Towards a Unity of the Human Behavioral Sciences*

Herbert Gintis
Central European University. Budapest
External Faculty, Santa Fe Institute. Santa Fe, New Mexico
hgintis@comcast.net

Abstract

Despite their distinct objects of study, the human behavioral sciences all include models of individual human behavior. Unity in the behavioral sciences requires that there be a common underlying model of individual human behavior, specialized and enriched to meet the particular needs of each discipline. Such unity does not exist, and cannot be easily attained, since the various disciplines have *incompatible* models and *disparate* research methodologies. Yet recent theoretical and empirical developments have created the conditions for unity in the behavioral sciences, incorporating core principles from all fields, and based upon theoretical tools (game theory and the rational actor model) and data gathering techniques (experimental games in laboratory and field) that transcend disciplinary boundaries. This paper sketches a set of principles aimed at fostering such a unity. They include: (a) evolutionary and behavioral game theory provides a transdisciplinary lexicon for communication and model-building; (b) the rational actor model, rooted in biology but developed in economic theory, applies to all the human behavioral disciplines. This model treats actions as instrumental towards satisfying preferences. However, the content of preferences must be empirically determined. Moreover, the rational actor model is based on a notion of preference consistency that is not universally satisfied, so its range of applicability must also be empirically determined; (c) controlled experiments have been underutilized in most behavioral disciplines. Game theory and the rational actor model can be used as the basis for formulating, deploying, and analyzing data generated from controlled experiments with human subjects.

Key words: behavioral sciences, game theory, experimental economics, rational actor model.

Resumen. *Hacia la unidad de las ciencias del comportamiento humano*

A pesar de que tienen objetos de estudio distintos, todas las ciencias del comportamiento humano cuentan con modelos de la conducta humana individual. La unidad de tales ciencias requiere un modelo común subyacente de comportamiento humano individual, especificado y enriquecido para satisfacer las necesidades particulares de cada disciplina. No existe tal unidad, y no puede ser fácilmente alcanzada, dado que que las diversas disciplinas tienen modelos *incompatibles* y metodologías de investigación *dispares*. Con todo, recientes desarrollos teóricos y empíricos han creado las condiciones para la unidad de las ciencias del comportamiento, incorporando principios centrales en todos los campos, y basándose en herramientas teóricas (como la teoría de juegos y el modelo de actor racional) y técnicas de recogida de datos (como los juegos experimentales de laboratorio y sobre el terreno) que transcienden las fronteras disciplinarias. Tales desarrollos incluyen: (a) la teoría de juegos

* I would like to thank the John D. and Catherine T. MacArthur Foundation for financial support.

evolucionaria y conductual, que proporciona un léxico transdisciplinar para la comunicación y la construcción de modelos; (b) el modelo de actor racional, anclado en la biología pero desarrollado por la teoría económica, que se aplica en todas las disciplinas del comportamiento humano. Este modelo trata las acciones como instrumentales, dirigidas a la satisfacción de las preferencias. Sin embargo, el contenido de las preferencias debe ser empíricamente determinado. Además, el modelo de actor racional está basado en una noción de la consistencia de las preferencias que no se satisface universalmente, de modo que su rango de aplicabillidad debe determinarse también empíricamente; (c) los experimentos controlados, que han sido infrautilizados en la mayoría de las ciencias del comportamiento. La teoría de juegos y el modelo de actor racional pueden ser usados como base para formular, desplegar y analizar datos generados a partir de experimentos controlados con sujetos humanos.

Palabras clave: ciencias del comportamiento, teoría de juegos, economía experimental, modelo de actor racional.

Contents

1. Introduction

The human behavioral sciences include economics, human biology, anthropology, sociology, behavioral psychology, and political science[1]. We may consider a set of disciplines as *unified* if they are (a) *consistent*, so that in cases where two disciplines deal with the same social phenomena, their models are equivalent, and *synergic*, each discipline being substantively enriched by the scientific content of the others. The natural sciences achieved unity with the development of quantum mechanics, elementary particle and solid state physics, and the big bang model of the universe. Such unity is lacking in the human behavioral sciences. Each behavioral discipline models *individual human behavior*, and construct models of aggregate social behavior compatible with, and often derived from, a model of individual behavior. Unity in the human behavioral sciences requires a common underlying model of individual behavior, which each discipline specializes and enriches for its particular purposes. No current model enjoys such transdisciplinary status.

1. By «human biology» I mean the application of biological techniques to modeling human behavior. I use the term «behavioral psychology» to mean social psychology and psychological decision theory.

Yet, recent developments reveal links across the behavioral sciences sufficiently deep to establish the preconditions for unity. Both sociology (Hechter and Kanazawa, 1997) and political science (Monroe, 1991), following the pioneering contributions of Coleman (1990), Downs (1957), Olson (1965), Buchanan and Tollison (1984) and others, have begun to adopt the rational actor model, previously espoused virtually exclusively in economics. Game theory, a central element of economic theory, was introduced to biology by Lewontin (1961), Hamilton (1967) and Maynard Smith and Price (1973), subsequently maturing into an invaluable behavioral tool (Alcock, 1993, Dugatkin and Reeve, 1998, Gintis *et al.*, 2001, Gintis, 2003a). In anthropology, the application of experimental game theory to understanding cultural variation is rather new, but quite promising (Henrich *et al.*, 2001, Henrich *et al.*, 2004). Conversely, increasing numbers of economists develop behavioral models of social interaction, and draw upon evidence from experimental game theory in modeling behavior. This development is evidenced by the Nobel prize in economics for the year 2002, awarded jointly to two experimentalists: psychologist Daniel Kahneman and economist Vernon Smith.

In this paper I will sketch a set of principles that express my current conception of unity. I will argue the following points:

a. Game theory provides a transdisciplinary behavioral lexicon for communication and model-building. For many years it was widely thought that game theory presupposes methodological individualism and a high level of cognitive functioning on the part of subjects. Were this the case, game theory would be inapplicable to settings where emotion, traditional, and heuristic behaviors are prominent, and where group-level processes and dynamic interactions are common. Contemporary evolutionary and behavioral game theory, however, extends classical game theory to cover such settings.
b. Evolutionary biology underlies all behavioral disciplines because *Homo sapiens* is an evolved species whose major characteristics are the product of its particular evolutionary history.
c. Evolutionary and behavioral game theory provides the substantive framework for the biology of human behavior.
d. The rational actor model, developed in economic theory, is a flexible tool that applies to all the human behavioral disciplines. This model treats actions as instrumental towards satisfying preferences. However, the content of preferences must be empirically determined, and individuals may have preferences over actions as well as their outcomes. Moreover, the rational actor model is based on a notion of preference consistency that is not universally satisfied, so its range of applicability must also be empirically determined.
e. Progress in modeling human behavior has been hampered by the *underutilization of controlled experiments*, which are common only in behavioral psychology. Game theory and the rational actor model can be used as the

basis for formulating, deploying, and analyzing data generated from controlled experiments in social interaction. Such controlled experiments are replicable across laboratories and foster cumulative knowledge relevant to all behavioral disciplines.

f. Progress in modeling human behavior has been hampered by the *artificially restricted range of social situations studied by behavioral scientists*. Only anthropology has systematically studied the effects of cultural differences *across societies* on human behavior, only sociology has systematically studied the effects of cultural differences *within societies* on human behavior, and only behavioral psychology has systematically studied the effects of *personality differences* on social interaction. A unified model of human behavior is fostered by taking controlled experiments to the field, and deploying such experiments in a variety of cross-cultural settings across and within societies.

g. The demographic success of *Homo sapiens* is due to the ability of humans to sustain a high level of cooperation among non-kin. Whereas biology and economics explains this ability in terms of exchange among self-interested individuals, the facts are in line with basic sociology and behavioral psychology: humans often display altruistically prosocial behavior, especially in a form that I will call *strong reciprocity* —a predisposition to cooperate and to punish non-cooperators at personal cost (Gintis *et al.*, 2005).

h. Prosocial behavior in humans can be modeled biologically using the tools of gene-culture coevolution, but the social mechanisms involved must include using the sociological notions of *socialization* and the *internalization of norms*.

Two *caveats* are in order. First, this set of unifying principles is incomplete and highly subject to revision. In particular, I make no mention of neuroscience or behavioral genetics. This is in part for lack of space, and in part because the general interconnection between these and other parts of behavioral science are unclear and our ideas thereupon are rapidly changing. Second, I claim that each behavioral discipline has developed core principles that are largely accurate, yet overlooked or denied by other disciplines. I do not assert that the ones I discuss are the *only* core principles of the discipline, or even the most important. Rather, they are principles central to the unity of the human behavioral sciences.

2. Game Theory as Behavioral Lexicon

Communication across disciplines presupposes a common language. *Game theory* is a universal behavioral lexicon that offers such a common language. In the language of game theory, *players* (or *agents*) are endowed with a set of available *strategies*, and have a range of *information* concerning the rules of the game, the nature of the other players and their available strategies, as well as the structure of payoffs. Finally, for each combination of strategy choices by

the players, the game specifies a distribution of *individual payoffs* to the players. If the game is accurately specified, we can predict the behavior of the players by assuming they attempt to maximize some preference function involving their personal payoffs, their chosen strategies, the personal payoffs to other agents, and the actions of the other agents (Gintis, 2000). Self-regarding agents maximize their personal payoffs, while other types of agents may care about fairness, the intentions of other agents, the sum of all payoffs, their relative personal payoff, and other aspects of the array of payoffs.

Developments within game theory in recent years have considerably enhanced its value to behavioral disciplines that have traditionally found little use for this analytical tool. First, it is now widely recognized that in many social interactions, individuals are not self-regarding, but rather care about the payoffs to and intentions of other players (Rabin, 1993, Bergstrom and Stark, 1993, Andreoni and Miller, 2002, Fehr and Gächter, 2002, Wood, 2003). Second, human actors care not only about material payoffs, but power, self-esteem, and behaving morally (Gintis, 2003b, Bowles and Gintis, 2005, Wood, 2003), goals that are recognized as central to many behavioral disciplines. Third, evolutionary and behavioral game theory do not require the extensive cognitive and information processing capacities of classical game theory, so disciplines in which it is recognized that cognition is a scarce and costly good can make use of game-theoretic models (Young, 1998, Gintis, 2000, Gigerenzer and Selten, 2001). Thus, individuals may consider only a restricted subset of strategies (Winter, 1971, Simon, 1972), and they may use by rule-of-thumb heuristics rather than maximization techniques (Gigerenzer and Selten, 2001). Game theory is thus a generalized schema that permits the precise framing of meaningful empirical assertions, but imposes no particular structure on the predicted behavior.

3. The Universality of the Rational Actor Model

The rational actor model assumes that individuals have *preferences* reflecting their wants and the tradeoffs among these wants, and that individuals maximize their *utility* by choosing from an *action set* that is limited by available information, material resources and time, cognitive capacity, and the agent's physical capacities. Choice is also contingent upon *beliefs* concerning the probabilities of various states of nature, the frequency distribution of types of individuals with whom they interact, and the relative effectiveness of different actions. The rational actor model is most highly developed in economics, but it applies to all the disciplines dealing with human behavior.

The rational actor model appears *prima facie* to apply only when extremely stringent conditions are satisfied. However, the model can be shown to apply over any domain in which the agent has *transitive preferences*, in the sense that if he prefers A to B and he prefers B to C, then he prefers A to C, and the agent can *make tradeoffs among outcomes* in the sense that for any finite set of outcomes $A_1,\ldots,A_n$, if A_1 is the least preferred and A_n the most preferred out-

come, then for any A_i, ≤ 1 $i \leq n$ there is a probability p_i, $0 \leq p_i \leq 1$ such that the agent is indifferent between A_i and a lottery that pays A_1 with probability p_i and pays A_n with probability $1-p_i$ (Kreps, 1990). Clearly, these assumptions are often extremely plausible. When applicable, the rational actor model's transitivity assumption strongly enhances explanatory power, even in areas that have traditionally abjured the model (Coleman, 1990, Kollock, 1997, Hechter and Kanazawa, 1997).

The rational actor model is ubiquitous because *any evolved life form is likely to conform to its consistency conditions over some range of actions*. This is because biological agents do not directly maximize fitness, but rather possess a genetically-rooted set of routines, involving needs, drives, pleasures, and pains, that determine how to respond to internal events (e.g., hunger) and external circumstances (e.g., temperature). This is precisely the agent's preference function, which will thus be transitive so long as actual choices reflect biological fitness, which is a linear variable. Evolutionary forces ensure that, under constant environmental conditions, maximizing this preference function will in fact come close to maximizing the agents' fitness. Since environmental conditions are not constant, however, preference consistency must always be empirical demonstrated rather than assumed *a priori*.

The rational actor model has been underutilized in some behavioral disciplines through several prominent misunderstandings. First, the rational actor model does not require that individuals be self-interested. There is no connection between the notion of the transitivity of preferences and the notion that preferences are purely self-regarding. Indeed, one can apply standard choice theory, including the derivation of demand curves, plotting concave indifference curves, and finding price elasticities, for such preferences as charitable giving and punitive retribution (Andreoni and Miller, 2002). Second, because the rational actor model treats action as instrumental towards achieving rewards, it is often inferred that action itself cannot have reward value. This is an unwarranted inference. For instance, the rational actor model can be used to explain the expressive motivation in rational action, including collective action that is precluded by the assumption that individuals act instrumentally towards satisfying their material needs (Olson, 1965), since individuals may place positive value on the process of acquisition (for instance, «fighting for one's rights»), and can value punishing those who refuse to join in the collective action (Moore, Jr., 1978, Wood, 2003). Third, the areas over which the transitivity postulate holds must be *empirically determined*.

For example, consider the *discount rate* —the rate at which individuals are willing to sacrifice present for future gains. In economics, «rationality» in the form of consistency of preferences across time implies that individuals use *exponential discounting*, in which the discount rate is constant across all periods. Assuming this consistency, the discount rate can be estimated empirically at about 3% per year (Huang and Litzenberger, 1988, Rogers, 1994). Animal studies find that non-human species have discount rates that are several orders of magnitude higher than this (Stephens *et al.*, 2002). Humans and other ani-

mals exhibit *hyperbolic discounting*, according to which discount rates for present versus near-future are much higher than discount rates for similar time periods starting in the more distant future (Herrnstein, 1961, Ainslie, 1975, Ainslie and Haslam, 1992, Laibson, 1997). This finding corresponds to the everyday notion that we are subject to «temptation» and «failure of will,» leading us to accept high long term penalties for small short-term pleasures (smoking, overeating, procrastinating). Formally, this means that preferences are not transitive across time, and most observers of this phenomenon (including the individuals who are subject to this inconsistency) agree that the 'distant future' discount rates more closely conform to the agent's welfare. Consonant with these findings, sociological theory stresses that *impulse control* —learning to favor long-term over short-term gains— is a major component in the socialization of youth (Strotz, 1955, Ainslie, 1975, Power and Chapieski, 1986, Grusec and Kuczynski, 1997). Time inconsistency does not imply that the rational actor model be rejected, but we must include parameters to deal analytically with preferences across time (Laibson, 1997).

Another misunderstanding is that embracing the rational actor model entails equating well-being with preference satisfaction. If rational agents have consistent preferences and are perfectly informed, such an inference may be warranted, but these conditions do not always hold. In particular, if individuals are excessively present-oriented (a major form of time inconsistency) then their choices will not reflect their long-term well-being. The time-inconsistent rational actor model may thus shed considerable light on such perverse phenomena as procrastination, substance abuse, undersaving for old age, and obesity (Elster, 1979, Akerlof, 1991, O'Donoghue and Rabin, 2001).

Broadening the rational actor model beyond its traditional form in neoclassical economics run the risk of developing unverifiable and *post hoc* theories, as our ability to theorize outpaces our ability to test theories. To avoid this, and following the lead of behavioral psychology, we must expand the use of controlled experiments, as suggested above. Often we find that the appropriate experimental design can generate new data to distinguish among models that are equally powerful in explaining the existing data (Tversky and Kahneman, 1981, Kiyonari *et al.*, 2000).

4. Biological Replicators

The analysis of living systems includes one analytical element that does not occur in the non-living world, and is not analytically represented in the natural sciences. This is the notion of a *replicator* (Schrödinger (1944) called this an «aperiodic crystal»), which is a physical system capable of drawing energy from its environment to make relatively accurate copies of itself. The dynamics of replicators are described by the evolutionary notions of replication, mutation, selection, and adaptation (Lewontin, 1974).

Biology plays a role in the behavioral sciences much like that of physics in the natural sciences. Just as physics studies the elementary processes that

underlie all natural systems, biology studies the general characteristics of replicators. In particular, genetic replicators account for the characteristics of species and their interactions, as well as their similarities and differences in morphology, physiology, and behavior. Just as one cannot deduce the character of natural systems (e.g., the principles of inorganic and organic chemistry, the structure and history of the universe, robotics, plate tectonics) from the basic laws of physics (e.g., quantum and statistical mechanics), similarly one cannot deduce the structure and dynamics of social life from basic biological principles.

The most natural setting for replicator dynamics is game theoretic. Replicators endow copies of themselves with a repertoire of strategic responses to environmental conditions, including information concerning the conditions under which each is to be deployed in response to character and density of competing replicators. Mutations included replacement of strategies by modified strategies, and the «survival of the fittest» dynamic (formally called a *replicator dynamic*) ensures that replicators with more successful strategies replace those with less successful ones (Taylor and Jonker, 1978).

Classical population biology, throughout much of the Twentieth century, did not employ a game-theoretic framework (Fisher, 1930, Haldane, 1932, Wright, 1931). However, Moran (1964) showed that Fisher's Fundamental Theorem, which states that as long as there is positive genetic variance in a population, fitness increases over time, is false when more than one genetic locus is involved. Eshel and Feldman (1984) identified the problem with the population genetic model in its abstraction from mutation. But how do we attach a fitness value to a mutant? Eshel and Feldman (1984) suggested that payoffs be modeled game-theoretically on the phenotypic level, and a mutant gene be associated with a strategy in the resulting game. With this assumption, they showed that under some restrictive conditions, Fisher's Fundamental Theorem could be restored. Their results were generalized by Liberman (1988), Hammerstein and Selten (1994), Hammerstein (1996), Eshel *et al*. (1998) and others. Thus it turns out that *at the most fundamental level of population biology* game theory is key to understanding evolutionary biology. Of course, game theory has also become the basic framework for modeling animal behavior (Maynard Smith, 1982, Alcock, 1993, Krebs and Davies, 1997).

5. Gene-Culture Coevolution

Genetic replicators transmit information encoded in DNA sequences, through a germ line that is unaffected by environmental conditions. Genetic adaptation to new environments then takes the form of shifts in allele frequencies, and promotion of mutations that better exploit the new environment. In the context of rapidly changing environments, there is a fitness benefit to the transmission of *epigenetic* information concerning the current state of the environment. Such epigenetic information is quite common (Jablonka and Lamb, 1995), but achieves its highest and most flexible form in *cultural transmission*

in humans and to a considerably lesser extent in other primates (Bonner, 1984, Richerson and Boyd, 1998).

There are several basic categories of culture: *conventions* (e.g., language use), *techniques and practices* (e.g., how to prepare food, how to make and use tools, how to treat illnesses), *ethical values* (e.g., norms of fairness, reciprocity, justice) and *transcendental beliefs* (e.g., sickness is caused by angering the gods, good deeds are rewarded in the afterlife). A transcendental belief is the assertion of a causal relationship or a state of affairs that has a truth value, but whose truth holders either cannot or choose not to test. There are of course other types of beliefs, but these appear to be subsumable under other cultural categories. For instance, one may believe that a certain convention exists, a certain technique is effective, or a certain ethical value is justifiable. To avoid confusion, we treat such beliefs as part of the conventions, techniques and practices, and values that they affirm.

Conforming to *conventions* is adaptive because it is payoff-maximizing to conform when all others are doing so. When an agent must determine the most effective of several alternative *techniques* or *practices*, and if experimentation is costly, it may be payoff-maximizing to copy others rather than incur the costs of experimenting (Boyd and Richerson, 1985, Conlisk, 1988). If everyone else experiments to find the superior technique, it will generally pay simply to follow the majority. By contrast, if everyone else conforms to a single technique in a situation where different techniques are best suited to different environments, then when the environment changes an individual who experiments may do better than the conformists. Thus, in general there will be a cultural equilibrium with a positive fraction of both conformists and experimenters. In this sense, the genetic machinery for a predisposition to conform to conventions and to imitate techniques is biologically adaptive.

It is plausible to extend this explanation to transcendental beliefs as well. Such beliefs affirm techniques where the cost of experimentation is extremely high or infinite, and the cost of making errors is high as well. This is, in effect, Blaise Pascal's argument for the belief in God and the resolve to follow His precepts. It is supported by Boyer (2001), who models religion as a set of cognitive beliefs that coexist and interact with our other more mundane and testable beliefs. In this view, one conforms to transcendental beliefs because their truth value has been ascertained by others (relatives, ancestors, prophets), and are as worthy of affirmation as the techniques and practices (such as norms of personal hygiene, that one accepts on faith, without personal verification).

Conventions, techniques, and beliefs are *instrumental* in the sense that they specify how best to achieve certain ends or goals. The remaining cultural category, *ethical norms and values*, is *final* in the sense of specifying what ends or goals to embrace. I discuss the place of cultural values in human behavioral theory below.

Whether or not cultural elements are replicators in a sense closely parallel to genetic replicators is the subject of much current controversy (Aunger, 2002)

but it is clear that cultural forms reproduce themselves from brain to brain and across time, mutate, and are subject to selection according to their effects on the fitness of their carriers (Parsons, 1964, Cavalli-Sforza and Feldman, 1982, Boyd and Richerson, 1985). Moreover, there are strong interactions between genetic and epigenetic elements in human evolution, ranging from basic physiology (e.g., the transformation of the organs of speech with the evolution of language) to sophisticated social emotions (e.g., empathy, shame, guilt, revenge-seeking). The analysis of the reciprocal action of genes and culture is known as *gene-culture coevolution* (Lumsden and Wilson, 1981, Durham, 1991, Feldman and Zhivotovsky, 1992, Bowles and Gintis, 2005).

6. Cooperation: Ethical Behavior or Self-interest?

The success of *Homo sapiens*, as measured by its broad geographical distribution and its considerable share of the Earth's biomass, is based on its unique capacity to use cultural forms to transmit technical knowledge accurately across generations, and its unique ability to sustain cooperation through space and across time among large numbers of unrelated individuals (Richerson and Boyd, 1998). How do we explain this cooperation?

Biologists maintain that cooperation can be sustained by *inclusive fitness*, or cooperation among kin (Hamilton, 1963), and by individual self-interest in the form of *reciprocal altruism* (Trivers, 1971). Reciprocal altruism occurs when an agent helps another agent, at a fitness cost to itself, with the expectation that the beneficiary will return the favor in a future period. The explanatory power of inclusive fitness theory and reciprocal altruism convinced a generation of biologists that what appears to be altruism —personal sacrifice on behalf of others— is really just long-run self-interest. Economics has developed a similar model of cooperation, based on the notion of long-term, enlightened self-interest (Arrow and Debreu, 1954, Axelrod and Hamilton, 1981, Fudenberg and Maskin, 1986), an idea that goes back to Bernard Mandeville's concept of «private vices, public virtues» (1924[1705]) and Adam Smith's notion of the «invisible hand» (2000[1759]).

Sociology, by contrast, has used *socialization* to explain cooperation among non-kin. According to Durkheim (1951), the division of labor in society involves assigning individuals to specific *roles*. Individuals are inculcated with *values* and *norms* that induce them to conform to the duties and obligations of the role-positions they occupy. This is altruism.

A key tenet of socialization theory is that a society's values are passed from generation to generation through the *internalization of norms* (Durkheim, 1951, Benedict, 1934, Mead, 1963, Parsons, 1967, Grusec and Kuczynski, 1997). In the language of optimization theory, internalized norms are accepted not as instruments towards and constraints upon achieving other ends, but rather as *arguments in the preference function that the individual maximizes*. Internalized norms are thus what we termed *ethical values* in our lexicon of cultural forms. In true gene-culture coevolutionary form, a variety of unique-

ly human prosocial emotions come into play, including prominently *shame*, *guilt*, and *empathy*, directly reinforcing internalized norms.

The programmability of the preference function appears in the form of the human *capacity to internalize norms*, which consists in an older generation instilling the values and objectives of a younger generation through an extended series of personal interactions, relying on a complex interplay of affect and authority. Individuals conform to an internalized norm because so doing is an end in itself, and not merely because of the material rewards that follow from norm compliance or punishments that follow from norm violation. For instance, an individual who has internalized the value of «speaking truthfully» will do so even in some cases where the net payoff to speaking truthfully would otherwise be negative. It follows that where people internalize a norm, the frequency of its occurrence in the population will be higher than if people follow the norm only instrumentally; i.e., when they perceive it to be in their narrow material interest to do so. The capacity to internalize is based on a distinctively human psychological predisposition, unrecognized in biology and economics.

7. The Evolutionary Basis for Norm Internalization

An «altruistic norm,» when acted upon, reduces the bearer's individual fitness or material well-being, but increases the fitness or well-being of other, unrelated, group members. The internalization of altruistic norms appears to be an evolutionary *curiosum* because individuals who internalize such norms should be at a fitness disadvantage in comparison with self-interested actors. A closer look at the cultural transmission process, however, offers a resolution to this problem (Gintis, 2003a).

Suppose there is an altruistic behavior A that imposes fitness cost s on those who embrace it. Suppose also that only a fraction of youth have the genetic capacity to accept ethical norms, and this fraction increases or decreases over time according to the biological fitness of its bearers. Suppose further that altruistic behavior A is transmitted to offspring with this genetic capacity by their parents in an unbiased manner (i.e., if both or neither parents embraces A, all of their genetically enabled offspring do the same, and if only one parent embraces A, half of such offspring embrace A). In addition, suppose there is *extraparental transmission* of A, in the form of social pressure (rumor, shunning, and ostracism), rituals (dancing, prayer, marriage, birth, and death), and in modern societies, formalized institutions (schools, churches, sacred texts). Such extraparental transmission is itself altruistic, since it will generally be individually costly while the benefits, in the form of a higher frequency of altruism in the group, accrue to unrelated others. We handle this, plausibly, we believe, by assuming that the altruistic norm is both to embrace A and to encourage others to embrace it as well, and we include the cost of extraparental transmission in s, the cost of altruism. We measure the strength of extraparental transmission by a parameter γ, such that if the fraction of altruists

in the older generation is p_A, then γp_A is the probability that a given non-altruistic child with the genetic capacity to acquire the altruistic norm, will in fact be induced to embrace the altruistic norm.

Suppose, further, that an altruist who meets a non-altruist, which we assume occurs with a probability proportional to the fraction of altruists, switches to the nonaltruist's behavior with probability α. Gintis (2003a) then shows that if α satisfies the inequality

$$\alpha < \frac{\gamma - s}{1 - \gamma} \qquad (1)$$

then the altruistic cultural equilibrium, in which all individuals have the genetic capacity to embrace ethical norms, and all actually embrace A, is evolutionarily stable. Note that (a) the larger the fitness cost s of altruism, and (b) the smaller the rate γ of oblique transmission, the lower the maximal rate of «moral defection» α to the nonaltruistic that is compatible with an altruistic cultural equilibrium. Note also that if γ is sufficiently large (specifically, if $\gamma > (1+s)/2$) then no rate of defection can undermine the altruistic equilibrium, because individuals rarely meet nonaltruists with whom they can compare their fitness.

The substantive questions, then, are (a) why γ might be positive and large, and (b) why the rate α of moral defection might be low. To address (a), note that the rate of extraparental transmission depends not only on the willingness of individuals to sacrifice on behalf of the group by engaging in extraparental socialization and by rewarding others who do the same, but also on the structure of social institutions that routinize cultural transmission. There is thus no guarantee that γ will be high, but societies that do effectively organize cultural transmission, and stress ethical norms that are heavily prosocial, will tend to grow and otherwise outcompete societies that do not (this process is referred to above as *weak* group-level selection).

To address (b), we must explain why individuals might not defect at a very high rate to fitness-maximizing behavior. The following argument suggests that the psychological constitution of *Homo sapiens* is conducive to a high rate of adherence to moral norms, and hence to the satisfaction of equation (1). While nearly everyone behaves amorally on some occasions, and some behave amorally much of the time, there is normally a sufficient reserve of moral behavior, including the motivation to punish the moral transgressions of others, to maintain a high level of conformity with group morality.

As we have noted, humans do not maximize fitness, but rather a preference function that is but a rough proxy for fitness under constant environmental conditions. The rapid pace of environmental change and cultural innovation over the past 100,000 years has produced a situation in which the set of needs, desires, drives, pleasures, and pains associated with the human preference function is out of line with the dictates of fitness maximization (Richerson and Boyd, 1998). Even a random deviation of the human preference function

from fitness maximization towards other goals, such as power, esteem, wealth, and pleasure, might be conducive to a relatively slow rate of rejection of moral norms, of which altruistic norms might figure prominently.

However, there is evidence of a more systematic force intervening between biological fitness and human preferences: as we have seen, the human preference function is, to some considerable extent *programmable*, in the sense that human goals can be altered by socialization. The notion of a programmable preference function is sufficiently unusual that such a mechanism must have arisen as an adaptation, and hence the content of socialization, the actual internalized norms themselves, must be, at least on balance, fitness enhancing. Yet standard sociological theory has not supplied an argument as to why it might be adaptive, and indeed have generally ignored evolutionary arguments altogether. We can, however, supply such an argument.

A programmable preference function is the most complex instrument facilitating epigenetic information flows, all of which represent means of transferring information across generations in a manner complementary to, and often more flexible than, genetic transmission (Bonner, 1984; Boyd and Richerson, 1985; Jablonka and Lamb, 1995, 1998). The form that this epigenetic transmission process takes in the case of the internalization of norms is a protracted series of interactions, controlled by parents and influential elders, undertaken at considerable cost, and reinforced by a complex web of informal sanctions. While cultural learning occurs in many species, programmability of goals is virtually limited to humans because the capacity to be socialized presupposes a high level of cognitive capacity (Tomasello, 1999), as well as specialized mental circuitry for valuing interpersonal relationships and making informed social judgments (Damasio, 1994), and specialized emotional capacities that enhance the individual's capacity to attain internalized goals, such as pride, shame, empathy, and remorse (Bowles and Gintis, 2005). The genetic basis for prosocial emotions is clear from the fact that the inability to experience prosocial emotions, associated with sociopathic personality types, is partially heritable (Mealey, 1995), and is deficient in individuals with damage to specific regions of the brain's frontal lobes (Damasio, 1994).

The capacity to program changes in the preference function *culturally* indeed has great adaptive value. By redirecting human goals, and thereby curbing, repressing, and channeling an agent's basic impulses, the agent will have higher fitness than another agent who lacks this capacity. Included among the norms that are commonly internalized are thus norms of personal hygiene, concern for the approval of others, control of temptation, cultivation of a work ethic, and maintaining a long time horizon in decision-making. Such norms are upheld and transmitted in virtually all societies (Brown, 1991), though a breakdown of cultural transmission in this area occurs in some poorly functioning societies (Edgerton, 1992).

For an example of the fitness-enhancing capacity of internalization, note that a sophisticated weapon, such as a sharp knife, may aid an individual in taking revenge upon a transgressor, but the spontaneous impulse to attack an

enemy may be fitness-reducing when such weapons are widely available. However, parents can instill in their offspring the norms of «love thy neighbor» and «be slow to anger.» Individuals who have acquired this genetic predisposition to internalize norms will pass both this capacity and its content —the conflict-limiting norm itself— to their offspring. For this reason, the internalization of norms may be fitness-enhancing.

For a second example, suppose someone invents an aerodynamic spear that is extremely effective in the hunt, but requires daily practice to hone the throwing skills needed to use the spear effectively (Calvin, 1983). Since individuals primordially prefer less expenditure of energy to more and have inappropriately short time horizons, they will skimp on daily practice. The hunter who internalizes the norm «good hunters like to practice» will have an adaptive advantage.

One might object that a non-internalizer could always mimic the behavior of internalizers when it suits his purposes, and do better by violating the norm strategically when it is in his interest to do so. In fact, the noninternalizer *could*, but will not *want* to emulate internalizers, in the sense that emulating their behavior simply does not maximize his preference function. To pursue the first example above, curbing one's violent tendencies may improve fitness, but the primordial preference function is not geared towards maximizing fitness, but rather towards a set of «fitness proxies» that entail being violent under earlier evolutionary but not contemporary circumstances. The noninternalizer will, of course curb his violence for prudential reasons, but not, because he in addition values peace, his neighbor's well-being, or even his biological fitness. In the second example, the noninternalizer will prefer the larger portion of meat, and the greater prestige that follows from a rigorous practice routine, but nevertheless, not enough to engage in such a routine.

Once genes for norm internalization are in place, there is nothing preventing altruistic norms from being culturally transmitted, internalized, and acted on in the same manner as personally fitness-enhancing norms. Altruism thus 'hitchhikes' on the personal fitness-enhancing capacity of norm internalization, and hence is an *exaption*, in the sense of Gould and Vrba (1981). It is for this reason that the rate of defection from altruistic norms might be sufficiently low that equation (1) might hold, as long as fitness costs are not too high and there is some positive level of oblique transmission[2].

It might be suggested that in a cultural equilibrium with internalized altruistic norms, a mutant family that teaches its children to internalize the personally fitness-enhancing norms but not the altruistic ones would out-compete families that transmit both personally fitness-enhancing and altruistic norms. However, if part of the ethic of altruism is to punish selfish types, even

2. By the same token, even antisocial norms can hitchhike on the internalization capacity, and they not infrequently do just that (Edgerton 1992). We show in Gintis (2003a) that the tendency of higher-fitness groups to out-compete lower fitness groups provides a strong tendency towards the circumscription of anti-social norms.

selfish types will act altruistically, so under plausible conditions, the mutant may have no adaptive advantage. Moreover, we show in Gintis (2003a) that, using the above notation, if

$$\alpha < \frac{\gamma - s}{1 + \gamma - s} \qquad (2)$$

selfish internalizers are positively disadvantaged with respect to altruistic internalizers.

8. Strong Reciprocity and Behavioral Game Theory

Recent experimental research supports the above synthesis[3]. We define *strong reciprocity* as a predisposition to cooperate with others, and to punish those who violate the norms of cooperation, at personal cost, even when it is implausible to expect that these costs will be repaid.

8.1. Evidence from the Ultimatum Game

In the ultimatum game, under conditions of anonymity, two players are shown a sum of money, say $10. One of the players, called the «proposer,» is instructed to offer any number of dollars, from $1 to $10, to the second player, who is called the «responder.» The proposer can make only one offer. The responder, again under conditions of anonymity, can either accept or reject this offer. If the responder accepts the offer, the money is shared accordingly. If the responder rejects the offer, both players receive nothing.

Since the game is played only once and the players do not know each other's identity, a self-interested responder will accept any positive amount of money. Knowing this, a self-interested proposer will offer the minimum possible amount, $1, and this will be accepted. However, when actually played, *the self-interested outcome is never attained and never even approximated*. In fact, as many replications of this experiment have documented, under varying conditions and with varying amounts of money, proposers routinely offer respondents very substantial amounts (50% of the total generally being the modal offer), and respondents frequently reject offers below 30% (Camerer and Thaler, 1995, Güth and Tietz, 1990, Roth *et al.*, 1991).

The ultimatum game has been played around the world, but mostly with university students. We find a great deal of individual variability. For instance, in all of the above experiments a significant fraction of subjects (about a quarter, typically) behave in a self-interested manner. But, among student subjects, average performance is strikingly uniform from country to country.

3. This material is developed in Gintis, Bowles, Boys and Fehr (2003).

To expand the diversity of cultural and economic circumstances of experimental subjects, Joseph Henrich, Samuel Bowles, Robert Boyd, Colin Camerer, Ernst Fehr, Herbert Gintis, and Richard McElreath (2001) undertook a large cross-cultural study of behavior in various games including the ultimatum game. Twelve experienced field researchers, working in twelve countries on four continents, recruited subjects from fifteen small-scale societies exhibiting a wide variety of economic and cultural conditions. These societies consisted of three foraging groups (the Hadza of East Africa, the Au and Gnau of Papua New Guinea, and the Lamalera of Indonesia), six slash-and-burn horticulturists (the Aché, Machiguenga, Quichua, and Achuar of South America, and the Tsimané and Orma of East Africa), four nomadic herding groups (the Turguud, Mongols, and Kazakhs of Central Asia, and the Sangu of East Africa) and two sedentary, small-scale agricultural societies (the Mapuche of South America and Zimbabwe farmers in Africa).

We can summarize our results as follows.

a. The canonical model of self-interested behavior is not supported in *any* society studied. In the ultimatum game, for example, in all societies either respondents, or proposers, or both, behaved in a reciprocal manner.
b. There is considerably more behavioral variability across groups than had been found in previous cross-cultural research. While mean ultimatum game offers in experiments with student subjects are typically between 43% and 48%, the mean offers from proposers in our sample ranged from 26% to 58%. While modal ultimatum game offers are consistently 50% among university students, sample modes with these data ranged from 15% to 50%. In some groups rejections were extremely rare, even in the presence of very low offers, while in others, rejection rates were substantial, including frequent rejections of *hyper-fair* offers (i.e. offers above 50%). By contrast, the most common behavior for the Machiguenga was to offer zero. The mean offer was 22%. The Aché and Tsimané distributions resemble American distributions, but with very low rejection rates. The Orma and Huinca (non-Mapuche Chileans living among the Mapuche) have modal offers near the center of the distribution, but show secondary peaks at full cooperation.
c. *Differences among societies in «market integration» and «cooperation in production» explain a substantial portion of the behavioral variation between groups*: the higher the degree of market integration and the higher the payoffs to cooperation, the greater the level of cooperation and sharing in experimental games. The societies were rank-ordered in five categories —«market integration» (how often do people buy and sell, or work for a wage), «cooperation in production» (is production collective or individual), plus «anonymity» (how prevalent are anonymous roles and transactions), «privacy» (how easily can people keep their activities secret), and «complexity» (how much centralized decision-making occurs above the level of the household). Using statistical regression analysis, only the first two

characteristics, market integration and cooperation in production, were significant, and they together accounted for 66% of the variation among societies in mean ultimatum game offers.

d. Individual-level economic and demographic variables did not explain behavior either within or across groups.
e. The nature and degree of cooperation and punishment in the experiments was generally consistent with economic patterns of everyday life in these societies.

In a number of cases the parallels between experimental game play and the structure of daily life were quite striking. Nor was this relationship lost on the subjects themselves. Here are some examples.

a. The Orma immediately recognized that the public goods game was similar to the *harambee*, a locally-initiated contribution that households make when a community decides to construct a road or school. They dubbed the experiment «the harambee game» and gave generously (mean 58% with 25% maximal contributors).
b. Among the Au and Gnau, many proposers offered more than half the pie, and many of these «hyper-fair» offers were rejected! This reflects the Melanesian culture of status-seeking through gift giving. Making a large gift is a bid for social dominance in everyday life in these societies, and rejecting the gift is a rejection of being subordinate.
c. Among the whale hunting Lamalera, 63% of the proposers in the ultimatum game divided the pie equally, and most of those who did not, offered more than 50% (the mean offer was 57%). In real life, a large catch, always the product of cooperation among many individual whalers, is meticulously divided into pre-designated parts and carefully distributed among the members of the community.
d. Among the Aché, 79% of proposers offered either 40% or 50%, and 16% offered more than 50%, with no rejected offers. In daily life, the Aché regularly share meat, which is being distributed equally among all other households, irrespective of which hunter made the kill.
e. The Hadza, unlike the Aché, made low offers and had high rejection rates in the ultimatum game. This reflects the tendency of these small-scale foragers to share meat, but with a high level of conflict and frequent attempts of hunters to hide their catch from the group.
f. Both the Machiguenga and Tsimané made low ultimatum game offers, and there were virtually no rejections. These groups exhibit little cooperation, exchange or sharing beyond the family unit. Ethnographically, both show little fear of social sanctions and care little about «public opinion.»
g. The Mapuche's social relations are characterized by mutual suspicion, envy, and fear of being envied. This pattern is consistent with the Mapuche's post-game interviews in the ultimatum game. Mapuche proposers rarely claimed that their offers were influenced by fairness, but rather by a fear

of rejection. Even proposers who made hyper-fair offers claimed that they feared rare spiteful responders, who would be willing to reject even 50/50 offers.

8.2. The Public Goods Game

The *public goods game* has been analyzed in a series of papers by the social psychologist Toshio Yamagishi (1986,1988), by the political scientist Elinor Ostrom and her coworkers (Ostrom *et al.*, 1992), and by economists Ernst Fehr and his coworkers (Gächter and Fehr 1999, Fehr and Gächter 2000, 2000). These researchers uniformly found that *groups exhibit a much higher rate of cooperation than can be expected assuming the standard economic model of the self-interested actor*, and this is especially the case when subjects are given the option of incurring a cost to themselves in order to punish free riders.

A typical public goods game consists of a number of rounds, say ten. The subjects are told the total number of rounds, as well as all other aspects of the game. The subjects are paid their winnings in real money at the end of the session. In each round, each subject is grouped with several other subjects —say 3 others— under conditions of strict anonymity. Each subject is then given a certain number of 'points,' say twenty, redeemable at the end of the experimental session for real money. Each subject then places some fraction of his points in a 'common account,' and the remainder in the subject's 'private account.' The experimenter then tells the subjects how many points were contributed to the common account, and adds to the private account of each subject some fraction, say 40%, of the total amount in the common account. So if a subject contributes his whole twenty points to the common account, each of the four group members will receive eight points at the end of the round. In effect, by putting the whole endowment into the common account, a player loses twelve points but the other three group members gain in total 24 ($= 8 \times 3$) points. The players keep whatever is in their private account at the end of the round.

A self-interested player will contribute nothing to the common account. However, only a fraction of subjects in fact conform to the self-interest model. Subjects begin by contributing on average about half of their endowment to the public account. The level of contributions decays over the course of the ten rounds, until in the final rounds most players are behaving in a self-interested manner (Dawes and Thaler, 1988, Ledyard, 1995). In a meta-study of twelve public goods experiments, Fehr and Schmidt (1999) found that in the early rounds, average and median contribution levels ranged from 40% to 60% of the endowment, but in the final period 73% of all individuals ($N=1042$) contributed nothing, and many of the remaining players contributed close to zero. These results are not compatible with the self-interested actor model, which predicts zero contribution on all rounds, though they might be predicted by a reciprocal altruism model, since the chance to reciprocate declines as the end of the experiment approaches. However this

is not in fact the explanation of moderate but deteriorating levels of cooperation in the public goods game.

The explanation of the decay of cooperation offered by subjects when debriefed after the experiment is that cooperative subjects became angry at others who contributed less than themselves, and retaliated against free-riding low contributors in the only way available to them —by lowering their own contributions (Andreoni, 1995).

Experimental evidence supports this interpretation. When subjects are allowed to punish noncontributors, they do so at a cost to themselves (Dawes, Orbell and Van de Kragt, 1986; Sato 1987; Yamagishi, 1988a,1988b,1992). For instance, in Ostrom *et al*. (1992) subjects interacted for twenty-five periods in a public goods game, and by paying a 'fee,' subjects could impose costs on other subjects by 'fining' them. Since fining costs the individual who uses it, but the benefits of increased compliance accrue to the group as a whole, the only Nash equilibrium in this game that does not depend on incredible threats is for no player to pay the fee, so no player is ever punished for defecting, and all players defect by contributing nothing to the common pool. However, the authors found a significant level of punishing behavior.

These studies allowed individuals to engage in strategic behavior, since costly punishment of defectors could increase cooperation in future periods, yielding a positive net return for the punisher. Fehr and Gächter (2000) set up an experimental situation in which *the possibility of strategic punishment was removed*. They used six and ten round public goods games with groups of size four, and with costly punishment allowed at the end of each round, employing three different methods of assigning members to groups. There were sufficient subjects to run between 10 and 18 groups simultaneously. Under the *Partner* treatment, the four subjects remained in the same group for all ten periods. Under the *Stranger* treatment, the subjects were randomly reassigned after each round. Finally, under the *Perfect Stranger* treatment the subjects were randomly reassigned and assured that they would never meet the same subject more than once. Subjects earned an average of about $35 for an experimental session.

Fehr and Gächter (2000) performed their experiment for ten rounds with punishment and ten rounds without[4]. Their results are illustrated in Figure 1. We see that when costly punishment is permitted, cooperation does not deteriorate, and in the Partner game, despite strict anonymity, cooperation increases almost to full cooperation, even on the final round. When punishment is not permitted, however, the same subjects experience the deterioration of cooperation found in previous public goods games. The contrast in cooperation rates between the Partner and the two Stranger treatments is worth noting, because the strength of punishment is roughly the same across all treatments.

4. For additional experimental results and analysis, see Bowles and Gintis (2002) and Fehr and Gächter (2002).

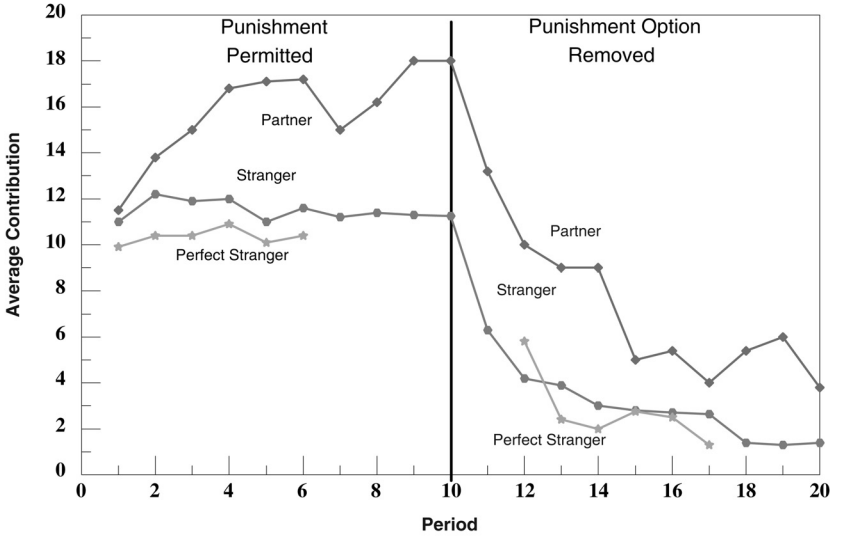

Figure 1. Average Contributions over Time in the Partner, Stranger, and Perfect Stranger Treatments when the Punishment Condition is Played First (adapted from Fehr and Gächter, 2000).

This suggests that the credibility of the punishment threat is greater in the Partner treatment because in this treatment the punished subjects are certain that, once they have been punished in previous rounds, the punishing subjects are in their group. The prosociality impact of strong reciprocity on cooperation is thus more strongly manifested, the more coherent and permanent the group in question.

9. Conclusion

Each of the behavioral disciplines contributes strongly to human behavioral science. Taken separately and at face value, however, they offer partial, conflicting, and incompatible models of human behavior. From a scientific point of view, it is scandalous that this situation was tolerated throughout most of the Twentieth Century. Fortunately, there is currently a strong current of unification based on both mathematical models and common methodological principles for gathering empirical data on human behavior and human nature.

The true power of each discipline's contribution to knowledge will only appear when suitably qualified and deepened by the contribution of the others. For instance, the economist's model of rational choice behavior must be qualified by a biological appreciation that preference consistency is the result of strong evolutionary forces, and where such forces are absent, consistency will

be imperfect and behavior must be augmented by empirical evidence. Moreover, *aprioristic* notions that preferences are self-regarding must be abandoned. These are the key tenets of behavioral economics. Second, the sociologist's notion of internalization of norms is generally rejected by the other behavioral disciplines because the ease with which diverse values can be internalized depends on human nature (Tooby and Cosmides, 1992, Pinker, 2002), and the rate at which values are acquired and abandoned depends on their contribution to fitness and well-being (Gintis, 2003b, Gintis, 2003a). Finally, there are often swift society-wide value changes that cannot be accounted for by socialization theory (Wrong, 1961, Gintis, 1975). When properly qualified, however, and appropriately related to the general theory of cultural evolution and strategic learning, the socialization theory is considerably strengthened.

Disciplinary boundaries in the behavioral sciences have been determined historically, rather than conforming to some consistent scientific logic. Perhaps for the first time, we are in a position to rectify this situation.

References

AINSLIE, George (1975). «Specious Reward: A Behavioral Theory of Impulsiveness and Impulse Control». *Psychological Bulletin* 82: 463-496.

AINSLIE, George; HASLAM, Nick (1992). «Hyperbolic Discounting». In: LOEWENSTEIN, George; ELSTER, Jon (eds.). *Choice Over Time.* New York: Russell Sage, p. 57-92.

AKERLOF, George A. (1991). «Procrastination and Obedience». *American Economic Review* 81,2: 1-19.

ALCOCK, John (1993). *Animal Behavior: An Evolutionary Approach.* Sunderland, MA: Sinauer.

ANDREONI, James (1995). «Cooperation in Public Goods Experiments: Kindness or Confusion». *American Economic Review* 85,4: 891-904.

ANDREONI, James; MILLER, John (2002). «Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism». *Econometrica* 70,2: 737-753.

ARROW, Kenneth J.; DEBREU, Gerard (1954). «Existence of an Equilibrium for a Competitive Economy». *Econometrica* 22,3: 265-290.

AUNGER, Robert (2002). *The Electric Meme: A New Theory of How We Think.* New York: Free Press.

AXELROD, Robert; HAMILTON, William D. (1981). «The Evolution of Cooperation». *Science* 211: 1390-1396.

BENEDICT, Ruth (1934). *Patterns of Culture.* Boston: Houghton Mifflin.

BERGSTROM, Theodore C.; STARK, Oded (1993). «How Altruism can Prevail in an Evolutionary Environment». *American Economic Review* 83,2: 149-155.

BONNER, John Tyler (1984). *The Evolution of Culture in Animals.* Princeton, NJ: Princeton University Press.

BOWLES, Samuel; GINTIS, Herbert (2002). «Homo Reciprocans». *Nature* 415: 125-128.

— (2005). «Prosocial Emotions». In: BLUME, Lawrence E.; DURLAUF, Steven N. (eds.). *The Economy As an Evolving Complex System III.* Santa Fe, NM: Santa Fe Institute.

BOYD, Robert; RICHERSON, Peter J. (1985). *Culture and the Evolutionary Process.* Chicago: University of Chicago Press.

BOYER, Pascal (2001). *Religion Explained: The Human Instincts That Fashion Gods, Spirits and Ancestors.* London: William Heinemann.
BROWN, Donald E. (1991). *Human Universals.* New York: McGraw-Hill.
BUCHANAN, James M.; TOLLISON, R.D. (1984). *The Theory of Public Choice.* Ann Arbor, MI: University of Michigan Press.
CALVIN, William H. (1983). «A Stone's Throw and its Launch Window: Timing Precision and its Implications for Language and Hominid Brains». *Journal of Theoretical Biology* 104: 121-135.
CAMERER, Colin; THALER, Richard (1995). «Ultimatums, Dictators, and Manners». *Journal of Economic Perspectives* 9,2: 209-219.
CAVALLI-SFORZA, Luca L.; FELDMAN, Marcus W. (1982). «Theory and Observation in Cultural Transmission». *Science* 218: 19-27.
COLEMAN, James S. (1990). *Foundations of Social Theory.* Cambridge, MA: Belknap.
CONLISK, John (1988). «Optimization Cost». *Journal of Economic Behavior and Organization* 9: 213-228.
DAMASIO, Antonio R. (1994). *Descartes' Error: Emotion, Reason, and the Human Brain.* New York: Avon Books.
DAWES, Robyn M.; THALER, Richard (1988). «Cooperation». *Journal of Economic Perspectives* 2: 187-197.
DOWNS, Anthony (1957). *An Economic Theory of Democracy.* Boston: Harper & Row.
DUGATKIN, Lee Alan; KERN REEVE, Hudson (1998). *Game Theory and Animal Behavior.* Oxford: Oxford University Press.
DURHAM, William H. (1991). *Coevolution: Genes, Culture, and Human Diversity.* Stanford: Stanford University Press.
DURKHEIM, Emile (1951). *Suicide, a Study in Sociology.* New York: Free Press.
EDGERTON, Robert B. (1992). *Sick Societies: Challenging the Myth of Primitive Harmony.* New York: The Free Press.
ELSTER, Jon (1979). *Ulysses and the Sirens: Studies in Rationality and Irrationality.* Cambridge, UK: Cambridge University Press.
ESHEL, Ilan; FELDMAN, Marcus W. (1984). «Initial Increase of New Mutants and Some Continuity Properties of ESS in two Locus Systems». *American Naturalist* 124: 631-640.
ESHEL, Ilan; FELDMAN, Marcus W.; BERGMAN, Aviv (1998). «Long-term Evolution, Short-term Evolution, and Population Genetic Theory». *Journal of Theoretical Biology* 191: 391-396.
FEHR, Ernst; GÄCHTER, Simon (2000). «Cooperation and Punishment». *American Economic Review* 90,4: 980-994.
— (2002). «Altruistic Punishment in Humans». *Nature* 415: 137-140.
FEHR, Ernst; SCHMIDT, Klaus M. (1999). «A Theory of Fairness, Competition, and Cooperation». *Quarterly Journal of Economics* 114: 817-868.
FELDMAN, Marcus W.; ZHIVOTOVSKY, Lev A. (1992). «Gene-Culture Coevolution: Toward a General Theory of Vertical Transmission». *Proceedings of the National Academy of Sciences* 89: 11935-11938.
FISHER, Ronald A. (1930). *The Genetical Theory of Natural Selection.* Oxford: Clarendon Press.
FUDENBERG, Drew; MASKIN, Eric (1986). «The Folk Theorem in Repeated Games with Discounting or with Incomplete Information». *Econometrica* 54,3: 533-554.
GÄCHTER, Simon; FEHR, Ernst (1999). «Collective Action as a Social Exchange». *Journal of Economic Behavior and Organization* 39,4: 341-369.

GIGERENZER, Gerd; SELTEN, Reinhard (2001). *Bounded Rationality.* Cambridge, MA: MIT Press.
GINTIS, Herbert (1975). «Welfare Economics and Individual Development: A Reply to Talcott Parsons». *Quarterly Journal of Economics* 89,2: 291-302.
— (2000). *Game Theory Evolving.* Princeton, NJ: Princeton University Press.
— (2003). «The Hitchhiker's Guide to Altruism: Genes, Culture, and the Internalization of Norms». *Journal of Theoretical Biology* 220,4: 407-418.
— (2003). «Solving the Puzzle of Human Prosociality». *Rationality and Society* 15,2: 155-187.
GINTIS, Herbert; SMITH, Eric Alden; BOWLES, Samuel (2001). «Costly Signaling and Cooperation». *Journal of Theoretical Biology* 213: 103-119.
GINTIS, Herbert; BOWLES, Samuel; BOYD, Robert; FEHR, Ernst (2003). «Explaining Altruistic Behavior in Humans». *Evolution & Human Behavior* 24: 153-172.
— (2005). *Moral Sentiments and Material Interests: On the Foundations of Cooperation in Economic Life.* Cambridge: The MIT Press.
GOULD, Stephen J.; VRBA, Elizabeth (1981). «Exaption: A Missing Term in the Science of Form». *Paleobiology* 8: 4-15.
GRUSEC, Joan E.; KUCZYNSKI, Leon (1997). *Parenting and Children's Internalization of Values: A Handbook of Contemporary Theory.* New York: John Wily & Sons.
GÜTH, Werner; TIETZ, Reinhard (1990). «Ultimatum Bargaining Behavior: A Survey and Comparison of Experimental Results». *Journal of Economic Psychology* 11: 417-449.
HALDANE, J.B.S. (1932). *The Causes of Evolution.* London: Longmans, Green & Co.
HAMILTON, William D. (1963). «The Evolution of Altruistic Behavior». *American Naturalist* 96: 354-356.
— (1967). «Extraordinary Sex Ratios». *Science* 156: 477-488.
HAMMERSTEIN, Peter (1996). «Darwinian Adaptation, Population Genetics and the Streetcar Theory of Evolution». *Journal of Mathematical Biology* 34: 511-532.
HAMMERSTEIN, Peter; SELTEN, Reinhard (1994). «Game Theory and Evolutionary Biology». In AUMANN, Robert J.; HART, Sergiu (eds.). *Handbook of Game Theory with Economic Applications.* Amsterdam: Elsevier, p. 929-993.
HECHTER, Michael; KANAZAWA, Satoshi (1997). «Sociological Rational Choice». *Annual Review of Sociology* 23: 199-214.
HENRICH, Joe; BOYD, Robert; BOWLES, Samuel; CAMERER, Colin; FEHR, Ernst; GINTIS, Herbert (2004). *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-scale Societies.* Oxford: Oxford University Press.
HENRICH, Joe; BOYD, Robert; BOWLES, Samuel; CAMERER, Colin; FEHR, Ernst; GINTIS, Herbert; MCELREATH, Richard (2001). «Cooperation, Reciprocity and Punishment in Fifteen Small-scale Societies». *American Economic Review* 91: 73-78.
HERRNSTEIN, Richard J. (1961). «Relative and Absolute Strengths of Responses as a Function of Frequency of Reinforcement». *Journal of Experimental Analysis of Animal Behavior* 4: 267-272.
HUANG, Chi-Fu; LITZENBERGER, Robert H. (1988). *Foundations for Financial Economics.* Amsterdam: Elsevier.
JABLONKA, Eva; LAMB, Marion J. (1995). *Epigenetic Inheritance and Evolution: The Lamarckian Case.* Oxford: Oxford University Press.
— (1998). «Epigenetic Interitance in Evolution». *Journal of Evolutionary Biology* 11: 159-183.

KIYONARI, Toko; TANIDA, Shigehito; YAMAGISHI, Toshio (2000). «Social Exchange and Reciprocity: Confusion or a Heuristic?». *Evolution and Human Behavior* 21: 411-427.

KOLLOCK, Peter (1997). «Transforming Social Dilemmas: Group Identity and Cooperation». In: DANIELSON, Peter (ed.). *Modeling Rational and Moral Agents*. Oxford: Oxford University Press.

KREBS, J.R.; DAVIES, N.B. (1997). *Behavioral Ecology: An Evolutionary Approach.* Oxford: Blackwell Science, fourth ed.

KREPS, David M. (1990). *A Course in Microeconomic Theory.* Princeton, NJ: Princeton University Press.

LAIBSON, David (1997). «Golden Eggs and Hyperbolic Discounting». *Quarterly Journal of Economics* 112,2: 443-477.

LEDYARD, J.O. (1995). «Public Goods: A Survey of Experimental Research». In: KAGEL, J.H.; ROTH, A.E. (eds.). *The Handbook of Experimental Economics.* Princeton, NJ: Princeton University Press, p. 111-194.

LEWONTIN, Richard C. (1961). «Evolution and the Theory of Games». *Journal of Theoretical Biology* 1: 382-403.

— (1974). *The Genetic Basis of Evolutionary Change.* New York: Columbia University Press.

LIBERMAN, Uri (1988). «External Stability and ESS Criteria for Initial Increase of a New Mutant Allele». *Journal of Mathematical Biology* 26: 477-485.

LUMSDEN, C.J.; WILSON, E.O. (1981). *Genes, Mind, and Culture: The Coevolutionary Process.* Cambridge, MA: Harvard University Press.

MANDEVILLE, Bernard (1924). *The Fable of the Bees: Private Vices, Publick Benefits.* Oxford: Clarendon [1705].

MAYNARD SMITH, John (1982). *Evolution and the Theory of Games.* Cambridge, UK: Cambridge University Press.

MAYNARD SMITH, John; PRICE, G.R. (1973). «The Logic of Animal Conflict». *Nature* 246: 15-18.

MEAD, Margaret (1963). *Sex and Temperament in Three Primitive Societies.* New York: Morrow.

MEALEY, Linda (1995). «The Sociobiology of Sociopathy». *Behavioral and Brain Sciences* 18: 523-541.

MONROE, Kristen Renwick (1991). *The Economic Approach to Politics.* Reading, MA: Addison Wesley.

MOORE, Jr., Barrington (1978). *Injustice: The Social Bases of Obedience and Revolt.* White Plains: M. E. Sharpe.

MORAN, P.A.P. (1964). «On the Nonexistence of Adaptive Topographies». *Annals of Human Genetics* 27: 338-343.

O'DONOGHUE, Ted; RABIN, Matthew (2001). «Choice and Procrastination». *Quarterly Journal of Economics* 116,1: 121-160.

OLSON, Mancur (1965). *The Logic of Collective Action: Public Goods and the Theory of Groups.* Cambridge, MA: Harvard University Press.

ORBELL, John M.; DAWES, Robyn M.; VAN DE KRAGT, J.C. (1986). «Organizing Groups for Collective Action». *American Political Science Review* 80: 1171-1185.

OSTROM, Elinor; WALKER, James; GARDNER, Roy (1992). «Covenants with and without a Sword: Self-Governance Is Possible». *American Political Science Review* 86,2: 404-417.

PARSONS, Talcott (1964). «Evolutionary Universals in Society». *American Sociological Review* 29,3: 339-357.
— (1967). *Sociological Theory and Modern Society.* New York: Free Press.
PINKER, Steven (2002). *The Blank Slate: The Modern Denial of Human Nature.* New York: Viking.
POWER, T.G.; CHAPIESKI, M.L. (1986). «Childrearing and Impulse Control in Toddlers: A Naturalistic Investigation». *Developmental Psychology* 22: 271-275.
RABIN, Matthew (1993). «Incorporating Fairness into Game Theory and Economics». *American Economic Review* 83,5: 1281-1302.
RICHERSON, Peter J.; BOYD, Robert (1998). «The Evolution of Ultrasociality». In: EIBL-EIBESFELDT, I.; SALTER, F.K. (eds.). *Indoctrinability, Idology and Warfare.* New York: Berghahn Books, p. 71-96.
ROGERS, Alan (1994). «Evolution of Time Preference by Natural Selection». *American Economic Review* 84,3: 460-481.
ROTH, Alvin E.; PRASNIKAR, Vesna; OKUNO-FUJIWARA, Masahiro; ZAMIR, Schmuel (1991). «Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study». *American Economic Review* 81,5: 1068-1095.
SATO, Kaori (1987). «Distribution and the Cost of Maintaining Common Property Resources». *Journal of Experimental Social Psychology* 23: 19-31.
SCHRÖDINGER, Edwin (1944). *What is Life?: The Physical Aspect of the Living Cell.* Cambridge: Cambridge University Press.
SIMON, Herbert (1972). «Theories of Bounded Rationality». In: MCGUIRE, C.B.; RADNER, Roy (eds.). *Decision and Organization.* New York: American Elsevier, p. 161-176.
SMITH, Adam (2000). *The Theory of Moral Sentiments.* New York: Prometheus [1759].
STEPHENS, W.; MCLINN, C.M.; STEVENS, J.R. (2002).«Discounting and Reciprocity in an Iterated Prisoner's Dilemma». *Science* 298: 2216-2218.
STROTZ, Robert H. (1955). «Myopia and Inconsistency in Dynamic Utility Maximization». *Review of Economic Studies* 23,3: 165-180.
TAYLOR, P.; JONKER, L. (1978). «Evolutionarily Stable Strategies and Game Dynamics». *Mathematical Biosciences* 40: 145-156.
TOMASELLO, Michael (1999). *The Cultural Origins of Human Cognition.* Cambridge, MA: Harvard University Press.
TOOBY, John; COSMIDES, Leda (1992). «The Psychological Foundations of Culture». In: BARKOW, Jerome H.; COSMIDES, Leda; TOOBY, John (eds.). *The Adapted Mind: Evolutionary Psychology and the Generation of Culture.* New York: Oxford University Press, p. 19-136.
TRIVERS, Robert L. (1971). «The Evolution of Reciprocal Altruism». *Quarterly Review of Biology* 46: 35-57.
TVERSKY, Amos; KAHNEMAN, Daniel (1981). «Loss Aversion in Riskless Choice: A Reference-Dependent Model». *Quarterly Journal of Economics* 106,4: 1039-1061.
WINTER, Sidney G. (1971). «Satisficing, Selection and the Innovating Remnant». *Quarterly Journal of Economics* 85: 237-261.
WOOD, Elisabeth Jean (2003). *Insurgent Collective Action and Civil War in El Salvador.* Cambridge,: Cambridge University Press.
WRIGHT, Sewall (1931). «Evolution in Mendelian Populations». *Genetics* 6: 111-178.
WRONG, Dennis H. (1961). «The Oversocialized Conception of Man in Modern Sociology». *American Sociological Review* 26: 183-193.

YAMAGISHI, Toshio (1986). «The Provision of a Sanctioning System as a Public Good». *Journal of Personality and Social Psychology* 51: 110-116.
— (1988). «The Provision of a Sanctioning System in the United States and Japan». *Social Psychology Quarterly* 51,3: 265-271.
— (1988). «Seriousness of Social Dilemmas and the Provision of a Sanctioning System». *Social Psychology Quarterly* 51,1: 32-42.
— (1992). «Group Size and the Provision of a Sanctioning System in a Social Dilemma». In: LIEBRAND, W.B.G.; MESSICK, David M.; WILKE, H.A.M. (eds.). *Social Dilemmas: Theoretical Issues and Research Findings.* Oxford: Pergamon Press, p. 267-287.
YOUNG, H. Peyton (1998). *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions.* Princeton, NJ: Princeton University Press.