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

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

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

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

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

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

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

Problems with internalization

Howard Rachlin

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

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

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

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

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

https://doi.org/10.1017/S0140525X05310119 Published online by Cambridge University Press

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

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

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

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

ACKNOWLEDGMENT

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

Behavioral (pico)economics and the brain sciences

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

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

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

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

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

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

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

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

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

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

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