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-

https://doi.org/10.1017/S0140525X05320115 Published online by Cambridge University Press

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

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

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

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

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

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

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

Federico Sanabria and Peter R. Killeen

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

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

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

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

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

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

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

https://doi.org/10.1017/S0140525X05320115 Published online by Cambridge University Press