Behavioral and Brain Sciences 30, 16-18, 2007

Game theory can build higher mental processes from lower ones[1]

Commentary on Gintis, H. (2007). A framework for the unification of the behavioral sciences. *Behavioral and Brain Sciences*, *30*(1), 1-15.

by George Ainslie

151 Veterans Affairs Medical Center, Coatesville, PA 19320
University of Cape Town, Rondebosch 7701, South Africa

george.ainslie@va.gov
www.picoeconomics.org

This material is the result of work supported with resources and the use of facilities at the Department of Veterans Affairs Medical Center, Coatesville, PA, USA. The opinions expressed are not those of the Department of Veterans Affairs of the U.S. Government. It is considered a work of the U.S. government and as such is not subject to copyright within the United States.

Abstract:
The question of reductionism is an obstacle to unification. Many behavioral scientists who study the more complex or higher mental functions avoid regarding them as selected by motivation. Game-theoretic models in which complex processes grow from the strategic interaction of elementary reward-seeking processes can overcome the mechanical feel of earlier reward-based models. Three examples are briefly described.

Text

Gintis's call for unification is well reasoned, but some behavioral scientists may resist it because of a largely unspoken rift that divides us into reductionist and anti-reductionist camps. The reductionists claim that people's various stated reasons for making choices – desire, duty, sympathy, ethics, and so on – ultimately depend on a unitary selective factor that operates in a single internal marketplace. The anti-reductionists do not have an alternative theory – pointedly – but shrink from the potential hubris of reductionist theories.

Reductionists infer the selective factor from the fact of choice itself (Premack 1959) and call it utility, satisfaction, reinforcement, reward, even "microwatts of inner glow" (Hollis 1983). Gintis follows the biologists in calling it fitness, or the expectation of fitness, but this usage confounds the selection of organisms – from which fitness is inferred – with the selection of behaviors within individuals (*proximate* as opposed to *ultimate* causality in his terms; see target article, sect. 1).[2] He is certainly a reductionist, but he does not say how the higher mental processes might be selected within individuals. For instance, his statements about internalized values being "constitutive," prevailing because of their moral

value, and depending "in part on the contribution of values to fitness and well-being" (sect. 7) leave the role of the internal marketplace in their selection unclear.

Anti-reductionists have the same concern that may move the proponents of free will in philosophy, the fear that

> reductionism is a plague that grows proportionally as our society gets more sophisticated at controlling human behavior. We come to experience and conceptualize ourselves as powerless victims of mechanism, and thereby enter into a self-fulfilling prophecy. (Miller 2003)

This fear is not entirely unfounded. For example, there is a lively debate about whether education in rational choice theory makes people less cooperative (Frank et al. 1996; Haucap & Just 2003). However, as Gintis points out, this education itself is probably erroneous. Likewise, the mechanical feel of reductionism may have come from some authors' procrustean application of simple experimental paradigms to complex human situations (e.g., Skinner 1948). Explicit hypotheses about how higher mental functions arise from lower ones might dispel robotic fantasies and clear the way for the unification Gintis envisions.

Elsewhere I have argued that rich human experience can be understood to arise from the interaction of simpler processes, without violence to its subtleties (Ainslie 2001; 2005). Hyperbolic discounting has the potential to motivate conflicting reward-based processes that can endure for long periods in a limited warfare relationship, giving an individual choice-maker many of the properties of a population of choice-makers. Just as "decision-making must be the central organizing principle of psychology" (target article, sect. 1.2.1), I submit that this limited warfare relationship among successively dominant interests in individuals must determine the basic nature of decision-making. The higher mental processes that are the starting point of cognitive psychology, sociology, and anthropology not only interact in ways that are clarified by game theory, as Gintis describes, but they also arise through game-theoretic mechanisms from simpler reward-seeking skills.

Three examples show the potential of this approach to go beyond the Skinner-box-writ-large: will, in the aspects of both strength (necessary for BPC's consistency; sect. 9.2) and freedom (necessary to meet antireductionist objections); vicarious reward, which interacts with will to motivate other-regarding preferences (sect. 10); and the construction of belief, for which Gintis seeks a mechanism in section 11 (see also sects. 6 and 7).

Will. Willpower can be understood as a person's[3] interpretation of her own choices in successive temptations as cooperations or defections in an intertemporal variant of repeated prisoner's dilemma (Ainslie 2001, pp. 78–104; 2005). Insofar as a person sees her choice about a current temptation as predicting how she will choose about similar future temptations, she adds the rewards for those choices to the rewards she can expect in the current choice – a perception that under hyperbolic but not exponential discounting gives her additional incentive to resist temptation. Given hyperbolic discounting, it is only by learning such perceptions that "the observed behavior of individuals with discount rates

that decline with the delay" can "[become] choice consistent" (sect. 9.2). Thus, the will can be interpreted as the perception of a bargaining situation among a person's successive selves rather than as a faculty with inborn complexities. Furthermore, the sensitive dependence of repeated prisoner's dilemmas on individual choices makes their outcomes unpredictable from mere knowledge of their contingencies – even by the person herself – thereby arguably reconciling the experience of free will with determinism. This kind of bridge from the bottom upward in the hierarchy of complexity will not reduce the study of higher mental functions to something more molecular, but it can supply a context that connects them to basic motivational science.

Vicarious reward. Whatever way altruism and social virtues are selected by fitness, putatively their ultimate cause (sect. 10), Gintis and his cited authors address their proximate causes (rewards) only in terms of reciprocity. Hyperbolic discounting suggests how vicarious experience can be rewarding in its own right. The piece that has been missing in utility-maximizing theories of social utility is emotion. In contrast to conventional, conditioned reflex models of emotion, hyperbolic discounting permits emotion to be seen as a motivated process that taps endogenous sources of reward – transient reward alternating with inhibition of reward in the case of negative emotions, reward attenuated by anticipation and habituation in the case of positive emotions (Ainslie 2001, pp. 164–174, 179–186; 2005). Emotional reward does not physically require stimuli from the environment, but it still needs them in practice because it will habituate to the level of a daydream unless occasioned by environmental events that are both of limited frequency and partially unpredictable.

Various kinds of gambles, challenging tasks, and fictional stories are among the patterns that can meet these criteria, but the most apt should be the actual experience of other people. My hypothesis is that the experiences of other people acquire value in the internal marketplace of reward insofar as they are good occasions for emotion, and that both social virtues and social vices acquire value insofar as they support strategies of occasioning emotion, respectively in the long run and short run. The rewarding properties of the various emotions are undoubtedly shaped in evolution by their contribution to fitness. In the individual, however, emotion is a reward-producing behavior that produces more or less depending on how occasions pace its occurrence over time. Thus, in addition to self-regarding reciprocity, the stuff of sociology and anthropology is woven by emotion-cultivating processes that develop complex social skills to avoid habituation.

Construction of belief. Finally, Gintis says that "beliefs directly affect well-being" (sect. 11), by which he means that, apart from their instrumental value in getting other rewards, beliefs are rewarding in their own right. Social constructionists have long made this point, but have not said what constrains motivated belief; that is, what makes belief different from make-believe. Elsewhere I have argued that the noninstrumental value of beliefs is to occasion emotion (Ainslie 2001, pp. 175–179; 2005) and that the two kinds of value are often confounded because the limited occurrence of instrumental success also qualifies information predicting it as a good occasion for emotion (Lea & Webley 2006; Ainslie 2006). "Transcendental beliefs" (sect. 6) are a large category of emotionally useful belief that is made unique for the individual not by instrumental accuracy but by cultural

consensus. Such beliefs have to be transmitted in "conformist" fashion lest they lose their uniqueness and thereby weaken their value as occasions for emotion – but they still survive only insofar as they produce individual reward. Likewise, although a person is apt to shed suggested norms that are not useful to her as boundaries against temptation (criteria for cooperation in her intertemporal prisoner's dilemmas; see my subsection *Will* above), she will find that the ones she believes to be uniquely dictated by fact ("internalizes" – sect. 7) are the most effective, as Gintis observes.

Just as societies are constructed by individuals interacting strategically, so too these individuals are constructed by basic reward-seeking processes that also interact strategically. However, maximizing reward implies neither selfishness nor determination by external contingencies.

Notes

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

2. Of course, the factor that selects behaviors within an individual must in turn have been selected in the species by its effect on fitness; but it may still lead her well astray from fitness, as witness cocaine and birth control.

3. It is possible, but doubtful, that some nonhuman animals have sufficient theory of mind to use their own current choices as predictive cues.

References

Ainslie, G. (2001) *Breakdown of will.* Cambridge University Press.

Ainslie, G. (2005) Précis of *Breakdown of will. Behavioral and Brain Sciences* 28(5), 635–673.

Ainslie, G. (2006) What good are facts? The "drug" value of money as an examplar of all non-instrumental value. *Behavioral and Brain Sciences* 29(2), 176–77.

Frank, R. H., Gilovich, T. & Dennis, R. (1996) Do economists make bad citizens? *Journal of Economic Perspectives* 10, 187–92.

Haucap, J. & Just, T. (2003) Not guilty? Another look at the nature and nurture of economics students. Discussion Paper 8, University of the Federal Armed Forces, Hamburg, Germany. Available at: www.ruhr-uni-bochum.de/wettbewerb/dlls/forschung/paper8.pdf

Hollis, M. (1983) Rational preferences. *The Philosophical Forum* 14:246–62.

Lea, S. E. G. & Webley, P. (2006) Money as tool, money as drug: The biological psychology of a strong incentive. *Behavioral and Brain Sciences* 29(2):161–209.

Miller, W. R. (2003) Comments on Ainslie and Monterosso. In: *Choice, behavioural economics, and addiction,* eds. R. Vuchinich & N. Heather, pp. 62-66. Pergamon Press.

Premack, D. (1959) Toward empirical behavior laws, I. Positive reinforcement. *Psychological Review* 66:219–34.

Skinner, B. F. (1948) *Walden two.* Macmillan.