising the opinion of other people – friends, family or your fellow anonymous alcoholics – who will think badly of you if you don't stick to your commitment. Their good opinion of you now becomes an extra reward which will

be lost with the reward of the larger later option itself if you don't stick to the intention you've advertised to them. Both types of arrangement, the physical and the social, have their uses, but also their limitations.

2. *Manipulating your attention.* The core phenomenon here is avoiding "information that would change your mind" (76). This covers the Freudian defence mechanisms of suppression, repression and denial (76-77; more fully treated in Ainslie 1982 and Ainslie 1984). Such mechanisms however impede your ability to have your actions informed by sufficient, and unbiased, information. Another major drawback of these devices is that they can also – indeed even more effectively – be used by short-term interests against long-term interests.

3. *Preparation of emotion.* Emotions have intrinsic momentum; once an emotion is experienced strongly, the corresponding appetite becomes stronger. When anger flares up, the desire to inflict physical or mental pain on the object of the anger flares up concomitantly, becoming harder to resist. We could reformulate 'a desire to *x* flares up' in Ainslie's language as 'the reward expected from *x*ing undergoes a temporary increase'. "If a person expects an emotion to make an otherwise unpreferred reward temporarily dominant, he may commit himself not to choose the reward through early inhibition of that emotion" (Ainslie 1992: 136). We remarked earlier that Ainslie regards any process or condition that can be modified by reward, as behaviour. To him emotions are therefore behaviours. If an emotion is experienced as a temptation, that is, if it is an appetite we are (partly) inclined to resist, then we can learn not to have it if we catch it early enough. (Such learning often happens where reward for the appetite is unavailable, or where following the appetite will be punished – many smokers lose their craving whenever they find themselves in standard, recurrent situations where smoking is forbidden, for instance). This can happen by suppressing all emotion, or by cultivating contrary emotions – the Freudian defence mechanisms of "isolation" and "reversal of affect". This strategy is again a two-edged sword, because it can be used by short-range interests at least as effectively as long-range ones.

4. *Personal rules.* There is "a fourth strategy to defend your long-range interest: the *personal rule* to behave alike to all the members of a category" (84). This is the strategy of *willpower*, and it has a wider scope than the previous ones. Willpower "allows a person to resist impulses while he is both attracted by them and able to pursue them" (Ainslie 1992: 143). It "seems to be the strongest and most versatile" of the four strategies for commitment (89) – partly because it does not involve *pre*commitment; in it a person's commitment is tested with her decision regarding every new case. As it "[leaves] perception and emotion undistorted" (Ainslie 1992: 144), it is able to avoid many of the downsides of the second and third strategy, mentioned above, due to which they are often viewed as 'pathological'.
   Ainslie summarises the two core components involved in willpower as follows: "*choosing rewards in aggregates rather than individually* gives later, larger

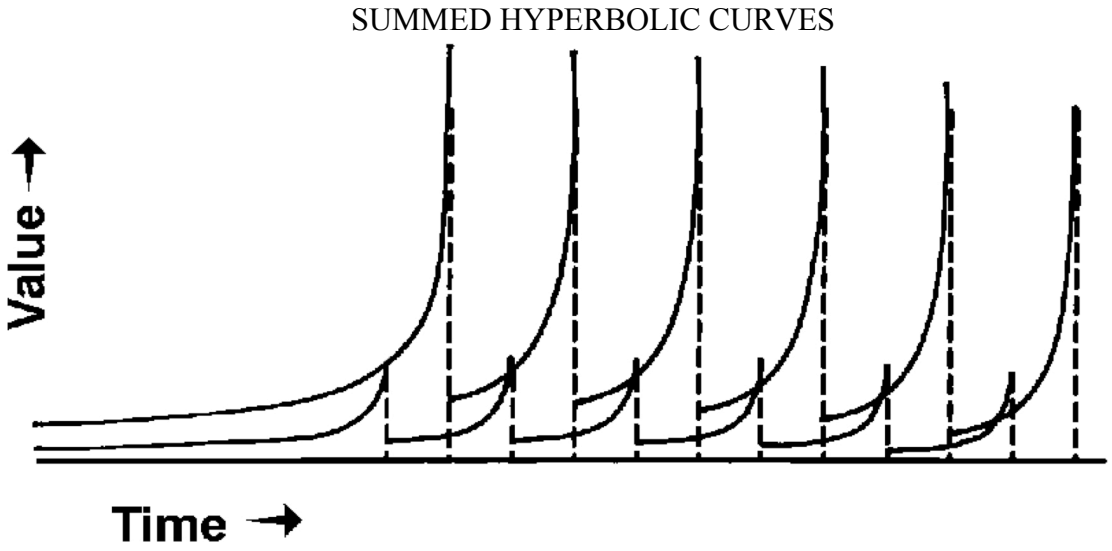
rewards a major advantage over smaller, earlier ones; and *the perception of one's current choice as a precedent predicting a whole series of choices* leads to just such aggregations" (Ainslie 1992: 144-5 – my italics).

We now look at each of the two components in greater detail.

A) *Choosing over categories.* For Ainslie personal rules are the core of will or willpower: choosing over categories rather than individual cases. (Over the centuries, many previous thinkers expressed ideas pointing in the same direction). The chances of giving up smoking become better if the smoker frames his choice as "Shall I smoke every day or not at all?" rather than as "Shall I light up a cigarette now or shan't I?" This follows from the nature of hyperbolic discounting curves, combined with the assumption that such curves are additive (an assumption for which there is some experimental evidence (81; 213n18)).

This can be represented in a graph: to see what a subject will choose when she has to choose between a series of Sooner Smaller and Larger Later Rewards: First draw the curves for the discounted values of the Sooner Smaller and Larger Later Reward as a pair of single options. Repeat this for the same two options moved further into the future from the moment of choice, etc. Then sum each of the two series of curves. The series with the highest summed curves at the moment of choice is the one that will be chosen, all else being equal.

Figure 3 (from Ainslie 2005:640)



This graph of Ainslie's can do with a gloss. It shows how summing a series of paired hyperbolic curves exactly like those in Figure 2, depicting a temporary preference for the sooner smaller reward, leads to the elimination of curve-crossing when the series becomes long enough. In the present figure, the curves for the final (i.e. right hand) pair of rewards are identical to those in Figure 2. The second pair of curves from the right is obtained by inserting a duplicate of the first pair of curves at an earlier point in time, and adding it to the discounted values which the right hand pair has at that earlier time. (The curves obtained by summing are thus slightly higher than the right

hand, unsummed curves). This process is iterated to keep on adding curves to the left, each pair slightly higher than the previous one, because the discounted values of all previously drawn curves are added to each new one.

Proceeding in this way with summing the curves will at some point lead to a cessation of curve-crossing. Figure 3 shows the last 6 items in such a series. For the last five pairs the curves still cross, that is, the sooner smaller option still dominates the larger later one at some moments. We are to conclude that *up to*, but not including, *the last five paired curves of this particular series, the larger later series will consistently dominate the smaller sooner one.* As the end of the series comes into sight there are periods in which the SS option dominates (the last five pairs of curves in the graph). The further the series extends into the future, the more closely the proportions of the discounted value for the two series approximate to the proportions of the undiscounted value of each of the two terms, so that the periods in which, and extent to which, the sooner smaller option dominates the larger later option progressively decrease (examples shown here, reading from right to left), and finally disappears (leftmost pair).

Choosing in categories or series thus stabilises choice, and stabilises it in the direction of the relative undiscounted values of the options concerned. The effect of both features is that we are less likely to sacrifice larger later rewards for sooner but disproportionately smaller rewards (disproportionate from the perspective of an exponential discounter), as our default hyperbolic discounting otherwise inclines us to do.

Were the curves to be exponential instead of hyperbolic, summing would not change their relative heights, regardless of whether in the one-off case the larger later or the sooner smaller option were chosen (Ainslie 2005:640). Choosing in categories or series, a feature that has long been identified as central to willpower, would thus not have made any difference – and thus, sense – had we been exponential discounters.

B) *Seeing one's current choice as a precedent.* Bundling choices into categories is one fundamental aspect of the mechanism of willpower. The other, more fundamental, aspect is *seeing one's current choice as a precedent for future choices*. If I 'yield to temptation' in making *this* choice, what reason do I have to think that I won't do the same thing in future? Conversely, the longer the series of choices in which I have successfully resisted temptation, the more confident I can be that I'll also be able to do so in future, and can thus count on harvesting the long-range rewards motivating my resistance to temptation. (There is however a major asymmetry between the effects of negative and positive precedents: a single lapse often undoes the effects of a long series of cases in which temptation was successfully resisted). We can only use the bundling strategy successfully if choices are viewed as precedents. Choice is a function of *expected* reward, and where my current choice acts as a *precedent* or *test-case* (Ainslie 2005:640), it is a (or *the*) major factor telling me what series of choices – and thus series of rewards – I can *expect* for the future:

> [H]ow does a person arrange to choose a whole series of rewards at once? … The values of the alternative series of rewards cannot depend on whether or not he will actually get them. … [T]he main element of uncertainty will be what he himself will actually choose. In situations where temporary preferences are likely, he is apt to be genuinely ignorant of what his own future choices will be. His best information is his knowledge of his past behavior under similar circumstances, with the most recent examples probably being the most informative. Furthermore, if he has chosen the poorer reward often enough that he knows that self-control will be an issue, but not so often as to give up the hope that he may choose richer rewards, his current choice is likely to be what will swing his expectation of future rewards one way or the other: If he makes an impulsive choice he will have little reason to believe he will not go on doing so, and if he controls his impulse he will have evidence that he may go on doing so.
>
> Consider, for example, the predicament of a person on a weight-reducing diet who has been offered a piece of candy. He knows that the calories in one piece of candy will not make any noticeable difference in his weight, and yet he is apt to feel that he should not eat the candy. What would it cost him? Common experience tells us: his expectation of sticking to the diet. He will face many chances to eat forbidden foods, and if he sees himself eating this one, it will not seem likely to him that he will refuse the others. If he succumbs to the temptation to eat the candy, it will cost him not only its small caloric burden but also the expectation of getting whatever benefits he had hoped for from his diet. And yet the very knowledge that he is in this predicament may make him refuse the candy where he would otherwise accept it (Ainslie 1992:149-152).

So, only if I can expect to stick to a rule can I bet with reasonable confidence on getting the rewards of the larger later option, and only if I can reasonably bet on them will I have a reason to forgo the sooner smaller reward, that is, not succumb to impulse or temptation. This describes a dynamic system in which feedback loops can at one moment engender stability, and at the next instability. "Personal rules are a *recursive* mechanism; they continually take their own pulse, and if they feel it falter, that very fact will cause further faltering" (88). My only reason to expect that I will stick to the rule is precedent. My current choice thus becomes especially important for its value as a precedent, even when the current rewards that are at stake in it are negligible.

This is where the extra motivation recruited by the strategy of will comes in: the incentives involved in the original choice are modified by the addition of a major incentive on the larger later side: safeguarding my expectation that I will keep on choosing in such a way that the series of larger later rewards is secured.[15] In negative

[15] As I discover the general advantages of the willpower move, a further incentive is added: the incentive of being the sort of person who in general is able to exercise willpower, over a wide variety of rewards, in a wide variety of situations. However, Ainslie (113-116) says that it will be the rare case in which a person stakes the credibility of her whole willpower on every single choice.

terms: I stand to incur a disproportionate loss if I cannot stick to the rule I have chosen, and knowing this drastically changes the terms of my choice.

Reward-bundling often gives insufficient motivation for resisting temptation – the current or next temptation; future temptations will have to be dealt with when *they* materialise. The added incentive of a choice's value as a precedent must be added to the bundled motivation. Because of hyperbolic discounting, even where bundling occurs, the immediately available sooner smaller option can still dominate the larger later option which succeeds it, and which is sacrificed by choosing for the former. So, having chosen for the series of larger later rewards over the series of sooner smaller rewards, there would be a recurrent situation in which the first member of the series of sooner smaller rewards was preferred as it became imminent, while looking beyond that into the future, choosing for the rule would be preferred. I would keep on choosing the rule for future cases, while reneging on it for the current choice. If that happened, I would soon realise that my choice for the series of larger later rewards was empty, because just as I had reneged on the rule in this case, or in the most recent series of cases, I would keep on reneging on it in future. I could only count on not reneging in future on preceding choices by establishing a precedent of not choosing the sooner smaller option now, and sticking to this precedent in future. The longer the series of precedents established in this way, the more credible my resolve to stick to my rule of choosing for the series of larger later options will become, and the better the chance that the whole series of larger later rewards will materialise. And the better that chance, the more likely that I'll be able to resist the temptation of each sooner smaller reward as it becomes imminent. (The less likely a reward is to materialise, the smaller the fraction of its consumption value that enters into the decision making process).

In this game, everything hinges on *credibility.* My credibility depends on a series of choices that function as precedents predicting my future behaviour. Credibility is undermined both by *lapses* – failure to stick to the rule – and by a slippery slope of *exceptions.* My credibility thus requires that I neither *break* the rules I have set for myself, nor hollow them out by negotiating an ever wider circle of admissible *exceptions*.

What is the difference between *lapses* – instances of rule-breaking – and *exceptions* – instances to which the rule does not apply? Willpower involves choosing over categories, but what exactly the categories are, and how exactly they apply to concrete situations, is often, or perhaps even usually, up for grabs. Long-range interests establish rules, while short range interests try to categorise the present case as an *exception* to the rule, rather than a *precedent of rule-breaking* (88). On the face of it, exceptions should be cases in which the rule does not apply, so that allowing the exception would not amount to breaking the rule. Because rules can be formulated in different ways, a short-range interest will root for those formulations of rules, or of exceptions, which allow it satisfaction. *Never* making an exception may mean major,

unnecessary losses of sources of reward.[16] Once a questionable choice has been made, it can be better to interpret it an exception, than to recognise a precedent of rule-breaking. (For a revealing example see Ainslie 2005: 669). But this is a risky strategy: if Thanksgiving is an exception as far as my diet is concerned, what about "my birthday …, the Fourth of July, St. Patrick's Day, Labor Day, Arbor Day, St. Swithin's Day, and Just This Once. This kind of logic can degrade a personal rule without ever breaking it" (87). Because what is initially seen as an exception to a rule can later be experienced as (equivalent to) a betrayal, there is no clear boundary separating *exceptions* from *lapses*.

My credibility is best served by linking any rule to a "*bright line*" (94ff), a demarcation that lacks, or seems to lack, the arbitrariness which opens the way to a rule's negotiability. To deal with temptations that I find really hard to resist, only bright lines will do. They help to stop the gradual erosion of a rule by an ever-widening circle of exceptions. There is a bright line between some drinking and no drinking, but no bright line between different diets. Ainslie sees this as part of the explanation why many alcoholics manage to stop drinking, while diets are hardly ever successful. Lines that are less bright can work where the value attached to an activity is not that high, or where the subject has attained sufficient skill at what Ainslie calls 'intertemporal bargaining between successive selves'.

Up to this point we have simply described Ainslie's model as if it assumes a unitary subject with a variety of interests, largely ignoring the range of concepts mentioned or implied in the words just quoted: 'intertemporal bargaining between *successive selves*'. It is now time to move beyond this simplifying fiction and to discuss Ainslie's essentially plural notion of the self. He arrives at this notion because he believes that game theoretic models of strategic negotiations *between multiple subjects* correctly describe – and explain the logic of – the situation in which different interests that become dominant at different times *within one subject* negotiate with each other. It is exactly the plurality of conflicting interests that makes it inappropriate to think of the subject as unitary. Each interest strategizes so as to outwit opposing interests; this often involves forestalling them so they won't get the upper hand later, when *they* become dominant.

By way of contrast with Ainslie's plural notion of the self, let us look at the unitary self we can expect in an exponential discounter. Barring changes of budget or information, the exponential discounter has a stable utility function: a clear hierarchy of preferences, presumably allied with some indication of cardinality (e.g. to account for the difference between my preference for pork over human flesh, and that of a cannibal who also happens to prefer pork). For an exponential discounter, game theory only covers interactions with *other* agents – 'inside' the person there is no plurality of selves that could play games with each other and try to outwit each other

[16] It is typical for Ainslie to be simultaneously aware of the advantages *and* disadvantages of every strategy.

strategically. All this will only happen with others. My interests will be completely integrated (an integration expressed in the consistency of my utility function), so that there will be no internal conflict, and thus no plurality of interests in the pregnant sense. To maximise my reward I will seek maximum information and apply my resources wisely. Maximising my reward will often involve playing games with (or against) other agents. I will make compromises to foster cooperation, enter into coalitions, try to pre-empt my competitors (for the best piece of land, a particular person as mate, a particular job), and so on.

In the above view the subject is unitary – there is no intrapersonal multiplicity; multiplicity is limited to interpersonal relations. Ainslie stands in a long tradition conceptualising the person as consisting of a society of interacting sub-agents. In Ainslie's particular version of such a plural conception of the subject, game playing and game theory also apply at the intrapersonal level. How does Ainslie come to his plural view of the self?

Ever since Watson a fundamental premise of behavioural psychology has been that rewards build (select, foster) behaviours to reap them. The more a behaviour is rewarded, the more entrenched it becomes, while behaviours that are never rewarded, extinguish over time.

In Ainslie's hands, the discovery that we are hyperbolic discounters adds a new twist to this fundamental behavioural premise, because the rewards in question now need not be compatible with each other. These incompatible rewards can build incompatible behaviours that become dominant at different times, with none of the incompatible behaviours permanently winning out over rival behaviours.

Behaviours built by one reward can undo the effects of behaviours built by another reward. (Recent US electoral politics gives a typical example: behaviours built in a presidential hopeful by the expected rewards of erotic adventures defeat the behaviours built by the expected rewards of high political office).

Given this danger, the reward-harvesting behaviour built by a particular reward will only be effective if it includes *strategic* behaviours that forestall the behaviours built by incompatible rewards. "[I]f a person is ever to see his current long-range plans realized, he must, in his current frame of mind, take into account the tendency of currently unpreferred goals to become preferred at a later time" (Ainslie 1992:93). (The presidential hopeful must take into account the danger that he'll blow his presidential chances by grasping a chance for adulterous sex).

What is needed to prevent this from happening is "some enduring commitment that will prevent the other reward from becoming dominant" (43). (The presidential hopeful (or his campaign manager), aware of his roving eye and the likelihood of erotic temptations on the campaign trail, could for instance arrange that an aide will ensure that he is never alone with women who are erotically tempting – a cumbersome arrangement, but perhaps nothing less cumbersome will work).

To talk about the conflict between incompatible rewards, and between the behaviours serving them, Ainslie uses the term "interests". This term clusters behaviours according to the reward they serve, e.g. "the gourmand interest", "the dieting interest". As in parliamentary politics, the term 'interests' is somewhat ambiguous: "Within the individual, both the particular motive [or equally: reward – AG] that gives rise to a set of behaviors and the set of behaviors themselves can be called interests" (Ainslie 1992: 90).

Ainslie's model of interests is bottom-up, not top-down; interests aren't options chosen by an overarching ego, but "independent opportunists that have grown to exploit particular sources of reward over particular time courses" (Ainslie 2005:637). "Regularly recurring rewards create internal interests in the same way that economic opportunities create businesses to exploit them" (Ainslie 1992: 90). We can only speak of interests by way of contrast with other, conflicting interests; where there is no conflict, speaking of interests is pointless (42).[17]

Ainslie (personal communication 2011) proposes clarifying the nature of interests, and how they relate to the whole person – the "self" in his parlance – via an interpersonal, political analogy: that of a sovereign legislature.

If some of the members of the legislature are supported by the arms industry, they will pursue available opportunities to act 'in the arms interest'. They could for instance seize upon a perceived military threat from outside to vote for an increase in the arms budget, and attempt to get a majority in the legislature to go along with this – an attempt which would involve actions like negotiating, logrolling, providing (dis)information and so on. If they were successful, then not only they, but the whole legislature would be acting 'in the arms interest'.

If those furthering 'the arms interest' foresaw a time in which their power would decline, they could, while still powerful enough, attempt to get the legislature (a majority, that is) to commit itself to sticking to a higher arms budget in future. The strategic precommitment in question can equally well be read as the actions of parts against parts – the currently dominant arms interest preventing rival interests from undoing the achievements of the arms interest when they become dominant – or of the current whole against future wholes – the current legislature preventing future legislatures from undoing its decisions. 'Members' in this legislature would translate to 'behaviours' in the whole person.

'Interest' is simply an analytic concept, and interests should thus not be reified into permanent entities; they "may coalesce or divide over time" (Ainslie 1992: 90).

[17] I have separate ice cream and diet interests, but do not have separate vanilla and chocolate ice cream interests. In the former case there is reason for each interest to try to forestall the other. In choosing between flavours, however, I just follow my momentary preference. Eating vanilla ice cream today does not undermine my chances of eating chocolate ice cream tomorrow, and vice versa. The consequences of eating vanilla ice cream do not differ from those of eating chocolate ice cream. The one does not present itself as a temptation in the way of obtaining the other. Last, but not least, these two tastes play themselves out over exactly the same time scale.

"Some of the 'arms' legislators … might also support good roads and fine art, in opposition to an interest in small government (that others of them might belong to), or in agreement with everyone. The members are the interest only in a manner of speaking, and they may be another interest later. The defining feature of the interest is the source of incentive; the members 'become' that interest to the extent that the incentive has motivated them to work for it" (Ainslie, personal communication 2011).

This model may sound implausible if we conceive of interests as necessarily being self-conscious. However, this need by no means be the case. Selection by reward is all that is needed for motives to make strategic moves, or to conflict with each other. This process will at best only become self-conscious at times and in part. [18]

Ainslie replaces the person as a unitary agent with a multiplicity in two different senses. On the one hand, as we have just seen, there is within one and the same person a multiplicity of conflicting interests – chunks of the person's behavioural repertoire which serve incompatible goals. The unitary self is thus replaced by an internal marketplace where various agents – "interests" – interact, and in the end the behaviour that promises the most reward, gets chosen. On the other hand, the priorities (preferences) of a person will change over time, for no better reason – given that the person is a hyperbolic discounter – than the passage of time. This is why Ainslie speaks of "successive selves", which take the place of the person whose unity over time is secured by the consistency of her preferences.

The multiplicity of interests is related to the multiplicity of successive selves in the following way: As different incompatible interests become dominant over time, this leads to a series of successive selves. 'Self' refers to the whole person at a particular time, and is "shorthand for 'the interest that is dominant at that moment' (Ainslie 2011: personal communication). 'Successive selves' is thus shorthand for the successive motivational configurations, each characterized by the interest which is dominant in it, which constitute the person's biography over time.

In economics an agent is something to which a consistent utility function can be attached. This chimes with our intuitive notion of a coherent agent as something consistently pursuing various goals of unequal priority. When we apply the standard economic criterion of agency impartially, a person who is prone to temporary

[18] One could think that selection by reward, the fundamental process described by behavioural psychology, must be a process too stupid to generate anything as smart as strategic behaviour. This would be a misconception. The fundamental problems faced by all vertebrates, and especially social animals such as the primates, are so often strategic in nature that a learning process which did not build strategic behaviours would be of very little value. Though such strategic behaviour *need* not be conscious, articulate or reflective, there is no reason why it *can*not in animals such as homo sapiens be (at times, partly) conscious, articulate or reflective. Note that even processes that are 'unsophisticated' in these respects can be highly sophisticated in terms of successful goal attainment – to use our previous terminology: successful reaping of reward. (Students of foraging theory for instance marvel at the efficiency – rationality, in a particular sense of 'rationality' – with which animals choose foraging behaviour that maximises their caloric intake, given the contingencies of their particular natural and social environment).

preferences does not at different times count as a single agent. In Ainslie's temporary-preference theory the person stops being the unitary estimator of standard utility theory, and instead becomes "a succession of estimators whose conclusions differ" (40). Ainslie thus replaces the unified person over time with a succession of selves, a diachronic multiplicity.

How does this parliament or marketplace of multiple selves come to achieve a semblance of unity over time, so that it becomes plausible to treat individual people as single (economic) agents? These interests are welded together by sharing a single, finite body, and a single, finite channel of attention in charge of it (41). Given these circumstances, each interest has to negotiate with others to secure as much of the reward it aims at, as possible; this negotiation is driven by scarcity, the central organising principle of economic theory.

The source of personal unity over time need not be any more substantial, any less virtual than this. The more bargaining breaks down, the less unity the person will show over time and the more zigzagging and backtracking there will be in the trajectory traced out by her behaviour.

Ainslie calls his approach picoeconomics: micro-microeconomics.[19] (*Breakdown of will* (2001) is a reworking of his 1992 book *Picoeconomics*). Microeconomics looks at economic behaviour at the level of interactions between individual economic agents, and takes the individual person (or other animal) as the smallest unit of analysis. Picoeconomics goes beyond microeconomics: to understand the behaviour of the individual person we analyse her as being the result of the strategic interaction of even more elementary agents. To deal with strategic behaviour economics employs game theory. According to Ainslie the game theory developed to understand negotiations between different agents (whether they be countries, companies or individual people) also allows us to understand the will, which he reads as not a faculty, but a bargaining situation between successive selves. Ainslie's big innovation here is to apply game theory to intrapersonal relations – how successive selves interact with each other strategically.

In applying game theory to the bargaining among successive selves, the game Ainslie uses to make his point is the Prisoner's Dilemma (PD), whose logic we will now examine.[20]

The matrix below (based on Ainslie 2001: 91) shows Country P's payoff for using or not using gas (for gas warfare) against Country Q in a one-shot Prisoner's Dilemma. In a "Prisoner's Dilemma" each player has to choose without knowing the

[19] Elsewhere, picoeconomics is described as the "study of [hyperbolic] curves and their consequences" (164). The two descriptions are compatible.

[20] Although the standard game Ainslie uses to analyse the pay-offs connected to various strategies is the Prisoner's Dilemma, according to Ross (2005:341) other types of game – "assurance games, pure and impure coordination games, inspection games", etc. – are equally relevant to will. Ainslie uses the PD because it is simple. However, we must keep in mind that it is unusual for a game to have a single equilibrium, like the one-off PD does.

other player's choice. (In a one-off Prisoner's Dilemma the cardinal values of the payoffs are irrelevant; only the preference ranking of the four options counts – in the current example 10 would thus be the most preferred, and 0 the least preferred option[21]):

Prisoner's Dilemma for Gas Warfare: Outcomes for Country P (91).

		If Country Q Chooses	
		Gas	No Gas
If Country P Chooses	Gas	2	10
	No Gas	0	6

Let us put ourselves in the two players' shoes. P must look at each of the options available to Q, and in the light of this make its own choice. Suppose Q chooses 'gas'. In that case P is better off choosing gas as well, as that secures a payoff of 2 instead of 0. Suppose Q chooses 'no gas'. Then P is also better off choosing gas – a payoff of 10 instead of 6. So whatever Q chooses, P's best choice is gas.[22] Because Q will reason in exactly the same way as P, both parties will end up using gas, and thus be much worse off than had they cooperated on not using gas. However, there would seem to be no road leading to such cooperation, desirable as it may be. The logic of the one-off Prisoner's Dilemma locks each party into a non-optimal choice.

If the game is going to be repeated indefinitely, things change dramatically, however. In this case P would have a motive not to use gas in this round, as this gives Q an incentive to cooperate – that is, not to use gas either (Q can read P's move of not using gas as an offer to cooperate (Axelrod 1990: Ch 4 [on WW1]))[23], so both get a whole series of 6 rather than 2 payoffs.

In a repeated Prisoner's Dilemma, the choice is whether to cooperate (go for a mutual compromise in which both players get only their second best option) or to defect (choose the option which for this one round is most attractive to you, but most repugnant to the other player). Cooperation is the best strategy for both players

[21] For an accessible short introduction to game theory see Ross (2010).
[22] In game theory a strategy is called 'strongly dominant' when it is the best one to follow, regardless of the strategy followed by the opposition. In the PD under discussion, 'gas' is strongly dominant.
[23] According to Ainslie "following suit is both the most obvious strategy and the most successful one in repeated Prisoner's Dilemmas" (92). This is not entirely right – though it *is* perhaps more true where the players are locked in the same room. Sometimes in a RPD one can achieve long runs of mutual cooperation. Everyone does better on average than for any other strategy. (Pareto optimal). However, the best possible strategy is to be an unpunished occasional defector – which in the intrapersonal case boils down to the same as our previous description in terms of making exceptions on a personal rule without losing credibility. (I owe this observation to a conversation with my UKZN colleague David Spurrett).

(assuming a two person game) in the sense of being Pareto optimal - this means that no alternative distribution is possible which would make one player better off without making the other worse off.[24] But mutual cooperation is not a Nash equilibrium, that is, not rational to pursue in a one-shot Prisoner's Dilemma – a Nash equilibrium is per definition my best strategy whatever the other player does. Defection 'strongly dominates'[25] in a one shot Prisoner's Dilemma. In a Repeated Prisoner's Dilemma cooperation is a good strategy for me only if I can believe that the other player will cooperate. If the other player defects, instead, my pay-off is less than it would have been had I defected as well. What can give me a reason to expect the other player to cooperate? Only the history of past rounds of the game. If the other player has hitherto not cooperated I won't expect her to cooperate in this round either. And the more consistently she has cooperated, especially during recent rounds, the more I can expect her to keep on cooperating.

The RPD model encapsulates the role of precedent in the establishment of will. Only if earlier self P cooperates with later self Q does self Q have any reason to choose the lower 'cooperation' payoff for this round, instead of going for the higher 'defection' payoff. A single defection by Q has a higher payoff for Q than a single instance of cooperation. But if P then also defects in subsequent rounds, both P *and* Q are worse off than had they both cooperated.

The 'will' of an individual is like the 'will' of nations in World War II not to use poison gas. "This will is a bargaining situation, not an organ" (90). For Ainslie the logic of intertemporal bargaining is essentially that of a repeated Prisoner's Dilemma. Wherever we can interconnect choices in a repeated Prisoner's Dilemma pattern, this gives us a means of self-control – the repeated PD is just a way of getting out of the inevitability of the sooner smaller option dominating (94). In this bargaining process the only information exchanged is that found in the actual moves of the players.[26] In the RPD, as in most games, the two sides are in a situation of 'limited warfare': they have goals that clash (each would prefer to win the war), but also goals in common (conserving resources; not wanting to become the victim of gas warfare). Similarly "a person *today* wants to stay sober tomorrow night and *tomorrow night* will want to get drunk, but from neither standpoint does she want to become an alcoholic" (90). Here 'limited warfare' defines the relation between successive selves; its incentive structure is that of a Repeated Prisoner's Dilemma. This points the way to a strategy to allow cooperation in the area of common goals.

[24] In a Repeated Prisoner's Dilemma the cardinal values of the outcomes *are* important for the definition of the game. If two rounds of mutual cooperation are worth less than, or the same as, a seesaw consisting of one round cooperate/defect and one round defect/cooperate, the parties may settle into a pattern of seesawing rather than cooperating.
[25] See note 22.
[26] It can be shown that exchanging other information will have no effect on the strategy that a 'rational player', out to maximise personal reward, will follow. It is thus 'cheap talk'. Any information not contained in the moves, is irrelevant.

How does the area of common goals relate to the choice for the larger later option we talked about previously in connection with the will? Different interests compete for the allocation of the scarce resources of the individual they all inhabit: time, energy, money. But they all benefit from these resources being as copious as possible to begin with. Each wants as large a fraction of the cake as possible, but they all want the cake to be as large as possible before the slicing begins. Most of our interests thus share the goal of safeguarding our health and survival, and maintaining or increasing other resources such as the amount of money we have. (A gambling habit is parasitic on ways of making and saving money in the first place. So a gambling interest and a family interest may cooperate in opposing extravagant gifts for a potential mistress). Building and maintaining our basic resources, as a shared goal between diverse interests, will thus be a large part of the larger later rewards with which sooner smaller rewards have to compete.

We now see how Ainslie's two observations about hyperbolic discounting articulate with this discussion of Prisoner's Dilemmas. Summation in hyperbolic discounting is linked to expected rewards from continued cooperation in Prisoner's Dilemmas. Just as summation in hyperbolic discounting gives an escape route out of the trap of a dominant sooner smaller reward in the one-off case, so the repeated Prisoner's Dilemma gives us a route out of the trap of the dominant non-optimal strategy in the one-off Prisoner's Dilemma.[27, 28]

The two cases are however asymmetrical in some respects: given a sooner smaller reward A which dominates a larger later reward B in a one-off choice, summation of curves in combination with the precedent function of our current choice changes the incentive structure so that the choice for the larger later reward B (the series of larger later rewards of which B is a part) is secured. Given a repeated Prisoner's Dilemma between two interests, one of which prefers the sooner smaller option A, the other the larger later option B, the outcome secured will sometimes be the larger later option B, and sometimes a compromise between A and B – some other goal or set of goals shared between the interests concerned.

---

[27] As said before, a repeated PD is just a way to get out of the inevitability of the sooner smaller option A dominating. The RPD incentive structure does not itself lead to the option B dominating as a necessary result. If that were the case, B would be the single equilibrium. But a RPD has infinitely many equilibria. So the RPD is just a way of getting out of domination of A – the single equilibrium in a one off PD. An outside observer can't predict the outcome of a RPD, the way a game with a single equilibrium can be predicted to settle into the equilibrium solution. All that can happen is that a player tries to steer the outcome by making a move that *can be interpreted as an offer to cooperate*. It need not be. There is no more pressing logic to the situation than this 'can be interpreted as' logic. What is needed to get the result heading in the direction of cooperation is a context in which a particular move is likely to be read as signalling willingness to cooperate, rather than just the act of a sucker. (I owe this footnote to a discussion with Don Ross and David Spurrett).

[28] Let me add a caveat for the sake of completeness: as the rewards of future moves will be discounted hyperbolically, differences in the expected timing of future rounds of the game can lead to differences in one's current move.

Such a compromise will often serve long-term interests, especially as the number of negotiating interests increases (the two player Prisoner's Dilemma is a simplification to show the logic involved), and thus will probably embody larger later goals – let us call them C. Realising that the process leads to a *de facto* choice between sooner smaller option A and larger later option C, takes us back to our original *type* of choice, viz. between a sooner smaller reward and a larger later one.[29] The possibility of a compromise as an outcome to the process therefore does not change anything essential.

Although a later self cannot punish an earlier one for defecting, the loss of reward as a result of 'defection' has the same effects as a punishment. Future selves don't (and can't) retaliate, but "merely choose on the basis of revised estimates of contingent outcomes" (Ainslie, personal communication 2011). "The threat that weighs on your current self's choice in a repeated Prisoner's Dilemma is not literally retroactive retaliation by a future self, but the risk of losing your own current stake in the outcomes that future selves obtain" (93). "By its current choice each self makes an offer to future selves—that is, the interest dominant at that moment makes what is in effect an offer, which will affect how interests that are dominant in the future make their choices, and, importantly, which interests may become dominant at future times" (Ainslie, personal communication 2011).

Downsides of will

Parts I & II of *Breakdown* could seem to reinforce the traditional Western tendency, starting with the ancient Greeks, to see the will as an unmixed blessing. The trick of willpower – bundling choices and taking your current choice as a precedent for future choices – seems to offer the rational person a simple, reliable recipe for maximising long-range interests (146): " 'the more willpower, the better' " (156).

Part III of *Breakdown* is an extended argument against this view. We are endangered both by a lack of willpower – impulsivity – and an excess of it – compulsivity – and there can be no recipe telling us how to cultivate just enough of it. "There's no formula for rationality" (154).

According to Ainslie "suspicions of willpower" are a recent phenomenon (143). He takes the dangers of willpower to be "the chief target" (145) of Freudian as well as most post-Freudian schools of psychotherapy, and also finds an awareness of these dangers in certain theologians and "existentialist" philosophers like Kierkegaard (144). Ainslie is here not rehashing a familiar, generally available history of ideas; it took his reconceptualisation of the will to see a wide variety of thinkers who apparently don't address the same issue – the will and willpower – as part of a single debate. (Both sides – willpower's cheerleaders as well as its detractors – turn out to have something essential to contribute). What typically happens, is that phenomena like "compulsiveness", "lack of spontaneity" or "lack of appetite" are identified as

[29] Our 'larger later' option B can itself turn out to be the 'sooner smaller' of the two options B and C.

major problems, with no awareness that they are side effects of using willpower. "In a dangerous split of awareness" (146) they are thus deplored by the same people who otherwise consider willpower as something of which one can never have too much.

For me Ainslie's overcoming of this split awareness is one of his major achievements. He first details how pervasive and debilitating our impulsiveness is before it is checked by strategies like willpower. Next, he shows how easily the will breaks down – how "volatile" or "brittle" it is. Finally, he argues that a strong will can be "a net liability" (147), with side effects that are even worse than a lack of willpower.[30]

But what are these bad side effects? They can mostly be summarised as compulsiveness: *having* to follow some personal rule, even where this reduces reward. Ainslie distinguishes four basic forms of this compulsiveness.

Firstly, "[r]ules [can] overshadow goods-in-themselves" (147), often outside our awareness and control. We can go to great lengths to adhere to a personal rule, just so as not to set the precedent of breaking it, even where the rewards to be gained or lost have little intrinsic value.

Secondly, to save our prospects of self-control, we sometimes exempt whole areas of behaviour ("lapse districts" (149)) from our rule, areas where repeated lapses have led us to give up any attempt at self-control. (We need again not be aware of this process). Examples would be public speaking, losing our temper with bungling clerks, or binging on certain foodstuffs. Impulses which otherwise could have been occasions for strategies of resistance, can then come to seem irresistible.

In the third place, "[r]ules motivate misperception" (149).[31] The precedent of breaking our personal rule can undermine our confidence of sticking to the rule in future. There is thus a strong motive not to perceive, or to misperceive, instances where the rule is broken. "As a result, money disappears despite a strict budget, and people who 'eat like a bird' mysteriously gain weight" (150). Like Freud (Gouws & Cilliers 2001) and Bentham (1999), Ainslie (175-179) treats perception and belief as behaviours: they are not simply or completely governed by considerations of accuracy, but modifiable by reward.

Finally, a rule which resists short-range interests (impulses), need not necessarily serve our long-range interests. It can instead serve midrange ("compulsion range" (50; 151)) interests, which often clash with our long-range interests. Midrange interests, such as doing one's job well, losing weight, or saving money, are more

[30] Contemporary post-industrial societies, governed by "the logic of an increasingly comprehensive marketplace" (157), rely more and more on the regulation of behaviour by willpower instead of by social pressures. In Part III of *Breakdown* Ainslie elaborates this idea into a whole diagnosis of our current cultural situation.

[31] This weakens the case for the advantages (page 16, above) willpower was supposed to have over "manipulating one's attention" (76; page 16, above), one of the other three strategies for precommitment outlined by Ainslie, which is also said to distort perception (Ainslie 1992:144).

circumscribed, concrete and measurable than the long range ones of happiness and general well-being. The strong-willed workaholic is undermining his long-range marital and individual happiness; the strong-willed anorexic is good at preventing weight gain, but not at securing health; the miser isn't choosing for long-range happiness, either. Because mid-range interests are easier to monitor than long-range ones – it is easier to specify "bright lines" for them – it takes little for them to eclipse the latter, especially if you and your culture adore willpower. The common dichotomisation of time frames into the "immediate" and the "long term" may make it sound as if resisting the lure of the immediate must necessarily serve long range interests, but Ainslie argues that this is not the case.

An important ramification of the problem of compulsiveness associated with willpower is that "[a]n efficient will undermines appetite" (161). *Ceteris paribus*, the less build-up of appetite there is before a potentially rewarding 'stimulus' presents itself, the less rewarding it will in fact be. When we describe something as having been done "with gusto," this suggests an experience more rewarding than one done *without* gusto.

In affluent societies, our main sources of reward are emotional in nature. In emotional reward the element of surprise is crucial. If on page 300 of your whodunit you suddenly find out that the person who dunit is somebody you had never suspected, this moment of surprise is the moment of maximum emotional reward, a reward that is increased because your appetite for it has been building up over 300 pages. If on p. 295 it suddenly occurs to you in a flash who it is that dunit, *this* becomes your moment of maximal emotional reward, your appetite – reward potential – is now radically decreased, and the reward you experience from the passage on p. 300, if any, is now probably but a fraction of what it would otherwise have been. And suppose that you already grasped on p. 5 who it was that dunit, you would find this much less rewarding than had it only occurred to you on p. 295 or 300. An efficient will generating situations meant as occasions for reward is rather like already knowing on the fifth page what the book ostensibly only reveals on p. 300. Though the occasion for the reward – what loosely speaking we would often simply call "the reward" – can be exactly the same, the actual reward – the experienced reward, the satisfaction – is totally different, and much smaller. The size of the reward is limited because there has been insufficient build-up of appetite – the "problem of premature satiation".

Spontaneous, pre-reflexive learning – the automatic result of exposure to the contingencies of emotional reward (possibly aided by some degree of reflexive awareness of what is going on) – suffices to teach us, over time, to reserve emotional rewards for occasions that are sufficiently scarce, outside our control, and unpredictable.[32] Such occasions can *per definition* not be directly secured by an

[32] According to Ainslie (1989: 17), the gradual diminution of people's reliance on fantasy in favour of "the facts" as they mature, is not so much as Freud would have it due to an avoidance of harm, thanks to the "instinct of self-preservation", but mainly the result of learning that an optimal pacing of rewards

efficient will. Where they occur, they are experienced as directly *causing* the emotional satisfaction. However, emotions count as *behaviours* in Ainslie's scheme, because they are modifiable by reward. We can experience them while daydreaming. But the rewards of daydreaming remain limited because, like the rewards of an efficient will, they soon become too predictable, abundant, and subject to our control. This feeds the mistaken notion that emotions are *passions* overcoming us from outside, without our participation.

---

requires that their occasions be sufficiently outside our control. The pacing function of facts is served more or less as well by any *belief* about the facts that makes occasions for emotion sufficiently rare. This is a powerful idea with wide ranging ramifications.

Discussion

STRENGTHS OF AINSLIE'S THEORY

Reordering the field of knowledge. One of the most striking features – and, I would argue, virtues – of Ainslie's approach is the way in which it changes the topography of the field of knowledge, so that what seems similar or dissimilar, adjacent or far apart, compatible or incompatible, changes. In undermining some of my prejudices, many of them widely shared by other philosophers, Ainslie has in effect redrawn my map of how *disciplines* and *theories* like behaviourism, economics/utility theory and psychoanalysis relate to each other. He also reorders the field of *mental phenomena*: an example is the innovation of regarding undesirable phenomena like compulsiveness or lack of appetite (for life) as predictable effects of the cultivation of willpower – something which is usually treated as an unmixed blessing. In Chapter 4 Ainslie develops a bold conjecture: that compulsions, "addictions" (or impulses), "itches" (or urges) and pains, whose phenomenology would never suggest that they can or should be treated together, are actually manifestations of the same mechanism of temporary preference, but at radically different time scales.[33, 34] Here Ainslie the Ockhamite systematiser again attempts to reduce the number of disparate, brute facts, which lie outside the purview of the basic principles of behavioural psychology, as exemplified especially by the Matching Law. This is also the context in which he motivates his systematic distinction between "reward" and "pleasure".

Combination of grand theory with conceptual, argumentative and explanatory rigour. The foregoing examples highlight the ambitious and surprising nature of Ainslie's theory. Beyond the will, it reconceptualises the nature of motivation and mental life generally[35], and beyond that, of the good life, as well as the value of prudence and rationality. Ainslie is catholic in his choice of interlocutors – they range from Skinner to Freud, from rational choice theorists to Kierkegaard. But at the same time, he is a determinist (129-134) and a reductionist, a fearless wielder of Ockham's razor, trying to save as many phenomena as possible as parsimoniously as possible. Dyed in the wool modernists will applaud the elegance and systematicity of his theory. Dyed in the wool postmodernists will applaud his descriptive constructivism about beliefs (though wishing he had also been a normative constructivist). By the

[33] I did not discuss this aspect of *Breakdown* above, as its truth or falsity, heuristic value or lack of it, do not determine those of Ainslie's basic theory of will. It does, however, become important when he develops this basic theory further (Ainslie 2009; Ainslie 2010; Ainslie 2011, for instance).

[34] In *Picoeconomics* Ainslie (1992, especially 319-327) similarly develops a systematic account of the emotions, partly based on the way in which in each emotion reality agrees or disagrees with a desired state. It is almost like a sketch for a periodic table of the emotions, outlining a system of parallels and contrasts between the different emotions, rather than treating them as a random collection of incommensurable brute facts.

[35] Ainslie (2009: 112) seems to think that Hanson (2009) "in accepting the competition of hyperbolically motivated interests as the whole picture of psychic life," has grasped his intention well.

same token, dyed in the wool representatives of each camp will also find many, though different, parts of his approach objectionable.

Throughout, Ainslie gives copious arguments for his conclusions and against competing theories, often based on experimental and other empirical findings. Arguments based on such premises better lend themselves to intersubjectively compelling debate pro and contra than the intuitions often appealed to by non-naturalist philosophers. When things start to get counter-intuitive – and thus really interesting – a good way to make philosophical mare's nests tractable is often to bring the tools, criteria and findings of science to bear on them, as Ainslie does.

The most crucial premises Ainslie appeals to are those deriving from the foundations of mature theories or disciplines, having a clear conceptual and technical elaboration, and often quantitative in nature. This means that despite the originality of Ainslie's theory, its foundations are the opposite of idiosyncratic. The reader's investment in understanding *Breakdown* will thus bear dividends by making a large body of work by others more accessible. Nevertheless Ainslie often argues for tweaks to received theories. Utility theory (rational choice theory) is of course modified so that future discounting is hyperbolic, not exponential. The version of behaviourism Ainslie subscribes to also covers covert behaviour ("mental processes"), rejects "classical conditioning" as an *additional* motivational principle next to operant conditioning, and rejects blank-slatism.[36]

One of the great virtues of *Breakdown* is the self-reflexive methodological clarity and sophistication Ainslie brings to bear on his project. (An example is his argument why dynamic systems models resist direct empirical testing. This leads him to take an explicitly philosophical tack instead, and argue that his theory allows us to resolve some well-known thought experiments – paradoxes of intentionality, devised by philosophers of mind – and that this gives strong support for its correctness).

My foregoing remarks could all be true, and Ainslie's theory still be wrong, even badly wrong. However, even then this theory would probably have great value in paving the way for a better one. Because Ainslie is so explicit in his arguments and so up front about his foundational premises and methodology, and because he formulates many of his central claims in quantitative terms, his theory is eminently open to criticism, and thus improvement.

Why philosophers should pay attention to Ainslie's work. Ainslie has described himself as a philosopher at heart. According to what are generally taken to be the qualities of a good philosopher Ainslie comes out very well. (Many of Ainslie's virtues outlined above are also *philosophical* virtues, even for philosophers who are not naturalists).

[36] See Ainslie's words quoted in footnote 45 for some of the ways in which he thinks the slate is *not* blank.

Ainslie's *empirical* theory of the will at the same time offers a *conceptual clarification* of the will – and of free will and the emotions (Ainslie 1992:319-327) too, for that matter. I used to think that one can give a perfectly good conceptual analysis of mental and behavioural concepts without appealing to empirical theory or the results of empirical research. Having now read Ainslie's empirically oriented account, I can't see how a philosopher could improve on its *conceptual* analysis of the will without appealing to empirical, doubtless quantitative, theories, or experimental and other empirical findings from science. Seeing the power of behavioural psychology and utility theory in Ainslie's hands also made me realise how often critics of these approaches are attacking straw men. There must be something wrong with critiques of these approaches which deny them any such power.

I have previously (Gouws 2010, apropos of Ross 2005) expressed some reservations about a narrowly conceived naturalism. These reservations do not apply to Ainslie, as is evident from the sheer variety of writers he engages with. He finds valuable insights in theology, existentialism, social constructivist accounts of belief, psychoanalysis, the "clinical lore" of mental health practitioners, "the accumulated wisdom of the society" (Ainslie 1992: 295), and lastly in religion as offering both aids to the necessary indirection with which some goals need to be pursued (188), and guises to conceptualise the contingencies of personal rules and current choices as precedents (107-108). The theory obtained by engaging with this variety of sources nevertheless seems cut from whole cloth; it does not come over as an unsystematic patchwork of found conceptual objects.

Philosophers often appeal to Harry Frankfurt's (1971) theory of the will. To my mind Ainslie's and Frankfurt's theories are simply in different intellectual leagues. Compared to Ainslie, Frankfurt's account of the will is myopic, unsystematic, lacking in historical contextualisation, methodologically inarticulate and unsophisticated, *ad hoc* relative to any scientific study of human motivation, and scantily argued for. To an inarticulate folk (or folk-like) theory of mind Frankfurt adds some *ad hoc* flourishes to safeguard the view that there is a faculty of reason with a power of choice which is not subject to any normal motivational processes. Frankfurt reminds me of somebody trying to develop from scratch a theory of acceleration, unaided by any worked out scientific methodology, quantitative techniques, or existing framework of physics – exactly the sort of mistake philosophical naturalism aims to avoid.

For most philosophers, reflecting on the will has been linked to issues of free will, moral responsibility and moral blameworthiness. In two recent publications Ainslie (2009; 2011) expands his rather condensed treatment of free will in *Breakdown*, to address the philosopher's problem of free will directly and systematically. What struck me about these articles is the lucidity with which he traverses a conceptual minefield, apparently without detonating a single mine.

Central to his account is the "chaotic" (in the technical sense of the term) nature of recursive self-prediction. If hyperbolic discounting had only linear manifestations, it too would be unable to account for free will. However, Ainslie's account of free will intrinsically involves recursivity: our current choice modifies our prediction of our own future choices, which can then modify our current choice, which then … This way we actively participate in our own choice – we "make" it – while our own future choices can nevertheless be genuinely opaque to ourselves. (Ainslie sees such active participation and opacity as features any theory of free will has to account for). Determinism in a recursive system is thus compatible with unpredictability, even to the agent herself. "The ultimate causes pre-exist, but they have by no means completed their activity when they have entered the person's motivation. Their dynamic interaction during intertemporal bargaining is what initiates choices" (Ainslie 2011: 73).

As for blame, Ainslie bases other-blame on self-blame, instead of the other way round – the usual approach. Self-blame should not be seen as a separate act, following upon self-perception. We blame ourselves when we perceive (obscurely, usually) that a choice we made has decreased our credibility to ourselves, and thus our potential for self-control. Self-blame is thus in essence "the perception of a loss that has already happened" (Ainslie 2011: 76) – the loss of prospects of future reward. Even if it originates in the internalisation of control by parental authority, it is maintained by its functional role in self-control (Ainslie 2011: 74ff). Self-blame is "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" (Ainslie 2011: 81). If the loss of self-trust is sufficiently large, it can incapacitate one radically, but in a way that is better modelled by bankruptcy than pathology (Ainslie 2011: 80).

Other-blame is then an empathic extension of self-blame (Ainslie 2009: 109). We blame or exculpate others when we would have blamed or exculpated ourselves in similar situations. This again is based on a perception of loss, now of a social nature – something like a loss of the trust needed for beneficial interactions with others.

He does not think that applying this account to questions of blame and responsibility will allow us to discern a bright line, a simple criterion demarcating blameworthy behaviours from others (Ainslie 2009: 110-112). The typical philosopher's demand for such a digitisation (Ainslie 2009: 95), sits badly with the analog nature of the distinctions involved in assigning responsibility or blame.

RESERVATIONS ABOUT AINSLIE'S THEORY

Can one ever fully master the theory?

To make my point here simply, I resort to a bit of autobiography. When I first discovered *Breakdown*, I was deeply impressed by the elegance and power of Ainslie's theory, and its promise of allowing numerous interesting applications to new fields. Looking back, at this time I probably envisaged mastering the theory

sufficiently in a year or two to be able to apply it to topics not treated or only sketchily treated in *Breakdown* itself.

It is now many years later, and my mastery is still incomplete, for instance when it comes to the notions of successive selves, and of interests negotiating with each other. (Hanson (2009) also struggled with this part of picoeconomics, judging by Ainslie's (2009) response). As always, it would have helped had I been smarter, and less set in my intellectual ways. Ainslie writes very clearly, so the problem does not seem to lie there. Nor did I skimp on exegetical effort; while repeatedly teaching the book, I carefully mapped many of Ainslie's arguments – an exercise which only made them more impressive to me. Rather than teaching Ainslie's book, I would have preferred being a student in *the* perfect course on it – perfect in that it would leave me with a complete mastery of Ainslie's theory and the scientific theories and experiments that form its background. Such a course would doubtless also have involved spending more time than I in fact did on mastering behavioural psychology and utility theory.

However, there are also aspects of the theory itself which make it hard to master. It is very much work in progress; its different parts have not all been spelled out or formalised to the same extent. (Somebody trying to write a toy computer model of it would have far less work to do on some parts than on others). This is not a gripe on my part. Rachlin (2005:658) however *does* make the complaint – or is it a back-handed compliment? – that "some of [Ainslie's] discussion takes the form of a literary essay (albeit finely wrought) rather than a scientific analysis" – for instance Ainslie's discussion of indirection (187-196). However, I think it is entirely to be expected that the different parts of a theory like Ainslie's will not all be equally explicitly spelled out and connected equally directly to experimental findings. In that case one wouldn't be able to master all the parts of the theory in the same way.

My dream of mastering Ainslie's theory was thus perhaps misguided from the start. Moreover, I perhaps envisaged approaching the theory too much like a traditional philosopher, for whom coherence and conceptual clarity are the prime criteria a theory should meet. However, as is always the case in cognitive science, to engage with this theory is to engage in the hand to hand combat required to situate and evaluate it vis à vis the ever expanding body of relevant experimental findings and scientific debates.

The difficulty and novelty of Ainslie's theory may explain why it has not had the impact it deserves. Admittedly, several writers have expressed admiration and even astonishment at his achievement (e.g. Dennett 2003: 207-213, Ross 2005, Ross et al 2008, Sanabria & Killeen 2005; Stanovich 2005). But rare is the philosopher or other academic who knows about his work, let alone shares my view of how important and exciting it is. Nor do I know of any careful attempt to refute his ideas – another way to acknowledge their importance. Neither, to my surprise, has anybody yet picked up Ainslie's rewriting of psychoanalysis and run with it.

Should Ainslie's theory have been presented differently?

I next address a related issue, this time regarding the systematicity of Ainslie's account, rather than its accessibility. Methinks that Ainslie should in future present his ideas in a different format, so that he makes preference instability – or curve-crossing – somewhat more central, and the hyperbolic nature of future discounting somewhat less so.

He has a mass of *evidence* on his side when he maintains that vertebrates are hyperbolic future discounters, and that this is the default mode of future discounting in humans as well. He also has *logic* and *maths* on his side when he maintains that this in itself will lead to instability of preference (or curve-crossing). Finally, his model of will in terms of personal rules (bundling and the present choice as precedent) is powerful and elegant. However, a considerable literature, some of which I will refer to below, indicates that other factors also contribute to instability of preference. Ainslie does acknowledge several such factors, but when he does, it is usually only as an afterthought, and often (e.g. 30; Ainslie 2005:665) immediately accompanied by an energetic affirmation that hyperbolic discounting on its own would predict preference instability, as if this warrants disregarding other sources of such instability.

I propose that in future accounts of his theory Ainslie include these other sources of preference reversals from the start, rather than adding them as an afterthought. (To what extent the proposed modification in presentation would entail a modification in substance, I cannot say). The modified account would go something like this:

Standard utility theory assumes stability of preference, in the absence of changes in the agent's information or budget, but many findings in behavioural economics contradict this. One of these findings is that the default future discounting function of humans (and other vertebrates) is hyperbolic, not exponential. [This is probably the most important one, but perhaps it isn't, and if it is, Ainslie should explain *why* it is]. Other studies indicate further factors which can independently lead to curve crossing even in an exponential future discounter, or can exacerbate curve crossing in a hyperbolic future discounter, for instance that

- *risk* discounting, like future discounting, is hyperbolic (Rachlin, Brown and Jay 2000) – this would lead to preference instability even if all goods were risk discounted at the same rate, but in fact
- the *rate of risk discounting* is different for different goods – a tautology, almost, but nevertheless an independent source of time-linked preference reversals
- different goods are *future* discounted at different rates[37]

[37] Zauberman & Lynch (2005) for instance argue that time and money are not discounted at the same rate.

- the rate at which the future is discounted can fluctuate in response to cues or circumstances[38]
- as buyers experimental subjects prefer high likelihood low amount lottery tickets to low likelihood high amount lottery tickets with the same utility according to standard theory, but as sellers prefer the opposite (Ross 2005:177). Ainslie is probably correct not to find this factor very significant as a contributor to preference instability.[39, 40]

In addition, the account would tell us which of these forms of preference instability can rightfully be considered to be instances of weakness of will. Ainslie could perhaps, as Ross (2005) thinks possible, argue that some of the factors which I present as *extra* sources of preference instability are in fact just consequences of the hyperbolic nature of our default future discounting function.[41]

Next Ainslie would tell us which of these forms of preference instability could be stabilised by willpower – the combination of bundling and considering the current choice as a precedent.[42] It is conceivable that willpower could here too stabilise

[38] Wilson and Daley (2003) for instance found that men discount the future more steeply after being exposed to photographs of attractive women, while Argo and Levav (2010) found that both men and women were more inclined to take financial risks – i.e., discount the future more steeply – after being touched subtly on the shoulder by a female. Phenomena such as these should also lead to more curve-crossing than hyperbolic discounting at a static, and lower, discount rate.

[39] The reason Ainslie does not find it very significant is that the relevant findings are far from robust – if our quantifications of values were somewhat less precise, this phenomenon would no longer be detectable (Ainslie, personal communication 2011). Ross (2005:184), in contrast, regards this as the "most important phenomen[on] discovered by behavioural economics," after hyperbolic discounting.

[40] I initially thought that the foregoing list should include the finding that humans discount smaller amounts more steeply than larger ones (Green & Myerson 2005:655; Ainslie 2005:665). However, after further reflection I now believe that an amount-linked difference in steepness of future discounting like this could only have constituted an independent source of preference reversals if the difference had gone in the opposite direction, that is, only if larger amounts had been more steeply discounted than smaller ones. As it is, this phenomenon would rather make a hyperbolic future discounter *less* inclined to preference reversals, as it opposes the very tendency which constitutes the hyperbolic future discounter's Achilles heel, namely the tendency to overvalue sooner smaller rewards relative to larger later ones. Moreover, in a recent publication, Ainslie (2010: 234) points out that the magnitude effect does not apply when both amounts are large, and, crucially, "has been reported only for amounts differing by a factor on the order of ten or more". Given such big differences, the preference for the larger reward is likely to be robust. In light of all these considerations this finding does not belong in this list.

[41] Ainslie (2005:665) seems to think that it is "the action of committing maneuvers on undiminished impulses" that lowers some people's future discounting rate, often quite radically. (He gives an elegant argument why we are unable to "bend the function itself" (Ainslie 2005:665) – if we were, we would be able to "coin value" ad lib). If so, this could be a or the source of the differences in people's discount rates for different goods, as some things (money, especially) better lend themselves to the will and other precommitment devices, and in the case of "lapse districts" people have moreover completely stopped trying to apply willpower to certain areas, and so lost out on the effect whereby willpower decreases the steepness of future discounting.

[42] Ross (2005), while setting great store by Ainslie's theory of the will, sees the self – to him the product of self-narration – as the most important way in which preferences are stabilised, so that people become predictable to themselves and others. Should self-narration thus be treated as a *fifth* commitment device to complement Ainslie's four (p. 16, above)? Ainslie (personal communication 2011) responds: "I count self-narration, or Bodner & Prelec's (1995) 'self-signaling,' as a form of

preference because many or most of the other factors listed above "would lead to the same limited warfare relationship among successive motivational states, which is the basis of intertemporal bargaining theory", to quote Ainslie's (2005:665) words in responding to an unrelated claim by Bach (2005).

Ainslie's modified account of his theory would also explicitly relate two parallel answers to the question why people are impulsive, linked to two different, but related, ways of defining 'impulsivity' (Ainslie 2009: 101). He is quite up front in positing the existence of impulsivity in both these senses, and they aren't mutually exclusive, but it would be clearer to treat them together systematically. The one answer says that people are impulsive$_1$ because their default discount function is hyperbolic; some people are more impulsive$_1$ than others because they are less good at the will and other precommitment devices. The other answer says that the more steeply people discount the future, the more impulsive$_2$ they are – "degree of impulsivity$_2$" is then directly proportional to, or synonymous with, "steepness of future discounting". (The notion of impulsivity$_2$ is also found in writers who don't treat future discounting as hyperbolic). Fleshing out the latter answer from Ainslie's perspective would presumably involve two aspects: in the first place, a will-less hyperbolic discounter A who discounts the future more steeply than another will-less hyperbolic discounter B, will *ceteris paribus* have more preference reversals (curve crossings) than B. In the second place, for people who do have a will, steeper discounters will need a longer (or wider or deeper)[43] series of bundled choices than for the shallower discounter before summation eliminates curve-crossing (see Figure 3 on p. 17, above), and thus stabilises preference. This means that the steeper future discounting is, the bigger the difference between a hyperbolic discounter and an exponential one.[44]

"*Seek simplicity, and distrust it"*

I have in the course of this article repeatedly praised Ainslie's Ockhamite striving for simplicity. There is a lot to be said for such a striving; in humankind's greatest scientific achievements an undreamt of simplicity is achieved by uniting under one law phenomena so diverse that they would on the face of it seem impossible to unite. Let me however express a general Wittgensteinian worry. The later Wittgenstein tirelessly warns us of the *dangers* of the desire for simplicity. Writing before Wittgenstein's later philosophy, Whitehead (1926:163) gives both Ockham and the Wittgenstein yet to be their due in his words: "Seek simplicity, and distrust it." (The words commonly – but erroneously (Sterling 2011: 57) – attributed

---

reward bundling. A narrative somewhat depends on your behaviour; to the extent that it is threatened by your current choice, it is the bundle that's at stake."

[43] Re a "wider" series: "Lapsed alcoholics notoriously try to recover by pledging more kinds of virtue, hoping to overcome the decrease in believability with volume".
Re a "deeper" series: "the wisdom of AA might be called creating a deeper series—Saying one is 'helpless' against alcohol is to give up the hope of negotiating some exceptions, the way one does when 'using willpower.' " (Ainslie, personal communication 2011).

[44] Ainslie seems to think that commitment strategies in effect make you discount the future less steeply. If so, this would be one of the reasons why willpower makes your behaviour when a sooner smaller reward is imminent more like that of an exponential discounter.

to Ockham in any case do not say that entities are not to be multiplied, but only that this should not happen *except if necessary*). One of the tasks of Ainslie's critics or friends is to assess whether accuracy demands that his theory become a bit less simple. It would have been simpler if all preference reversals were the result of hyperbolic discounting. However, if the more complicated picture that arises when we see preference instability as having multiple sources fits the evidence better, accuracy will trump simplicity.[45]

SO WHAT? WHAT CAN ONE DO *WITH AINSLIE'S THEORY?*

In the foregoing I gave some suggestions regarding where Ainslie's account of his theory needs to be modified. Given some hypothetical true future theory of the same field, I do not expect Ainslie's theory to be so definitively refuted that it will in retrospect not even have been a quick and dirty first approximation to the truth. Questions of its truth or approximation to truth will ultimately only be answered on the basis of on-going debates referring to empirical research, computer models (possibly) and alternative hypotheses. An article like this cannot take the place of this on-going reception of the theory by the scientific and philosophical community.

Assuming that Ainslie's theory will have a good enough track record over time, another important question to ask, is: what can we *do* with it? What light does it cast beyond its own (apparently) circumscribed domain? To round off this article I explore two possible applications of Ainslie's theory and general approach to issues which seem to lie outside narrower questions about willpower: the foundations of psychoanalysis, and making sense of Bataille.

Psychoanalysis. Above I related Ainslie's model to psychoanalysis repeatedly. Here I want to say something more general about how the two approaches can be connected.

As a philosopher I have long tried to frame or interpret psychoanalysis in a way that would supply a rationale for *not* dismissing its on-going intellectual importance, despite the plethora of arguments in the literature for writing it off completely on philosophical, scientific or methodological grounds. Over the years I investigated a number of approaches, many of them fairly convoluted. In the end I did not find any of them very satisfactory, so that I was left without a rejoinder to those who thought it obtuse of me to *still not* dismiss psychoanalysis.

Today my philosophical orientation has changed to an unconvoluted form of naturalism. To the extent that psychoanalysis contributes to knowledge, the extent that

[45] Ainslie (personal communication, 2011) is worried that my account oversimplifies his position by making the Matching Law the sole determinant of motivation: "I acknowledge many other motivational factors, e.g. preparedness (inborn grooves in the blank slate), absorptiveness (inborn tendency to focus attention), and social sensitivity (ability to use, or inability not to use, pressure from peers). The simplicity is that all these factors have to have their effects in a single marketplace of motivation."

there is truth or an approximation to truth in what it has to say, *sooner or later* there will have to be consilience with the rest of the body of our most justified beliefs. To be true *is* to chime with the rest of what is true. What that means is that whatever there is in psychoanalysis that *in the long run* just cannot be fitted into our best science, cannot be true. Now according to this criterion a lot of psychoanalysis could be true even if it were to contradict just about everything else we currently *believe* to be true, also regarding the results of science. Explaining how this could be the case would again probably be very convoluted – a claim for instance that all of our current so-called science is just a systematic body of untruths, which is not a view I endorse.

Naturalists tend to be sceptical or even dismissive about psychoanalysis, so that naturalism as a philosophical home for my sympathetic probing of psychoanalysis is hardly any more comfortable than any of the other philosophical approaches I tried in the past. Before I encountered the work of Ainslie I had come to doubt that one could ever justify taking psychoanalysis seriously from a hardnosed naturalist perspective, or philosophy of science perspective – a conclusion which, if true, psychoanalysis should find worrying.[46]

Enter Ainslie, and a surprising new way to vindicate the continued intellectual and scientific relevance of psychoanalysis presents itself. Ainslie has suggested a new metapsychology for psychoanalysis, meaning that it is in principle compatible with his amalgam of (tweaked) utility theory and (tweaked) behavioural theory. This leaves the detail – and much more than just the detail – up for grabs; a lot of psychoanalysis is doubtlessly wrong, even badly wrong – something which had always been part of the equation for me. The first step in the naturalist rereading of Freud for me thus would be: what happens when we replace the appropriate pieces of the Freudian edifice – especially the metapsychology – with pieces of Ainslie's behaviourist/behavioural economic amalgam?

Ainslie's model may help reunite psychoanalysis with empirical science, which would fit well with Freud's own style of theorising, but not with most of what happens in psychoanalysis as it is currently institutionalised.

Breakdown contains scattered references to Freud, but for Ainslie's systematic engagement with Freud one has to go to earlier texts (Ainslie 1989; 1982; 1984; 1992). As part of his suggested new metapsychology for psychoanalysis, Ainslie gives a new reading of several central psychoanalytic concepts, such as the ego (Ainslie 1992:328-334) and the mechanisms of defence (Ainslie 1982; 1984; 1992: 128-9, 142, 175, 187, 190-192, 206-207, 174-179).

[46] A seeming exception was the detailed parallelism between Freud's *Project* (Freud 1975/1895) and the field of computing called parallel distributed processing (PDP), which Paul Cilliers (1990) and I explored (Gouws and Cilliers 2001). The problem with this is that the brain does not *in fact* work *just like* a conventional PDP network, i.a. because of the essential role played by neurotransmitters and hormones in its functioning. However, this role can perhaps be modelled in a PDP network by global shifts in weightings – a possibility that we did not explore.

Ainslie's theory combines behavioural psychology and utility theory, and this link to the utilitarian tradition makes its relevance to Freud less surprising. Freud's thought had deep roots in utilitarianism – he was for instance the German translator of some of John Stuart Mill's works, and the many parallels between Freud's thought and Bentham's have been pointed out by Watson (1958). Ross (1999) convincingly argues that today's mainstream economics is the descendant, and explicit mathematical elaboration, of the fundamental quantitative intuitions found in Bentham. (This mathematical elaboration was not a trivial task; it took generations of economists to devise, one laborious step at a time, strategies for refining Bentham's concepts in such a way that his programme was realised via a mathematical apparatus). Psychoanalysis and modern utility theory thus share a common ancestor historically preceding the founding fathers each officially recognizes.

Ainslie's proposed new metapsychology for psychoanalysis again takes seriously the quantitative model of motivation that Freud often explicitly espoused (Freud 1975/1895; Gouws and Cilliers 2001), but that elsewhere could seem little more than a metaphor. Many commentators have seen this quantitative model as an embarrassing nineteenth century mechanistic holdover, which will be overcome when psychoanalysis is reconceptualised as an exclusively hermeneutic discipline. However, behavioural psychology and utility theory have shown the conceptual and empirical power of quantitative models, based moreover on motivational pulls rather than mechanical pushes – while being as uncompromising in their determinism as Freud.

This suggests that such a quantitative model may be the promising future of psychoanalysis, rather than its discredited past. A two way exchange between psychoanalysis and utility theory would thus become possible: psychoanalysis could draw on the example and results of a century and a half's worth of conceptual and mathematical elaboration of Bentham's programme[47], while psychoanalytic perspectives could leaven behavioural economics by suggesting further ways in which actual human behaviour and motivation departs from what standard economics predicts – or by suggesting ways in which the downstream implications of (a slightly, rather than revolutionarily modified) economics are different from what one would expect. Freud, who did not have a numerically inclined mind, did not go far in the direction of a mathematical formulation of his metapsychology. Similarly Bentham, the father of utilitarianism, did not go far in turning his quantitative intuitions into a

[47] When Freud discusses the libido his language is perhaps closer to that of economists than anywhere else (Ainslie 1989: 13). A good example is his treatment of different bodily parts or bodily actions as substitutable for each other. I have previously argued (Gouws 1998) that Freud is unable to give us a criterion whether something is sexual – in his extended sense of the term – or not. If I am correct, one possible conclusion is that, rather than there being a specific libidinal economy, which is distinct from the general motivational economy, any libidinal economy would be part of a generalised motivational economy. (This would chime well with one of Ainslie's central contentions: the claim that there is a single dimension of reward). As such the lack of a criterion for distinguishing between the sexual and the non-sexual would cease being a problem. By the same token, however, in controversial cases it would not be possible to insist that something either *was*, or was *not*, sexual, as Freud often wants to.

mathematical formulation of the principles of the utilitarian calculus; this was something others had to do. The example of utility theory suggests that giving even the soundest quantitative intuitions a workable mathematical elaboration may require several generations – and even then "work in progress" signs will need to be posted everywhere.

As in most attempts at unification in the sciences, it is unlikely that an Ainsliean rereading of psychoanalysis would do no more than vindicate an existing variant of the target theory. (Such an outcome would also contribute less to the growth of knowledge). The likely outcome of a good rereading of psychoanalysis is that some aspects of it will be vindicated, and some aspects justly rejected or revised.

Both Ainslie's critique of an equation of reward with pleasure, and the way in which economics developed out of *Bentham's* embryonic programme, suggest that the hedonistic conceptualisation of *all* motivation in terms of pain and pleasure in *Freud's* embryonic programme would be replaced by more abstract categories, like "Quantity" in Freud's (1975/1898) own "Project", "utility"[48] in utility theory, or "reward" in Ainslie. Freud's rationale for supplementing his pleasure principle with the death drive would thus be (partly?) undermined. (See note 4).

Bataille. Ainslie does not refer to the work of the French thinker Georges Bataille (1897-1962). However, Bataille is a natural – if extreme – extension of a series of thinkers (e.g. existentialists and theologians) to whom Ainslie *does* refer in discussing the downsides of will. These thinkers invariably come from the 'Romantic' tradition, a term I use loosely to indicate the varied and diffuse movement that arose in reaction to the Enlightenment.

Bataille (1998; 1995) admits that because of scarcity humankind will never be able to dispense with work, thrift and prudence. Nevertheless, he rails against the narrowness of the bourgeois life – a life lived according to the dictates of work, thrift and prudence, in which expenditure is only acceptable if it is productive. In its stead he celebrates a life of loss, reckless expenditure, gambles, surprises and intense experiences. He argues for the subsumption of conventional economics under a more general economics which recognises the value, and even necessity, of the things he celebrates as "expenditure without reserve". Among the phenomena Bataille is especially fascinated by are sacrifice, potlatch and other forms of overtly or covertly courted loss. Such phenomena seem anomalous to a utility theory which assumes that an agent striving to maximise gain will actively avoid such 'unproductive' forms of loss.

Ainslie's concepts and theses have a very different flavour from Bataille's, and it could seem that it will never be possible to make the two thinkers part of the same conversation. However, what Ainslie does with his concepts turns out to be oddly

[48] "Utility" is a *subjective* concept of value (Ross 1999:4): *anything* an agent strives to obtain; as a term of art in economics its everyday meaning of "usefulness" is irrelevant.

reminiscent of Bataille. Willpower is indispensable. It makes us more prudent. Thanks to willpower we are better able to work, as well as to save money and other resources. But it can easily become a curse. The relentless systematisation of life by the will (reminiscent of the relentless systematisation of life by work in Bataille) needs to be undermined if life is to become surprising again, and we are to reap the rewards of intense emotional experiences.

For Ainslie (161) "an efficient will undermines appetite". He sees appetite itself as a good – an increase in appetite is an increase in reward potential, and without appetite there is *no* reward potential. He quotes Benjamin Franklin: "The poor man must walk to get meat for his stomach, the rich man a stomach for his meat" (161). Franklin's words suggest that appetite becomes the more important as it becomes scarcer in affluent societies. To maximise reward, it is insufficient to maximise access to or control over goods that can be rewarding – reward is limited by available appetite. Therefore a major part of maximising reward will concern how appetite is managed or "paced". There will for instance be less appetite, and thus reward potential, if *ceteris paribus* a good is consumed too soon after a previous satiation. This goes for food, sex, and experiences of every sort.

These notions suggest a partial explanation for why societies and individuals pursue or court loss in the ways which fascinate Bataille: loss regenerates appetite.

The ideally rational person of standard economic theory will avoid loss, the build-up of unsatisfied appetite, gambling, etc. She will treat unsatisfied appetite as negative, not a good. Her ideal life would be one of perfect comfort, the predictable and immediate satisfaction of even the tiniest appetite.

However, a life without toil, struggle, the overcoming of obstacles, surprises, setbacks or challenges would be so short on appetite as to be utterly boring. So *ceteris paribus* the stronger and more successful your will is, and the more things go according to plan, the more boring your life is likely to be. Where things always go according to plan, extra instances of this pattern do not have the status of *events* – "nothing happens"; there is no *story* to tell.

Losing money (or other goods) by gambling may be a way for some people to remedy the limitations of the will – a way to refresh their jaded appetites, and thereby increase their reward potential. According to Ainslie, gambles in the broader, metaphoric sense are essential to pacing appetite, even if we don't go for gambling in the casino or lottery sense. In general, the emotional rewards involved in interacting with other people depend on precisely such gambles.

In a further parallel with Bataille, Ainslie argues that standard utility theory, which assumes exponential discounting, must be seen as a special, limiting case of a more generalised economic approach, in which hyperbolic discounting is the default option.

I have long been intrigued by Bataille's writings, which deeply influenced a whole generation of post-60s French thinkers, but simultaneously irritated by them because I could never discover method in their seeming madness – Bataille's conceptualisation and explanation of the phenomena he investigates to me never seemed compelling. Ainslie's reflections on the downsides of the will and the advantages of undermining the systematisation of life by the will, together with Ross's (2005; compare Gouws 2010) rereading of neoclassical economics[49] and a good dollop of evolutionary thinking[50], suggest just such method. "Method" here being: consilience with current neoclassical economics, behavioural economics and evolutionary psychology; and "madness" the "general economy" which Bataille thinks is needed to account for the phenomena of "expenditure without reserve", which seem so incompatible with standard economics – "restricted economy". Bataille may also suggest ways in which Ainslie could extend his own thinking on the downsides of the will.

In closing

The ultimate conceptual and empirical building blocks of Ainslie's theory are fairly simple and largely familiar to those versed in behavioural psychology and utility theory. However, in his hands their downstream consequences turn out to be highly surprising: we seem to move seamlessly from irrationality and impulsivity to the will, free will, and the downsides of the will; from pigeons pecking keys to some of the most complex human behaviour and experience imaginable, all without needing to invoke any miracles or deny determinism. This I find a formidable scientific, philosophical and stylistic achievement.

I refer the reader to Ainslie's own writings, which contain a wealth of material which has either been skipped or just skimmed over here. The reader will there find Ainslie's basic ideas applied to further topics and illuminated by a greater number of examples – and, I expect, be as impressed as I am by the consistent lucidity with which he addresses each of his topics.

Acknowledgements

[49] Ross, like Ainslie, denies that economics is exclusively or even predominantly about material goods. That gift-giving, which for Bataille is a prime example of a phenomenon defying standard economics, increases status and creates obligation (as Bataille himself recognises) can thus easily be accommodated by conventional economics. Moreover, where gift-giving *does* spring from simple disinterested altruism, this is not a problem either, as nothing in the technical apparatus of economics assumes that economic agents are egoistic.

[50] From an evolutionary perspective reproductive success trumps personal survival – 'inclusive fitness' is the name of the game. From this perspective demonstrating one's ability to afford extravagant expenditures of resources functions as a form of "expensive signalling" of one's fitness to potential mates and rivals. Examples are the peacock's tail and using a $100 bill to light a cigar. Such displays of "unproductive loss" can thus 'make economic sense', even though they wouldn't if personal survival and flourishing were all. In the bigger – reproductive or evolutionary – context the loss involved would then not be "unproductive" after all.

The final editing of this article was mostly done during a stay at STIAS, the Stellenbosch Institute for Advanced Study, which provided a good foretaste of the heaven which presumably awaits intellectuals who have led virtuous lives in their earthly existence. In writing this article I benefited greatly from correspondence and discussions with George Ainslie, as well as with my colleagues in the then School of Philosophy and Ethics at the University of KwaZulu-Natal, Durban Campus, especially Julia Clare, John Collier, Deepak Mistrey and David Spurrett. They graciously put their time and expertise at my disposal, way beyond the call of duty. I thank them all. Much as I regret it, I am well aware that I alone am to blame for the remaining defects of this article.

References

Ainslie, George. 1982. A behavioral economic approach to the defense mechanisms: Freud's energy theory revisited. *Social Science Information* **21**, 735-779.

Ainslie, George. 1984. Behavioral economics II: motivated, involuntary behaviour. *Social Science Information* **23**, 247-274.

Ainslie, George. 1985. Behaviour is what can be reinforced. *Behavioral and Brain Sciences* **8**, 53-54.

Ainslie, George. 1989. Freud and picoeconomics. *Behaviorism*, **17,** 11-19.

Ainslie, George. 1992. *Picoeconomics: The strategic interaction of successive motivational states within the person.* Cambridge: Cambridge Univ. Pr.

Ainslie, George. 2001. *Breakdown of will*. Cambridge: Cambridge Univ. Pr.

Ainslie, George. 2005. Précis of Breakdown of will. *Behavioral and Brain Sciences* **28**, 635-650.

Ainslie, George. 2009. Responsibility in a reductionist model, in Craig Hanson & George Ainslie: *Thinking about addiction: hyperbolic discounting and responsible agency.* Amsterdam: Rodopi, 95-113.

Ainslie, George. 2010. The core process in addictions and other impulses: Hyperbolic discounting versus conditioning and cognitive framing, in: Don Ross, Harold Kincaid, David Spurrett, & Peter Collins (Eds.): *What Is Addiction.* Cambridge, Mass.: MIT Press, 211-246.

Ainslie, George. 2011. Free will as recursive self-prediction: Does a deterministic mechanism reduce responsibility? in George Graham & Jeffrey Poland (Eds.) *Addiction and responsibility.* Cambridge, Mass.: MIT Press,

Axelrod, R. 1990. *The evolution of co-operation*. Harmondsworth: Penguin.

Baars, B.J. 1986. *The cognitive revolution in psychology.* New York: Guilford.

Bach, K. 2005. Three other motivational factors. *Behavioral and Brain Sciences* **28**, 651-652.

Bataille, G. 1995. *The accursed share.* (Vol. II & III). Robert Hurley (Trans.). New York: Zone Books.

Bataille, G. 1998. *The accursed share.* (Vol. I). Robert Hurley (Trans.). New York: Zone Books.

Bentham, J. 1999. The psychology of economic man, in Don Ross: *What people want: The concept of utility from Bentham to game theory*. Cape Town: University of Cape Town Press, pp. 34-57.

Bodner, R. & Prelec, D. 1995. The diagnostic value of actions in a self-signaling model. Paper delivered at the Norwegian Research Council Working Group on Addiction, Oslo, Norway, May 26, 1995.

Cilliers, F.P. 1990. The brain, the mental apparatus and the text: A post-structural neuropsychology. *S.Afr.J.Philos.* **9** (1):1-8.

Delacroix, E. 2009. Quoted in http://www.rogersandall.com/category/the-arts/. Accessed December 4, 2009.

Dennett, D. 1981. Why the law of effect will not go away, in *Brainstorms: philosophical essays on mind and psychology*. Cambridge, Mass.: MIT.

Dennett, D. 1991. *Consciousness explained.* Boston, Mass.: Little Brown.

Dennett, D. 1996. *Kinds of minds: Towards an understanding of consciousness.* New York: Basic Books.

Dennett, D. 2003. *Freedom evolves.* New York: Viking.

Ferrera, L. 2005. The will: Interpersonal bargaining versus intrapersonal prediction. *Behavioral and Brain Sciences* **28**, 654-655.

Frankfurt, H. 1971. Freedom of the will and the concept of a person. *J. Philosophy* **68**: 5-20.

Freud, Sigmund. 1905. Three essays on the theory of sexuality. *The standard edition of the complete psychological works of Sigmund Freud.* James Strachey (Trans.). London: The Hogarth Press and the Institute of Psycho-Analysis, VII: 123-243.

Freud, Sigmund. 1920. Beyond the pleasure principle. *The standard edition of the complete psychological works of Sigmund Freud.* James Strachey (Trans.). London: The Hogarth Press and the Institute of Psycho-Analysis, XVIII: 1-64.

Gailliot, M.T., Baumeister, R.F, DeWall, C.N., Maner, J.K., Plant, E.A., Tice, D.M., Brewer, L.E. & Schmeichel, B.J. 2007. Self-control relies on glucose as a limited

energy source: Willpower is more than a metaphor. *J. Personality & Social Psychology* **92**: 325-336.

Gouws, A.S & Cilliers, F.P. 2001. Freud's "Project", distributed systems, and solipsism. *S.Afr.J. Philosophy* **20**(3): 1-21.

Gouws, A.S. 2010. Setting the scientistic cat among the humanist pigeons. Review article of Don Ross: *Economic theory and cognitive science: Microexplanation. S.Afr.J. Philosophy* **29** (1), 28-56.

Grace, R. C. 1994. A contextual model of concurrent chains choice. *J. Experimental Analysis of Behavior* **61**:113–29.

Green, L., Fry, A. & Myerson, J. 1994. Discounting of delayed rewards: A lifespan comparison. *Psychological Science* **5**:33–36.

Green, L. & Myerson, J. 2005. Hyperbola-like discounting, impulsivity, and the analysis of will. *Behavioral and Brain Sciences* **28**, 655-656.

Hanson, C. 2009. *Thinking about addiction: hyperbolic discounting and responsible agency.* Amsterdam: Rodopi.

Herrnstein, R.J. 1961. Relative and absolute strengths of response as a function of frequency of reinforcement. *Journal of the Experimental Analysis of Behavior.* **4**, 267-272.

Kirby, K. N. 1997. Bidding on the future: Evidence against normative discounting of delayed rewards. *J. Experimental Psychol.*: *General* **126**:54–70.

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

Knoch, D. & Fehr, E. 2007. Resisting the power of temptations: The right prefrontal cortex and self-control. *Ann. N.Y. Acad. Sci.* 1104: 123–134.

Laibson, D. 2005. Intertemporal Decision Making. In: L. Nadel (Ed.): *Encyclopedia of Cognitive Science*. New York: Wiley. Accessed in preprint version: http://www.economics.harvard.edu/faculty/laibson/files/ecsmar2.pdf; accessed 16 July 2009.

Levav, J & Argo, J.J. 2010. Physical contact and financial risk taking. *Psychological science* **21**: 6, 804-810.

Mazur, J. E. 1987. An adjusting procedure for studying delayed reinforcement. In: *Quantitative analyses of behavior V: The effect of delay and of intervening events on reinforcement value*, ed. M. L. Commons, J. E. Mazur, J. A. Nevin & H. Rachlin. Hillsdale, NJ: Erlbaum, 55-73.

Mazur, J. E. 1997. Choice, delay, probability, and conditioned reinforcement. *Animal Learning & Behavior* **25**:131–47.

Muraven, M. Tice, D.M. & Baumeister, R.F. 1998. Self-control as a limited resource: regulatory depletion patterns. *J. Personality & Social Psychology* **74**: 774-789.

Rachlin, H. 2005. Problems with internalisation. *Behavioral and Brain Sciences* **28**, 658-659.

Rachlin, H., Brown, J., & Cross, D. 2000. Discounting in judgments of delay and probability. *Journal of Behavioral Decision Making* **13**: 145-159.

Ross, Don. 2005. *Economic theory and cognitive science: Microexplanation*. Cambridge, Mass.: MIT Press.

Ross, Don. 2010. "Game Theory", *The Stanford Encyclopedia of Philosophy (Spring 2006 Edition)*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2010/entries/game-theory/>. Accessed 1 November 2010.

Ryle, G. 2002. *The concept of mind.* Chicago: Univ. of Chicago Pr.

Sanabria, F. & Killeen, P.R. Freud meets Skinner: Hyperbolic curves, elliptical theories and Ainslie Interests. *Behavioral and Brain Sciences* **28**, 660-661.

Stanovich, K.E. 2005. On the coexistence of cognitivism and intertemporal bargaining. *Behavioral and Brain Sciences* **28**, 660-661.

Sterling, G. 2011. Ockham's razor. In: Michael Bruce & Steven Barbone (Eds.): *Just the arguments: 100 of the most important arguments in Western philosophy.* Oxford: Blackwell, pp. 57-58.

Whitehead, A.N. 1926. *The concept of nature.* Cambridge: Cambridge Univ. Pr.

Wilson, M. & Daly, M. 2004. Do pretty women inspire men to discount the future? *Proc. R. Soc. Lond. B* **271**: 177-179.

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

Zauberman G. & Lynch, J.G. 2005. Resource slack and propensity to discount delayed investments of time versus money. *J. Experimental Psychol.* 134 (1): 23–37.