Hyperbolic discounting has confronted conventional utility theory with the likelihood that it doesn't describe elementary principles of choice, but represents a higher-order cultural invention that doesn't necessarily operate in all people or in all situations. Preferences that are temporary aren't aberrations any more, but the starting place for a strategic understanding of functions that used to be thought of as organs: the ego, the will, even the self.

Processes that pay off quickly tend to be temporarily preferred to richer but slower-paying processes, a phenomenon that can't be changed by insight per se. However, where people come to look at their current choices as predictors of what they will choose in the future, a logic much like that in the familiar bargaining game, repeated prisoner's dilemma, should recruit additional incentive to choose the richer processes. This mechanism predicts all the major properties that have been ascribed to both the power and freedom of the will. Further examination of this mechanism reveals how the will is apt to create its own distortion of objective valuation. Four predictions fit commonly observed motivational patterns: A choice may become more valuable as a precedent than as an event in itself, making people legalistic; signs that predict lapses tend to become self-confirming, leading to failures of will so intractable that they seem like symptoms of disease; there will be motivation not to recognize lapses, which creates an underworld much like the Freudian unconscious; and distinct boundaries will marshal motivation better than subtle boundaries, which impairs the ability of will-based strategies to exploit emotional rewards.

Furthermore, hyperbolic discounting suggests a distinction between short-lived reward and the more durable pleasure that allows us to account for the often-observed seductiveness of pain and "negative" emotions. Conversely, the likelihood that this discounting pattern hastens our consumption of a reward where slower consumption would be richer explains why we seek external occasions for rewards that are otherwise at our disposal. The existence of both strong lures to entertain aversive mental processes and intrinsic constraints on freely available, pleasurable processes makes it possible to do without the hoary theory of classical conditioning. Instead, emotions and hungers (together, "appetites") recur to the extent that they are rewarded. This means that the "conditioned stimuli" for appetites are not automatic triggers, but signs that emitting these appetites will be more rewarding, at least in the very short run, than not emitting them. These cues don't *release* appetites, they *occasion* them.

The urge to prematurely satisfy appetite teaches efficiency of reward-getting, but brings about the decline of pleasures once they have become familiar. This problem provides a primary motive for the separation of belief from fantasy. Instrumental needs aside, beliefs determined by relatively rare events that are outside of your control are better occasions for feeling than your own arbitrary constructions, and hence come to be experienced as more meaningful. However, uniquely well-established social constructions may function about as well as objective facts in this regard. Similar logic explains the value of empathic interaction with other people, apart from any motives for practical cooperation. To gamble, in effect, on the experiences of others keeps your occasions for emotion surprising, and thus counteracts learned habituation.

Finally, there is an inevitable clash between two kinds of reward-getting strategies. Belief in the importance of appetite-satisfying tasks – amassing wealth, controlling people, discovering knowledge itself – leads to behaviors that rush to completion; but a tacit realization of the vulnerability of appetite motivates a search for obstacles to solutions, or for gambles that will intermittently undo them. Consciousness of the second task spoils the very belief in the first task that makes the first task strict enough to be an optimal pacer of reward. Thus, the task of restoring appetite tends to be learned indirectly, and to be culturally transmitted via beliefs that seem superstitious or otherwise irrational to conventional utility analysis.

ACKNOWLEDGMENTS

Thanks to Lynne Debiak for corrected artwork, and to John Monterosso, David Spurrett, and Andries Gouws for comments on this précis.

NOTES

The author of this Précis is employed by a government agency and, as such, this Précis is considered a work of the U.S. government and not subject to copyright within the United States.

1. Even if these elements are governed by different brain centers, neurophysiologists Shizgal and Conover have pointed out that there has to be an "evaluative circuitry" that reduces them to a common currency: "For orderly choice to be possible, the utility of all competing resources must be represented on a single, common dimension" (Shizgal & Conover 1996).

2. For an analogous problem in social organization, see Sunstein (1995).

3. Figure 6B is Figure 10B in the target book. In Figure 10B the slope of increasing appetite is steeper than it is in Figure 10A, whereas to illustrate my point it has to be the same as in Figure 10A. Figure 6B has thus been corrected here.

Open Peer Commentary

Models of preference reversals and personal rules: Do they require maximizing a utility function with a specific structure?

Horacio Arló-Costa

Philosophy Department, Carnegie Mellon University, Pittsburgh, PA 15213.
hcosta@andrew.cmu.edu
http://www.phil.cmu.edu/faculty/arlocosta/

Abstract: One of the reasons for adopting hyperbolic discounting is to explain preference reversals. Another is that this value structure suggests an elegant theory of the will. I examine the capacity of the theory to solve Newcomb's problem. In addition, I compare Ainslie's account with other procedural theories of choice that seem at least equally capable of accommodating reversals of preference.

One of the ambitions of Ainslie's book is to develop a *descriptive* theory of choice and intentionality (Ainslie 2001). The theory abandons some of the main tenets of Rational Choice Theory (RCT) although it is still compatible with explanations in terms of maximization of time-indexed utility.

The main departure from RCT is based on Ainslie's conviction that discount curves are not only nonexponential, but also "specifically hyperbolic" (p. 31). Although this hypothesis has some cur-

https://doi.org/10.1017/S0140525X05220113 Published online by Cambridge University Press

rency among economists, the idea is not completely consensual. The phenomenon of preference reversals is certainly an empirically robust fact. But as Rubinstein (2003) and other economists have recently remarked, the adoption of hyperbolic discounting does far more than just modeling the psychological phenomenon that the present has a special status. "It assumes the maximization of a utility function with a specific structure and as such it misses the core of the psychological decision-making process. Thus, I find it to be no more that a minor modification to the standard discounting approach" (Rubinstein 2003, p. 1215).

The objection raised by Rubinstein is that the same sort of evidence provided by preference reversals can also reject hyperbolic discounting as well; although most of these phenomena can be explained in terms of a decision procedure based on similarity relations, Rubinstein's own proposal constitutes an even more severe departure from RCT (e.g., transitivity is not satisfied). It has nevertheless the merit of delivering an account of other important choice anomalies that inspired the main theories of nonexpected utility.

Of course, a sophisticated theory of rationality cannot be based on an unmodified version of RCT. This is so in virtue of *normative* reasons alone. For example, some phenomena, like Ellsberg's, seem to require a normative relaxation of RCT of the sort advocated by Amartya Sen (2002) or Isaac Levi (1986). An adequate descriptive theory of choice should also be able to explain problematic patterns of choice of this kind. But it seems doubtful that the adoption of a particular functional form for utility can accomplish that. The so-called theories of nonexpected utility, which also relied on the use of special utility curves, have been relatively successful in explaining only a limited range of recalcitrant phenomena (e.g., Ellsberg's puzzle cannot be explained by appealing to the special utility and probability curves of prospect theory; especial and questionable ad hoc hypotheses have been used instead by Fox and Tversky [1991] – see Arló-Costa and Helzner [2005] for a critical appraisal of this account).

In Chapter 4 of Ainslie's book, hyperbolic discounting is applied via ingenious arguments to develop accounts of emotions, pain, and other aversive behavior. The extension requires the neat separation of the notions of reward and pleasure. Therefore, the theory is quite different from the standard hedonic articulations of rational choice. In addition, the model presents an account of emotions as "pulled" by reward, which seems, prima facie, controversial.

But the bulk of the book is devoted to showing that one can develop a model of the will as emerging from a process of intertemporal bargaining among units called *interests*. This is one of the most imaginative and interesting parts of the book. Perhaps the most charitable account of interests is as time slices of the self.

Stroz concluded in his seminal paper (Stroz 1955) that "the individual always decides what to do on the assumption that he has no authority over his future self" (p. 180). Since then, many philosophers have questioned this assumption by postulating that agents can bind themselves through the operation of their wills. This strategy *postulates* rather than explains intentionality. Ainslie suggests instead that when someone seems to be choosing according to principle (or by using a personal rule), what literally happens is that his successive selves form a repeated prisoner's dilemma relationship, which is solved in the manner of interpersonal bargainers. So the idea here is to explain the will away, rather than invoke it to explain commitment.

Ainslie concedes that the existence of the resulting internal feedback process is probably impossible to study via controlled experiments. Nevertheless, he claims that postulating hyperbolic discounting (and therefore recursive decision-making) solves some well-known philosophical conundra on intentionality and choice. I only have space here to focus on one of them: the so-called Newcomb's problem.

As originally formulated, Newcomb is a single-play situation and the puzzle is to determine what is *normatively* required in this case. Ainslie seems to think (citing Nozick 1969) that in this case RCT requires choosing both boxes (p. 134). But Nozick's argument insists that if one maximizes "evidential" expected utility (EU), one should choose one box. Dominance is invoked to rationalize the two-box solution. Later on, deviant versions of RCT (causal versions) have been developed in order to articulate the latter solution (Joyce 1999). Presystematically, Ainslie seems to be a two-boxer (p. 137). But this conclusion is deeply controversial (see, e.g., Levi 1975; Meek & Glymour 1994), and it does not seem to be based on using hyperbolic discounting or recursive decision-making.

In addition, the author argues that the temptation to hedge on a personal rule could be modeled as a problem with the same arithmetic structure as a repeated version of Newcomb. In his model, he argues, diagnostic acts are also causal acts and one can then explain the sequential analogue of the "evidential" solution in the single-play case (cooperating). But this does not seem to explain why Ainslie defends the solution analogous to mutual defect in the one-shot case of Newcomb. And this is the gist of Newcomb's problem.

Ainslie's very ambitious idea is to open the impenetrable black box of intentionality, which is then modeled as a sort of brokerage process. But this process is hard to dissect on account of its recursive nature. I do not think that the appeal to philosophical puzzles like Newcomb's adds credibility to it. Many of these puzzles share with Ellsberg's example the feature of being thought experiments designed to uncover *normative* inadequacies of RCT. And the theory presented in the book, like many other descriptive theories of choice, seems unable to deal with this type of problems. Ultimately, the plausibility of Ainslie's theory seems to rest on how well it fares in comparison to other attempts to "open the black box of decision."

The final chapters of the book (on the "dyscontrol" symptoms that can be induced by bargaining strategies) are full of fascinating insights. Probably, Ainslie offers in this book one of the most complete and theoretically unified theories of the will available today. The resulting overall picture is certainly quite impressive.

Three other motivational factors

Kent Bach

Department of Philosophy, San Francisco State University, San Francisco, CA 94132. **kbach@sfsu.edu** **http://online.sfsu.edu/~kbach**

Abstract: Ainslie uses his hyperbolic discount model to explain a dazzling array of puzzling motivational phenomena. In so doing, he assumes that the motivational force of a given option at a given time is directly proportional to its discount-adjusted reward as assessed at that time. He overlooks three other factors which, independently of the perceived reward, can affect motivational force.

Ainslie (2001) assumes that the motivational force of a given option at a given time is directly proportional to its discount-adjusted reward as assessed at that time. Evidently, he rejects any independent role for cognition in mediating or arbitrating competing rewards or in deliberating and deciding what to do when. Rather, he conceives of these interests as a population of quasi-independent agents engaged in tacit bargaining, each aimed at its own temporally discounted reward. He argues that the curves representing this discounting are "highly bowed," hyperbolic rather than exponential in form, thereby allowing for temporary reversals of preferences. He briefly mentions four other discounting patterns (p. 208), but, mathematically speaking, there are countless others consistent with temporal reversal. Indeed, perhaps the discounting takes different shapes for different rewards, and maybe the discounting is sometimes non-monotonic, as with highly unstable desires. But let's assume that Ainslie's contention that they are hyperbolic is not hype and that he is not taking us for pigeons.

https://doi.org/10.1017/S0140525X05220113 Published online by Cambridge University Press