Review of Behavioral Economics, 2015, 2: 1–59

Homo Socialis: An Analytical Core for Sociological Theory

Herbert Gintis[1] and Dirk Helbing[2*]

[1] *Santa Fe Institute, USA*
[2] *ETH Zürich, Switzerland*

> Social life comes from a double source, the likeness of consciences and the division of social labor.
>
> Emile Durkheim

> We are caught in an inescapable network of mutuality, tied in a single garment of destiny.
>
> Martin Luther King

> How selfish soever man may be supposed, there are evidently some principles in his nature, which interest him in the fortune of others, and render their happiness necessary to him, though he derives nothing from it except the pleasure of seeing it.
>
> Adam Smith

ABSTRACT

We develop an analytical core for sociology. We follow standard dynamical systems theory by first specifying the conditions for social equilibrium, and then studying the dynamical principles that govern disequilibrium behavior. Our general social equilibrium model is an expansion of the general equilibrium model of economic theory, and our dynamical principles treat the society as a complex adaptive system that can be studied using evolutionary game theory and agent-based Markov models based on variants of the replicator dynamic.

Keywords: Social theory

*We would like to thank Samuel Bowles and Barkley Rosser for insightful suggestions. Dirk Helbing acknowledges support by the ERC Advanced Investigator Grant 'Momentum' (Grant No. 324247).

ISSN 2326-6198; DOI 10.1561/105.00000016

1 Introduction

Modern societies are complex dynamical systems in which social institutions are modified through high-level political decision-making and popular collective action (Helbing *et al.*, 2005). We offer here an analytical framework for modeling the structure and dynamics of modern societies. We follow standard dynamical systems theory by first specifying the conditions for social equilibrium, and then studying the dynamical principles that govern disequilibrium behavior. Our *general social equilibrium model* is patterned after the highly successful Walrasian general equilibrium model (Arrow and Debreu, 1954), and our dynamical principles can be modeled using evolutionary game theory (Weibull, 1995; Helbing, 1995; Gintis, 2009b) and agent-based Markov models based on variants of the replicator dynamic (Helbing, 1996, 2010; Gintis, 2013).

Talcott Parsons initiated the formal modeling of modern societies in *The Structure of Social Action* (1937) and *Toward a General Theory of Social Action* (1951). This brilliant effort foundered, however, for reasons unrelated to the scientific value of his project. First, Parsons lacked analytical decision theory, stemming from Savage (1954), as well as game theory, which developed following Nash (1950). He also lacked an appreciation for general equilibrium theory, which came to fruition in the mid-1950's (Arrow and Debreu, 1954). These powerful tools together allow us to formulate an analytical core for sociology. Second, Parsons followed Vilfredo Pareto (1896, 1906) in maintaining a strict separation between preferences over *economic values*, based on material self-interest on the one hand and *social, political, and moral values*, involving concern for social life in the broadest sense, on the other. This separation precludes any general model of rational choice and social action (Lindenberg, 1983, 2004; Fehr and Gintis, 2007; Gintis, 2009a).

2 Summary: A Core Analytical Model

We draw on several scientific traditions for creating a core analytical model for sociology. The first is the work of sociologists Max Weber, Emile Durkheim, George Herbert Mead, Ralph Linton, Talcott Parsons and others, whose insights have so far largely escaped analytical expression and are little known, despite their extreme relevance, beyond the sociology discipline. The second is our model of individual choice behavior, which is a broadened version of rational decision theory (Savage, 1954; Fishburn, 1970; Gintis, 2009a). The two behavioral disciplines that include a core analytical model, biology and economics, are built around the notion of rational choice. This theory is useful in conjuction with game theory which, while widely applied in sociobiology in general (Alcock, 1993; Krebs and Davies, 1997; Dugatkin and Reeve, 1998), is especially important for humans (Bowles and Gintis, 2011; Wilson, 2012)

because *Homo sapiens* is not only *Homo socialis*, but also *Homo ludens*: our species has the capacity to construct novel games with great flexibility and its members possess the cognitive and moral requirements for game-playing. Our major innovation in this respect is to expand on Thomas Schelling's notion of a *focal point equilibrium* (Schelling, 1960) by proposing the *correlated equilibrium*, rather than the more standard *Nash equilibrium*, as the basis of an analytical model of social norms (Aumann, 1987a; Gintis, 2009a).

The third tradition is the general economic equilibrium model of Walras (1874), Arrow and Debreu (1954), and others, which is analytically rigorous and mathematically elegant. Despite its appearance of extreme abstraction, it is in fact capable of a surprisingly straightforward and plausible extension to a general social equilibrium model of considerable sophistication.

Modeling social dynamics is significantly more challenging than modeling social equilibrium because human society has the key characteristic of a complex dynamical system: it consists of many structurally similar, strongly interacting and intricately networked units (*social actors*), operating in parallel with little centralized structural control (Miller and Page, 2007). Such complex systems generically exhibit emergent properties at the macrosystem level that resist analytical derivation from the behavior of the individual parts (Morowitz, 2002).

The fourth intellectual strand upon which we draw, a field that did not exist until recently, is epistemic and evolutionary game theory (Aumann, 1987b; Weibull, 1995; Gintis, 2009b; Grund *et al.*, 2013) and the use of agent-based simulations and the study of Markov models of stochastic behavior for empirical validation (Helbing, 1995, 2012; Gintis, 2009b, 2013).

Our fifth foundational element is behavioral game theory (Camerer, 2003; Gintis, 2009a), based on laboratory and field experimentation into choice and social interaction, which provides the empirical basis for the generalization of rational choice theory to include moral, social and *other-regarding* values (Camerer and Fehr, 2004; Fehr and Gintis, 2007).

2.1 Gene-Culture Coevolution

The predominant mode of acquisition of behaviors in the biological world is genetic transmission, sometimes followed by learning through experience. Some non-human mammals also transmit behaviors that they have learned to their off-spring, although such learned behaviors have a tenuous existence and tend to last for only a few generations at best (Bonner, 1984; Cavalli-Sforza, 1986).

Humans are distinct in the world of social species by having developed means to preserve cultural forms across generations by embodying them in tools, physical constructions, and language (Boyd and Richerson, 1985). In addition, humans have specialized cognitive structures, including the capacity to infer the mental states and intentions of others, that strengthen the cultural

transition process (Tomasello *et al.*, 2005). Because humans have enjoyed cumulative culture for much of their evolutionary existence, they have evolved complex social structures that serve as the background conditions for genetic evolution. It follows that individual fitness in humans depends on the structure of social life.

Because culture is both constrained and promoted by the human genome, human cognitive, affective and moral capacities are the product of an evolutionary dynamic involving the interaction of genes and culture. We call this dynamic *gene-culture coevolution* (Boyd and Richerson, 1985, 2004; Cavalli-Sforza and Feldman, 1982; Dunbar, 1993).

This coevolutionary process has endowed us with the cognitive capacities and predispositions to adopt culturally-fostered personal preferences that go beyond the *self-regarding* concerns emphasized in traditional economic and biological theory. This coevolutionary process also endows us with a social epistemology that facilitates the sharing of intentionality across minds. Gene-culture coevolution is responsible for the salience of such other-regarding values as a taste for cooperation, fairness and retribution, the capacity to empathize, and the ability to value such *character virtues* as honesty, hard work, piety and loyalty (Boehm, 1999; Fehr and Gintis, 2007).

2.2 Rational Decision Theory

We model choice behavior using the rational actor model, according to which individuals have a time-, state-, and social context-dependent *preference function* over *actions*, *payoffs*, and *beliefs* concerning the probabilistic effects of actions on outcomes. Individuals value payoffs besides the material goods and services depicted in economic theory, including aspects of actions that are valued for their own sake. For example, there are character virtues, including honesty, loyalty, trustworthiness, courage, and considerateness, that have intrinsic value for many individuals, independent of their effect of their application on others and in addition to any other welfare effects on themselves, including personal reputation effects. Moreover, social actors generally value not only self-regarding payoffs such as personal income and leisure, but also other-regarding payoffs, such as the welfare of others, environmental integrity, fairness, reciprocity, and conformance with social norms.

The rational choice model expresses but does not explain individual preferences. Understanding the content of preferences requires rather deep forays into the psychology of goal-directed and intentional behavior (Haidt, 2012), social evolutionary theory (Cosmides and Tooby, 1992), and problem-solving heuristics (Gigerenzer and Todd, 1999). Moreover, the social actor's preference function will generally depend both on his current motivational state, including his previous experience and future plans, and on the social situation that he faces.

The first principle of rational choice is that in any given situation, which may be time-, state-, and social-context dependent, the decision-maker, whom we will call Alice, has a *preference relation* $\succ$ over choices such that Alice prefers x to y if and only if $x \succ y$. The conditions for the existence of such a relation, developed in Section A1, are quite minimal, the main point being that Alice's choices must be *transitive* in the sense that if the choice set from which Alice must choose is X with $x, y, z \in X$, then if Alice prefers x to y, and also prefers y to z, then Alice must prefer x to z as well. An additional requirement is that if Alice prefers x to y when the choice set is X, she must continue to prefer x to y in any choice set that includes both x and y. This condition can fail if the choice set itself represents a substantive social context that affects the value Alice places upon x and y. For instance, Alice may prefer fish (x) to steak (y) in a restaurant that also serves lobster (z) because the fish is likely to be very fresh in this case, whereas in a restaurant that does not serve lobster, the fish is likely to be less fresh, so Alice prefers steak (y) to fish (x). In cases such as this, a more sophisticated representation of choice sets and outcomes both satisfies the rationality assumptions and more insightfully models Alice's social choice situation.

Note that the preference function derived from the above simple axioms does not suggest that Alice chooses what is in her best interest or even what gives her pleasure. There are simply no utilitarian implications of these axioms. Nor does the analysis assume that Alice is in any sense selfish, calculating, or amoral. Finally, the rationality assumption does not suggest that Alice is "trying" to maximize utility or anything else. The maximization formulation of rational choice behavior is simply an analytical convenience, akin to the least action principle in classical mechanics, or predicting the behavior of an expert billiards player by solving a set of differential equations.

The second principle of rational choice applies when Alice's choice involves probabilistic payoffs. Suppose also that we have a set of alternative possible *states of nature* E with elements e_1, e_2, e_3, and so on, that can possibly materialize, and a set of payoffs X. A *lottery* is a mapping that specifies a particular payoff $x \in X$ for each state $e \in E$. We write the set of such lotteries as $\mathcal{L}$, so any lottery $\pi \in \mathcal{L}$ gives Alice payoff $x_1 = \pi(e_1)$ in case e_1 occurs, $x_2 = \pi(e_2)$ in case e_2 occurs, $x_3 = \pi(e_3)$ in case c occurs, and so on. By our first rationality assumption, Alice has a consistent preference function over the set of lotteries $\mathcal{L}$. We add to this a few rather innocuous assumptions concerning Alice's preferences (see Section A1) that together imply that Alice has a consistent preference function $u(x)$ over the various outcomes in X and also Alice attaches a specific probability $p(e)$ to each of the events in E. We call this probability distribution Alice's *subjective prior* over the events in E. Moreover, given the preference function $u(x)$ and the subjective prior $p(e)$, Alice prefers lottery π to lottery ρ, that is $\pi \succ \rho$, precisely when the expected utility of π exceeds that of lottery ρ (see equation A1).

There are only two substantive assumptions in the above derivation of the expected utility theorem. The first is that Alice does not suffer from *wishful thinking*. That is, the probability that Alice implicitly attaches to a particular outcome by her preference function over lotteries does not depend on how much she stands to gain or lose should that outcome occur. This assumption is certainly not always justified. For instance, believing that she might win the state lottery may give Alice more pleasure while waiting for it to happen that the cost of buying the lottery ticket. Moreover, there may be situations in which Alice will underinvest in a desirable outcome unless she inflates the probability that the investment will pay off (Benabou and Tirole, 2002). In addition, Alice may be substantively irrational, have excessive confidence that the world conforms to her ideological preconceptions.

The second substantive assumption is that the state of nature that materializes is not affected by Alice's choice of a lottery. When this fails, we must interpret the subjective prior as a *conditional probability*, in terms of which the expected utility theorem remains valid (Stalnaker, 1968). This form of the expected utility theorem is developed in Section A1.

Of course, a social actor may be rational in this decision-theoretic sense, having transitive preferences and not engaging in wishful thinking, and still fail to conform to higher canons of rationality. Alice may, for instance, make foolish choices that thwart her larger objectives and threaten her well-being. She may be poorly equipped to solve challenging optimization problems. Moreover, being rational in the decision-theoretic sense does not imply that Alice's beliefs are in any way reasonable, or that she evaluates new evidence in an insightful manner.

The standard axioms underlying the rational actor model are developed in Savage (1954). We discuss the plausibility and generality of these axioms in Section A1. In Section 4 we explore the implications of replacing Savage's assumption that beliefs are purely personal "subjective probabilities" with the notion that the individual is generally embedded in a network of social actors over which information and experience concerning the relationship between actions and outcomes is spread. The rational actor thus draws on a network of beliefs and experiences distributed among the social actors to which he is informationally and socially connected. By the sociological principle of *homophily*, social actors are likely to structure their network of personal associates according to principles of social similarity, and to alter personal tastes in the direction of increasing compatibility with networked associates (McPherson *et al.*, 2001; Durrett and Levin, 2005; Fischer *et al.*, 2013).

2.3 The Social Division of Labor

The social division of labor is a network of interacting *social roles* (Mead, 1934; Linton, 1936; Parsons and Shils, 1951). The *content* of a social role is a set of

rights, duties, expectations, material and symbolic rewards, and behavioral norms. In equilibrium, the content of all social roles is public information shared by all members of society, and this content influences the mutual expectations of individuals involved in social interaction. In periods of social change, by contrast, the content of particular roles may be subject to contrasting expectations and the process of re-establishing a common understanding of role-contents involves dialog, collective action, cultural conflict, and the exercise of political power. For instance, the role of secretary may in one period include the restriction that role-occupants be females who make coffee and run personal errands for their superiors. In a period of feminist social action, these aspects of the content of a secretary's role may be dropped, while other aspects, such as preparing documents and managing appointments may remain or even become enhanced.

Role-occupants are *actors* who fill many different and contrasting roles in the course of performing their daily activities. An individual may perform as spouse preparing breakfast, as parent advising children on the day's activities, as sales manager in an enterprise, as school committee and church member, and as voter.

Actors are in general rational decisions-makers who maximize their preference functions subject to the content of the social roles they occupy, and given a belief system that is context-dependent and governed by the expectations defined by the actor's social location. These decisions determine the social actors' role-specific behaviors. For instance, when one engages a taxi in a strange city, both the driver and the client may know exactly what is expected of each, so no time or energy is wasted bargaining or otherwise adjudicating mutually acceptable behavior.

The distribution of social roles and the association of social actors with particular social roles can be modeled by appropriately enriching the *Walrasian general equilibrium model* of economic theory (Walras, 1874; Arrow and Debreu, 1954). In this general economic equilibrium model, actors are owners of productive resources, which they supply to firms, and they are consumers of the goods and services produced by firms. Productive resources include capital goods, raw materials, and various sorts of labor services. Firms in this model combine the productive resources to generate marketable commodities. Firms choose a pattern of inputs and outputs to maximize profits, given the price structure they face. Social actors in this model choose their pattern of consumption, as well as their supply of services to firms, to maximize their preference functions at given prices. An economic equilibrium occurs when prices are such that the plans of all agents are simultaneously met at the posted prices.

The general economic equilibrium model becomes the special case of a *general social equilibrium model* by identifying the firm with a set of social roles and identifying suppliers of services to firms with social actors who occupy

these roles. This sociological broadening of general economic equilibrium is quite natural, because it is reasonable to view a position in the firm as a social role whose content includes not only the salary and the employee's obligation to come to work, but also a set of rights, behavioral norms, as well as a pattern of symbolic rewards and sanctions determined by the culture of the firm and the larger society. While interpreters generally stress the price system as the key element in adjudicating among the interests of economic actors, the theory becomes more powerful if we view the general content of social roles adjusting when out of equilibrium (Granovetter, 1985, 1995; DiMaggio, 1994, 1998; Hechter and Kanazawa, 1997; Hedström and Bearman, 2009).

The general economic equilibrium model recognizes only one social institution: profit-maximizing firms. Families in this model are treated as "black boxes," as is government, if it is treated at all. For the general social equilibrium model we must add at a minimum *families*, *communities*, as well as *public institutions* and *private associations*, such as governmental, religious, scientific, charitable, and cultural organizations. These organizations are constrained in their internal organization of social roles to maintain a positive balance sheet, but otherwise can determine their organization of social roles according to criteria other than profitability. A theory of the family, for instance, would suggest how the limits of family membership are determined, what social roles are occupied by family members, and how content of these roles is determined.

The general economic equilibrium model assumes that in equilibrium all agents have perfect information concerning the nature of the goods and services they exchange and the prices at which they exchange. The same must be true of a general social equilibrium model. Out of equilibrium, however, the content of social roles, including their material, social, and moral attributes, are statistical distributions over which individuals have subjective and networked probability distributions. This corresponds to the fact that in the general economic equilibrium model, out of equilibrium there is no basis for forming price expectations except by networked experience, which may differ significantly across economic agents (Gintis, 2007a). For instance, in deciding whether to take a job at wage w, the worker must consider the return to continuing job search, which will depend on the statistical distribution of demand for labor in the economy. The worker has only his networked experience to estimate this distribution, and such experience can vary widely among workers with similar credentials and demographics.

In general social equilibrium, each actor maximizes his preference function in the sense that no change of role will increase his expected payoff, taking into account possible search and relocation costs, and the pattern of supply and demand for social roles will be such that expected payoffs will not change over time. In addition, if there are institutions, such as firms, hospitals, families, communities, or governments, these institutions may have certain social

conditions that must be satisfied in equilibrium, such as a balance between expenditures and receipts, or achievement of certain institutional goals.

In proposing the actor/role model, sociologists have traditionally held that the major difference between social and economic roles is that social roles function properly only by virtue of the moral commitments of role-occupants, whereas economic roles function independently from role-occupants' social conscience and moral commitments. To achieve its purported independence from moral commitment, general economic equilibrium models make the implausible assumptions of *complete contracts*, meaning that any contract between individuals, however complex, covers all possible contingencies and can be enforced by a third party (the judicial system) at no cost to the contracting parties (Bowles and Gintis, 1993). When we drop this assumption from the general economic equilibrium model, moral commitments become as salient in economic life as they are in social life in general (Bowles and Gintis, 1993; Brown *et al.*, 2004; Gintis, 2009a).

The major effect of conceiving of the general network of social roles as an expansion of the general economic equilibrium model is the clarification it lends to the distinction between equilibrium and dynamic models of society. The general economic equilibrium model is a static construct that gives no suggestion as to how equilibrium might be attained. This is a critical limitation, just as is the parallel limitation of the general social equilibrium model developed in this paper. While Gintis and Mandel (2012) provide a plausible dynamic for the general economic equilibrium model and prove the stability of equilibrium for this dynamic, this proof does not extend to the general social equilibrium model.

2.4 The Socio-Psychological Theory of Norms

Durkheim (1902) was the first to recognize the social tension in modern society caused by an increasingly differentiated social role structure — the social division of labor — created the need for a common base of social beliefs and values, which he terms *collective consciousness*, to promote social harmony and efficient cooperation. Durkheim's theme was developed into a theory of *social norms* by Linton (1936) and Mead (1934), and integrated into a general social theory by Parsons (1937). Social norms are often promulgated by a nexus of system-wide cultural institutions and social processes that in equilibrium produce a consistent set of expectations and normative predispositions across all social actors. The *socio-psychological theory of norms* models this social subsystem and accounts for their effectivity. Other social norms govern well-defined subsets of the population, such as religious groups, professional associations, and sports. Out of equilibrium, conflicting social norms often vie for dominance, and cultural dynamics are often the result of these conflicts (Winter *et al.*, 2012).

In the first instance, the complex of social norms has an *instrumental* character devoid of normative content, serving merely as an informational device that coordinates the behavior of rational agents (Lewis, 1969; Gauthier, 1986; Binmore, 2005; Bicchieri, 2006). Social norms thus supply the general factual descriptions of the content of many standard social roles (employer, worker, mother, judge, traffic cop, taxi driver, and the like), allowing social actors to coordinate their behavior even when dealing with unfamiliar social partners in novel situations. Social norms thus create common subjective priors that facilitate general social cooperation.

However in many social roles high level performance requires that the actor have a personal commitment to role performance that cannot be captured by the self-regarding "public" rewards and penalties associated with the role (Conte and Castelfranchi, 1999; Gintis, 2009a). For instance, a physician may be obligated to ignore personal gain when suggesting medical procedures, only the most egregious of violations of which will incur serious social sanctions. The need for a normative content to social roles follows from the fact that (a) a social actor may have private, publicly inaccessible payoffs that conflict with the public payoffs associated with a role, inducing him to act counter to appropriate role-performance given by the content of the social role (e.g., corruption, favoritism, and aversion to specific tasks); (b) the signal used to determine the public payoffs may be inaccurate and unreliable (e.g., the performance of a teacher or physician); and (c) the public payoffs required to gain compliance by self-regarding actors may be higher than those required when there is at least partial reliance upon the moral commitment of role incumbents (e.g., it may be less costly to employ personally committed rather than purely materially motivated physicians and teachers). In such cases, self-regarding actors who treat social norms purely instrumentally will behave in a socially inefficient and morally reprehensible manner.

The normative aspect of social roles is motivating to social actors because to the extent that social roles are considered legitimate, role-occupants normally place an intrinsic positive ethical value on role-performance (Andreghetto *et al.*, 2013). We may call this the *normative bias* associated with role-occupancy (Bicchieri, 2006; Gintis, 2009a). Second, human ethical predispositions include character virtues, such as honesty, trustworthiness, promise-keeping, and obedience, that may increase the value of conforming to the duties associated with role-incumbency (Aristotle, 350BC, Ullman-Margalit, 1977). Third, humans are predisposed to care about the esteem of others even when there can be no future reputational repercussions (Smith 1759; Masclet *et al.*, 2003), and take pleasure in punishing others who have violated social norms even when they can gain no personal advantage thereby (Güth *et al.*, 1982; Gintis, 2000; Fehr and Fischbacher, 2004). These normative traits by no means contradict rationality, because individuals trade off these values against material reward,

and against each other, just as described in the economic theory of the rational actor (Andreoni and Miller, 2002; Gneezy and Rustichini, 2000).

2.5 Socialization and the Internalization of Norms

Society is held together by moral values that are transmitted from generation to generation by the process of *socialization*. These values are instantiated through the *internalization of norms* (Parsons, 1967; Grusec and Kuczynski, 1997; Nisbett and Cohen, 1996; Rozin *et al.*, 1999), a process in which the initiated instill values into the uninitiated, usually the younger generation, through an extended series of personal interactions, relying on a complex interplay of affect and authority. Through the internalization of norms, initiates are supplied with moral values that induce them to conform voluntarily and even at times enthusiastically to the duties and obligations of the role-positions they are expected to occupy. In addition, the adherence to social norms is socially reinforced by the approval and rewards offered by prosocial individuals, and the decentralized punishment of norm violation by concerned individuals (Gintis, 2000; Fehr and Fischbacher, 2004). Moreover, humans acquire social norms simply through the action of homophily, imitating behavior and acquiring the value of social peers (Kandel, 1978; McPherson *et al.*, 2001; Durrett and Levin, 2005).

The internalization of norms of course presupposes a genetic predisposition to moral cognition that can be explained only by gene-culture coevolution (Boyd and Richerson, 1985, 2004; Gintis, 2003a, 2011; Haidt, 2001).

It is tempting to treat some norms as inviolable constraints that lead the individual to sacrifice personal welfare on behalf of morality. But virtually all norms are violated by individuals under some conditions, indicating that there are tradeoffs that could not exist were norms merely constraints on action. In fact, internalized norms are accepted not as instruments towards achieving other ends, but rather as ends in themselves—*arguments in the preference function that the individual maximizes*. 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. Such fundamental human emotions as shame, guilt, pride, and empathy are deployed by the well-socialized individual to reinforce these prosocial values when tempted by the immediate pleasures of such deadly sins as anger, avarice, gluttony, and lust.

The human openness to socialization is perhaps the most powerful form of epigenetic transmission found in nature. This preference flexibility accounts in considerable part for the stunning success of the species *Homo sapiens*, because when individuals internalize a norm, the frequency of the desired behavior will be higher than if people follow the norm only instrumentally—i.e., when they perceive it to be in their best interest to do so on self-regarding grounds. The

increased incidence of prosocial behaviors are precisely what permits humans to cooperate effectively in groups (Gintis *et al.*, 2005).

There are, of course, limits to socialization (Wrong, 1961; Gintis, 1975; Tooby and Cosmides, 1992; Pinker, 2002), and it is imperative to understand the dynamics of emergence and abandonment of particular values, which in fact depend on their contribution to fitness and well-being, as economic and biological theory would suggest (Gintis, 2003ab). Moreover, there are often swift society-wide value changes that cannot be accounted for by socialization theory (Wrong, 1961; Gintis, 1975). For instance, movements for gender and racial equality have been highly successful in many countries, yet initially opposed all major socialization institutions, including schools, churches, the media, and the legal system.

2.6 Social Norms as Correlated Equilibria

Many social norms involve social actors behaving appropriately in roles in which social interaction is absent. For instance, a mother's maintaining proper maternal health during pregnancy, the humane treatment of animals, or littering in public spaces are of this nature. In general, however, social norms regulate the strategic interaction of social actors. Several influential theorists have modeled social norms in such cases as Nash equilibria of games played by rational agents, including David Lewis (Lewis, 1969), Taylor (1976, 1982, 1987), Sugden (1986, 1989), Bicchieri (1993, 2006), and Binmore (1993, 1998, 2005).

The insight underlying this approach is that if agents play a game with several Nash equilibria, a social norm can serve to choose the most socially desirable among these equilibria. While this insight applies to several important social situations, it is insufficiently broad for a core analytical model of social norms. We suggest that the broader concept of correlated equilibrium (Aumann, 1974, 1987a) better captures the notion of a social role. A correlated equilibrium consists of a *correlating device*, which we sometimes call the *choreographer*, that sends a signal indicating a suggested action to each social actor, such that the actor, for both material and moral reasons, does best by obeying the choreographer's suggestion, provided the other relevant social actors do so as well. While the notion of a choreographer accurately captures the effect of a correlating device's fostering of social cooperation, we generally reject the connotation of the choreographer as a dictator who rules by force. Social norms generally will not be followed when they are not considered legitimate, whatever the social sanctions entailed by the discovery of violations. Moreover, social norms generally are instantiated and changed through collective action, so that the choreographer itself is the product of a social will (Gintis, 1975; Winter *et al.*, 2012).

The model of social norms as correlated equilibria has an attractive property lacking in the notion of social norms as Nash equilibria: the conditions under which rational agents play Nash equilibria are generally complex and implausible, whereas rational agents in a very natural sense play correlated equilibria, provided they have common knowledge of the behavior of the correlating device. For instance, Thomas Schelling's notion of a *focal point* equilibrium can be interpreted as a correlated equilibrium. Consider the situation of two friends who agree to have lunch in the city but fail to state exactly where and at what time to meet. There are an infinite number of Nash equilibria for this situation, one for each time and place in the city. Moreover, the chances the two friends will agree on which Nash equilibrium to implement are extremely small.

However, suppose the choreographer is the social convention "the default time for lunch is noon and the default place is the most frequented place in the city." Assuming both friends know the social norm and have the same prior concerning the most frequented spot in this city, there will be a unique correlated equilibrium, and this equilibrium will be socially efficient. We develop the general argument for social norms as correlated equilibria in Section B1.

3 Gene-Culture Coevolution

Gene-culture coevolution is the application of *sociobiology* (Wilson, 1975; Brown, 1991; Cosmides *et al.*, 1992), the general theory of the social organization of biological species, to humans — a species that transmits culture in a manner that leads to quantitative growth across generations. This is a special case of *niche construction*, which applies to species that transform their natural environment so as to facilitate social interaction and collective behavior (Odling-Smee *et al.*, 2003). In the case of gene-culture coevolution, the environmental change is that of the social structure within which individuals live out their lives. The natural environment may be involved as well, as when settled agriculture alters the density of disease-carrying insects, and hence selects for individuals who are relatively immune to these diseases (Laland *et al.*, 2000).

Because of their common informational and evolutionary character, there are strong parallels between models of genetic and cultural evolution (Mesoudi *et al.*, 2006). Like biological transmission, culture is transmitted from parents to offspring, and like cultural transmission, which is transmitted horizontally to unrelated individuals, so in microbes and many plant species, genes are regularly transferred across lineage boundaries (Jablonka and Lamb, 1995; Abbott *et al.*, 2003; Rivera and Lake, 2004). Moreover, anthropologists reconstruct the history of social groups by analyzing homologous and analogous cultural traits, much as biologists reconstruct the evolution of species by the analysis

of shared characters and homologous DNA (Mace and Pagel, 1994). Indeed, the same computer programs developed by biological systematists are used by cultural anthropologists (Holden, 2002; Holden and Mace, 2003). In addition, archeologists who study cultural evolution have a similar modus operandi as paleobiologists who study genetic evolution (Mesoudi *et al.*, 2006). Both attempt to reconstruct lineages of artifacts and their carriers. Like paleobiology, archaeology assumes that when analogy can be ruled out, similarity implies causal connection by inheritance (O'Brien and Lyman, 2000). Like biogeography, the study of the spatial distribution of organisms (Brown and Lomolino, 1998), behavioral ecology studies the interaction of ecological, historical and geographical factors that determine distribution of cultural forms across space and time (Winterhalder and Smith, 1992).

Gene-culture coevolution is an empirical fact, not a theory. However, it is a complex and variegated process that takes many forms. Modeling gene-culture coevolution began with Feldman and Cavalli-Sforza (1976), followed by their book Cavalli-Sforza and Feldman (1981), in which they modeled vertical (parent to child), oblique (non-parental elders to youngers) and horizontal (peer to peer) cultural transmission. Lumsden and Wilson (1981) presented an alternative model, as did Boyd and Richerson (1985). For enlightening contemporary reviews of these pioneers, see Lewontin (1981) and Maynard Smith and Warren (1982). As a concrete example of gene-culture coevolution, we present an overview of the evolution of the physiology of speech in Section 3.2.

3.1 Cultural and Institutional Evolution

An organism's genome encodes information that is used both to construct a new organism and to endow it with instructions for transforming sensory inputs into decision outputs. Because learning is costly and time-consuming, efficient information transmission will ensure that the genome encodes those aspects of the organism's environment that are constant, or that change only very slowly through time and space, as compared with an individual lifetime. By contrast, environmental conditions that vary rapidly can be dealt with by providing the organism with phenotypic plasticity in the form of the capacity to learn. For instance, suppose the environment provides an organism with the most nutrients where ambient temperature is highest. An organism may learn this by trial and error over many periods, or it can be hard-wired to seek the highest ambient temperature when feeding. By contrast, suppose the optimal feeding temperature varies over an individual's lifetime. Then there is no benefit to encoding this information in the individual's genome, but a flexible learning mechanism will enhance the individual's fitness.

There is an intermediate case, however, that is efficiently handled neither by genetic encoding nor learning. When environmental conditions are positively but imperfectly correlated across generations, each generation acquires

valuable information through learning that it cannot transmit genetically to the succeeding generation, because such information is not encoded in the germ line. In the context of such environments, there is a fitness benefit to the epigenetic transmission of information concerning the current state of the environment; i.e. transmission through non-genetic channels. Several epigenetic transmission mechanisms have been identified (Jablonka and Lamb, 1995), but cultural transmission in humans and to a lesser extent in other animals (Bonner, 1984; Boyd and Richerson, 2004) is a distinct and extremely flexible form. Cultural transmission takes the form of vertical (parents to children), horizontal (peer to peer) and oblique (elder to younger), as in Cavalli-Sforza and Feldman (1981), prestige (higher influencing lower status), as in Henrich and Gil-White (2001), popularity-related as in Newman *et al.* (2006), and even random population-dynamic transmission, as in Shennan (1997) and Skibo and Bentley (2003). The parallel between cultural and biological evolution goes back to Huxley (1955), Popper (1979), and James (1880)—see Mesoudi *et al.* (2006) for details. The idea of treating culture as a form of epigenetic transmission was pioneered by Dawkins (1976), who coined the term *meme* in *The Selfish Gene* to represent an integral unit of information that could be transmitted phenotypically. There quickly followed several major contributions to a biological approach to culture, all based on the notion that culture, like genes, could evolve through replication (intergenerational transmission), mutation and selection.[1] Cultural elements 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 (Cavalli-Sforza and Feldman, 1982; Parsons, 1964). 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, including empathy, shame, guilt and revenge-seeking (Ihara, 2011; Zajonc, 1980, 1984).

Perhaps the most common criticism of the analogy between genetic and cultural evolution is that the gene is a well-defined, discrete, independently reproducing and mutating entity, whereas the boundaries of the unit of culture are ill-defined and overlapping. In fact, however, this view of the gene is outdated. We now know that overlapping, nested and movable genes have some of the fluidity of cultural units, whereas quite often the boundaries of a cultural unit (a belief, icon, word, technique, stylistic convention) are quite delimited and specific. Similarly, alternative splicing, nuclear and messenger RNA editing, cellular protein modification and genomic imprinting, which are quite common, undermine the standard view of the insular gene producing

[1]Dawkins recognized that the extended phenotypic expression of a genotype should affect the fitness of that genotype, but opposes considering that this expression can also have the nicheconstructive effect of modifying the selective environment for other genotypes (see Dawkins, 2004).

a single protein, and support the notion of genes having variable boundaries and strongly context-dependent effects. Moreover, natural selection requires heritable variation and selection, but does not require discretely transmitted units.

Dawkins (1982) added a second fundamental mechanism of epigenetic information transmission in The Extended Phenotype, noting that organisms can directly transmit environmental artifacts to the next generation, in the form of such constructs as beaver dams, bee hives and even social structures (e.g. mating and hunting practices). The phenomenon of a species creating an important aspect of its environment and stably transmitting this environment across generations, known as niche construction, is a widespread form of epigenetic transmission (Odling-Smee *et al.*, 2003). Niche construction includes gene-environment coevolution, because a genetically induced environmental regularity becomes the basis for genetic selection, and gene mutations that give rise to novel niche elements will survive if they are fitness-enhancing for their constructors.

An excellent example of gene-environment coevolution is the honeybee, in which the origin of its eusociality probably lay in a high degree of relatedness, but which persists in modern species despite the fact that relatedness in the hive is generally quite low, due to multiple queen matings, multiple queens, queen deaths and the like (Gadagkar, 1991; Seeley, 1997; Wilson and Hölldobler, 2005). The social structure of the hive, a classic example of niche construction, is transmitted epigenetically across generations, and the honeybee genome is an adaptation to the social structure laid down in the distant past.

Gene-culture coevolution in humans is a special case of gene-environment coevolution in which the environment is culturally constituted and transmitted (Feldman and Zhivotovsky, 1992). The key to the success of our species in the framework of the hunter-gatherer social structure in which we evolved is the capacity of unrelated, or only loosely related, individuals to cooperate in relatively large egalitarian groups in hunting and territorial acquisition and defense (Boyd and Richerson, 2004; Boehm, 1999). While some contemporary biological and economic theorists have attempted to show that such cooperation can be supported by self-regarding rational agents (Alexander, 1987; Fudenberg *et al.*, 1994; Trivers, 1971), the conditions under which their models work are implausible even for small groups (Boyd and Richerson, 1988; Gintis, 2009a). Rather, the social environment of early humans was conducive to the development of prosocial traits, such as empathy, shame, pride, embarrassment and reciprocity, without which social cooperation would be impossible (Sterelny, 2011).

Neuroscientific studies exhibit clearly the genetic basis for moral behavior. Brain regions involved in moral judgments and behavior include the prefrontal cortex, the orbitalfrontal cortex and the superior temporal sulcus (Moll *et al.*, 2005). These brain structures are virtually unique to or most highly developed

in humans and are doubtless evolutionary adaptations (Schulkin 2000). The evolution of the human prefrontal cortex is closely tied to the emergence of human morality (Allman *et al.*, 2002). Patients with focal damage to one or more of these areas exhibit a variety of antisocial behaviors, including the absence of embarrassment, pride and regret (Beer *et al.*, 2003; Camille, 2004), and sociopathic behavior (Miller *et al.*, 1997). There is a probable genetic predisposition underlying sociopathy, and sociopaths comprise 3.4% of the male population, but they account for between 33 and 80 per cent of the population of chronic criminal offenders in the United States (Mednick *et al.*, 1977). It is clear from this body of empirical information that culture is directly encoded into the human brain with symbolic representations in the form of cultural artifacts. This, of course, is the central claim of gene-culture coevolutionary theory.

3.2 The Physiology of Communication

The evolution of the physiology of speech and facial communication is an excellent example of gene-culture coevolution. The increased social importance of communication in human society rewarded genetic changes that facilitate speech. Regions in the motor cortex expanded in early humans to facilitate speech production. Concurrently, nerves and muscles to the mouth, larynx and tongue became more numerous to handle the complexities of speech (Jurmain *et al.*, 1997). Parts of the cerebral cortex, Broca's and Wernicke's areas, which do not exist or are relatively small in other primates, are large in humans and permit grammatical speech and comprehension (Belin *et al.*, 2000; Binder *et al.*, 1997).

Adult modern humans have a larynx low in the throat, a position that allows the throat to serve as a resonating chamber capable of a great number of sounds (Relethford, 2007). The first hominids that have skeletal structures supporting this laryngeal placement are the *Homo heidelbergensis*, who lived from 800,000 to 100,000 years ago. In addition, the production of consonants requires a short oral cavity, in whereas our nearest primate relatives have much too long an oral cavity for this purpose. The position of the hyoid bone, which is a point of attachment for a tongue muscle, developed in Homo sapiens in a manner permitting highly precise and flexible tongue movements.

Another indication that the tongue has evolved hominids to facilitate speech is the size of the hypoglossal canal, an aperture that permits the hypoglossal nerve to reach the tongue muscles. This aperture is much larger in Neanderthals and humans than in early hominids and non-human primates (Dunbar, 2005). Human facial nerves and musculature have also evolved to facilitate communication. This musculature is present in all vertebrates, but except in mammals it serves feeding and respiratory functions alone (Burrows, 2008). In mammals, this mimetic musculature attaches to the skin of the face,

thus permitting the facial communication of such archetypal emotions as fear, surprise, disgust and anger. In most mammals, however, a few wide sheetlike muscles are involved, rendering fine information differentiation impossible, whereas in primates, this musculature divides into many independent muscles with distinct points of attachment to the epidermis, thus permitting higher bandwidth facial communication. Humans have the most highly developed facial musculature by far of any primate species, with a degree of involvement of lips and eyes that is not present in any other species.

In short, humans have evolved a highly specialized and very costly complex of physiological characteristics that both presuppose and facilitate sophisticated aural and visual communication, whereas communication in other primates, lacking as they are in cumulative culture, goes little beyond simple calling and gesturing capacities. This example is quite a dramatic and concrete illustration of the intimate interaction of genes and culture in the evolution of our species.

4 Networked Minds and Distributed Cognition

There are many plausible ways to model the cognition of social actors as networked across a range of significant others (Coleman, 1988; Rauch, 1996; Bowles and Gintis, 2004; Di Guilmi *et al.*, 2012; Gintis, 2013). The following model is offered simply as an illustration of how this might be accomplished.

Suppose there are social actors $i = 1, \ldots, m$ and there is a network of information flows among the actors. Let $\mathcal{P}_i$ be the set of actors to whom actor i is directly linked. Suppose there are n traits, such as gender, ethnicity, occupation, religion, social position, physical attributes, family relationship, cultural beliefs and demographic characteristics. We assume each social actor has a *social trait vector* $a = (a_1, \ldots, a_n)$ where each a_j takes the value zero and one. We interpret $a_j = 0$ as meaning that the individual does not have trait j , and $a_j = 1$ means the individual has trait j. An actor i with personal traits vector $a^i \in A$ has available a set of *trait filters*, where a trait filter $b_i \in A$ represents the set of traits that i considers relevant in polling others in a particular decision context. We interpret $b_{ij} = 1$ as meaning members of $\mathcal{P}_i$ satisfying the filter have trait j, and $b_{ij} = 0$ as meaning that members of $\mathcal{P}_i$ may or may not have trait j. For some decisions, i will consider only other actors with the same personal characteristics, so $b_i \leq a_i$, in the sense that $b_{ij} \leq a_{ij}$ for all traits j . However, in other cases i may defer to experts or highly experienced network members with personal traits that differ in important ways.

In facing a particular decision, actor i evaluates information from other social actors in his network $\mathcal{P}_i$, using a trait filter b_i that is dependent on the nature of the decision. The *strength* $\rho(b_i)$ of a trait filter b_i is the number of positive entries in b_i. The stronger the trait filter, the closer others must be in

social space for their experience to count in the actor's decision. The strength of a trait filter is a partial order on A in the obvious sense. We write $b_i(\mathcal{P}_i)$ for the set of network links to i that conform to the filter b_i.

Let k_i be the number of actors in $\mathcal{P}_i$, and let $k_i(b_i)$ be the number of actors in $\mathcal{P}_i$ who conform to the filter b_i, which is decreasing in the strength of the filter b_i. Thus $k_i(b_i)/k_i$ is the fraction of social actors in i's network who have the traits b_i. Let $q_i(b_i)$ be the probability that a social actor with traits b_i provides correct information allowing i to choose an action that maximizes i's payoff. Because the use of a stronger filter cannot improve the decision-maker's information unless it also increases the probability of receiving correct information, we may safely assume that for a given decision problems, decision-maker i considers only filters that belong to a totally ordered sequence of increasingly strong filters $b^{i1}, b^{i2}, \ldots$ such that $q_i(b^{ij})$ is increasing in j. Let q_i^* be the probability that i chooses correctly without information.

We suppose individual i queries a particular member of his network with traits b_i , who tells him the correct action if he knows it, which occurs with probability $q_i(b_i)$. Otherwise the queried actor gives no information. We can then express the probability that the individual receives the correct information as

$$p_i(b_i) = \alpha(b_i)q_i(b_i) + (1 - \alpha(b_i))q_i^*, \tag{1}$$

where $\alpha(b_i) = k_i(b_i)/k_i$. The decision-maker can then choose the filter b_i to maximize the probability of obtaining useful information (Bowles and Gintis, 2004).

5 Socio-Psychological Theory of Norms

A key tenet of socialization theory is that a society's values are passed from generation to generation through the internalization of norms (Durkheim, 1902; Mead, 1934; Parsons, 1967; Grusec and Kuczynski, 1997). In the language of optimization theory, internalized norms are accepted not as constraints upon achieving other ends, but rather as *arguments in the objective function that the individual maximizes.*

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, is based on a distinctive psychological predisposition.

For analytical specificity, we study the dynamics of a single altruistic norm that has a payoff disadvantage for those who adopt it, but is transmitted vertically by parents and obliquely through socialization institutions. We allow altruism to be either beneficial or harmful to the group, and we admit four types of cultural change. This model is fully developed in Gintis (2003ab).

- Individuals mate and have offspring. Families who use lower payoff strategies have fewer offspring (biologically adaptive dynamics).
- Families pass on their cultural traits, self-interested or altruistic, to their off-spring (vertical transmission) through internalization.
- A fraction of self-interested offspring are induced to adopt altruistic norms by socialization institutions (oblique transmission).
- Some of the resulting population change their cultural values to conform to the behavior of other individuals who have higher payoffs (replicator dynamics).

Our model yields two general conclusions.

- In the absence of oblique transmission of the altruistic norm, altruism is driven out by self-interested behavior. When oblique transmission of altruism is present, a positive frequency of altruism can persist in cultural equilibrium.
- A high level of cooperation can be sustained in cultural equilibrium by the presence of a minority of agents who adopt the altruistic norm of what we call *strong reciprocity*: cooperating unconditionally and punishing defectors at a personal cost, the remaining agent being self-interested.

The first assertion states what might be called the *Fundamental Theorem of Sociology: extra-familial socialization institutions are necessary to support altruistic forms of prosociality.* The second assertion expresses the insight that cooperation is robustly stable when antisocial behavior is punished by the voluntary, and largely decentralized, initiative of group members (Gintis, 2003ab; Helbing *et al.*, 2010).

Because social norms generally have a strong moral component, constructing dynamic models of the evolution of social norms is an inherently complex and ill-understood process. For instance, social norms concerning gender roles or interethnic relationships can persist for many generations and then change extremely rapidly. Such changes are virtually unpredictable given the current state of social theory. Conventions, by contrast, may be more or less desirable on social efficiency grounds but because they lack a moral component, they are more easily modeled and understood.

A *convention* is a correlated equilibrium of a coordination game. A *coordination game* is defined as follows. Suppose there is some social activity that requires the cooperation of one or more types of social actor. For instance, the activity may be building a wall. The types of social actor may be "bricklayer" and "assistant." Cooperation is successful when the bricklayer asks for a piece of building material and the assistant provides the proper material. The social

convention may be that the bricklayer shows one finger when he wants a brick, two fingers when he wants some mortar, and three fingers when he wants a bucket of water. A second convention may be to show one finger for a bucket of water, two fingers for some water, and to say "ladrillo" for a brick. It does not much matter what the particular sign is for each of the three possibilities, just so both the bricklayer and his assistant agree, and the assistant has some incentive to obey the requests of the bricklayer.

There are several plausible models of the evolution and transformation of conventions (Kandori *et al.*, 1993, Young, 1993, 1998) based on the notion of a Markov process. We provide a simple but representative example of this approach to modeling the evolution of conventions in Section C1.

5.1 A Model of Cultural Evolution

Consider a group in which members can either adopt, or fail to adopt, a certain cultural norm A. We shall call those who adopt norm A *altruists* because we assume that following the norm improves the mean payoffs of group members, although at a cost to the altruist. We call those who do not adopt norm A *self-interested* types, or "B-types." Altruism is costly, in that self-interested types have fitness 1, as compared with altruists, who have fitness $1-s$, where $0<s<1$.[2] We assume in each period that agents pair off randomly, mate, and have offspring in proportion to their fitness, after which they die (we call this a *biologically adaptive dynamic*). Families pass on their cultural norms to their offspring, so offspring of AA parents are altruists, offspring of BB parents are self-interested, and half of the offspring of AB-families are altruists, the other half self-interested (we call this *vertical transmission*). We also assume that the self-interested offspring of AB- and BB-families are susceptible to influence by community institutions promoting altruistic norms, a fraction of such offspring becoming altruists (we call this *oblique transmission*).

For the first stage, suppose there are n males and n females at the beginning of the period. If the fraction of altruists is α, there will be $n\alpha^2$ AA-families, who will have $n\alpha^2(1-s)^2\beta$ offspring, all of whom are altruists, where we choose β so that population size is constant. There will also be $2n\alpha(1-\alpha)$ AB-families, who will have $2n\alpha(1-\alpha)(1-s)\beta$ offspring, half of whom are altruists. Finally there will be $n(1-\alpha)^2$ BB-families who will have $n(1-\alpha)^2\beta$ offspring. Adding up the number of offspring, we see that we must have $\beta=1/(1-s\alpha)^2$. Thus the frequencies of AA, AB, and BB offspring are given by

$$f_{AA}=\frac{\alpha^2(1-s)^2}{(1-\alpha s)^2},\quad f_{AB}=\frac{2\alpha(1-\alpha)(1-s)}{(1-\alpha s)^2},\quad f_{BB}=\frac{(1-\alpha)^2}{(1-\alpha s)^2}. \qquad (2)$$

[2]Note that altruists may reach a higher fitness than non-altruists when they predominantly interact with other altruists (Bowles and Gintis, 2011; Grund *et al.*, 2013). Therefore, spatial assortativity and social segregation are other mechanisms that can stabilize or promote the emergence of altruism, even if everyone is self-interested in the beginning.

Second, a fraction $\alpha\gamma$ of offspring of AB- and BB-families who are self-interested switch to being altruists under the influence of the oblique transmission of cultural norm A, where y is a measure of the strength of the oblique transmission process. Note that we have made the conservative assumption that oblique transmission is proportional to the level of altruism. It is easy to check that the change in the fraction of altruists in the next generation is given by

$$\frac{d\alpha}{dt} = f(\alpha) = \frac{\alpha(1-\alpha)(\gamma - s)}{1 - s\alpha}. \tag{3}$$

Third, each group member i observes the fitness and the type of a randomly chosen other member j , and changes to j's type if j's fitness is higher. However, information concerning the difference in fitnesses of the two strategies is imperfect, and agents' objective functions do not perfectly track fitness, so it is reasonable to assume that the larger the difference in the payoffs, the more likely the agent is to perceive it, and change. Specifically, we assume the probability p that an agent using A will shift to B is proportional to the fitness difference of the two types, so $p = \sigma s$ for some proportionality constant $\sigma > 0$.

The expected fraction α' of the population using A after the above shifts is then given by

$$\alpha' = \alpha - \sigma\alpha(1-\alpha)s,$$

which, expressed in differential equation form, is

$$\alpha' = -\sigma\alpha(1-\alpha)s. \tag{4}$$

This is a special case of the *replicator dynamic* in cultural evolution. We now combine these two sources of change in the fraction of altruists, giving

$$\frac{d\alpha}{dt} = h(\alpha) = f(\alpha) - \sigma\alpha(1-\alpha)s \tag{5}$$

where σ now represents the relative strength of the replicator dynamic, which is biased against the altruistic norm, in comparison with the cultural transmission mechanisms, which may favor this norm. In reduced form, we now have

$$\frac{d\alpha}{dt} = \frac{\alpha(1-\alpha)}{1 - s\alpha}(\gamma - s - s\sigma(1 - s\alpha)). \tag{6}$$

We call the situation $\frac{d\alpha}{dt} = 0, \alpha \in [0,1]$ a *cultural equilibrium* of the dynamical system. We then have

Theorem 1. *Let us assume $\gamma \geq 0$ is given and fixed throughout and define $s_{min} = \frac{\gamma}{1+\sigma}$ and*

$$s_{max} = \frac{1}{2\sigma}\left\{1 + \sigma - \sqrt{(1+\sigma)^2 - 4\gamma\sigma}\right\}.$$

1. *If $s < s_{min}$ then $\alpha = 1$ is a globally stable altruistic equilibrium.*
2. *If $s_{min} < s < s_{max}$ then both $\alpha = 0$ and $\alpha = 1$ are locally stable equilibria of the system and there is third unstable equilibrium $\alpha^* \in (0, 1)$ separating the basins of attraction of the two stable equilibria: both self-interested and altruistic equilibria are stable.*
3. *If $s > s_max$, then $\alpha = 0$ is a stable self-interested equilibrium of the system.*

Proof. There are three zeros of (6), of which two are $\alpha = 0$ and $\alpha = 1$. The third is $\alpha^* = (s(1+\sigma) - \gamma)/s^2\sigma$. If $s < s_{\min}$, then $h'(0) > 0, \alpha^* < 0$, and $h'(1) > 0$, so the unique stable equilibrium is $\alpha = 1$, proving (a). If $s_{\min} < s < s_{\max}$, then $\alpha^* \in (0, 1), h'(0), h'(1) < 0$, so both $\alpha = 0$ and $\alpha = 1$ are stable. α^* must then be unstable, proving (b). Finally, if $s > s_{\max}, \alpha^* > 1, h'(0) < 0$, and $h'(1) > 0$, so $\alpha = 0$ is the only stable cultural equilibrium, proving (c). □

Theorem 1 might logically be called the *Fundamental Theorem of Sociology.*

Corollary 1. *With the above assumptions, altruistic norms persist in a cultural equilibrium only if there is a strictly positive rate of cultural transmission of altruism* via *social institutions.*

Proof. If $\gamma = 0$, then $s_{\max} = 0$, so $s > s_{\max}$. Then, according to Theorem 1, $\alpha = 0$ is the only stable cultural equilibrium. □

Theorem 1 shows that the higher the personal cost of altruistic behavior, the more stringent the conditions under which altruism will emerge. This result illustrates the power of a theory that models the tension between prosocial socialization institutions and the psychological mechanism of norm internalization on the one hand, and the replicator dynamic that induces agents to shift to higher payoff behaviors, despite their effect on general social well-being, on the other hand. This tension is also revealed in the following:

Corollary 2. *We say the replicator dynamic is* weak *if σ satisfies*

$$\sigma < \frac{\gamma - s}{s},$$

is moderate *if*

$$\frac{\gamma - s}{s} < \sigma < \frac{\gamma - s}{s(1-s)},$$

and is strong *if*

$$\frac{\gamma - s}{s(1-s)} < \sigma,$$

If the replicator dynamic is weak, then the altruistic cultural equilibrium is globally stable. If the replicator dynamic is moderate, then both the self-interested

and the altruistic cultural equilibria are locally stable, and the basin of attraction of the altruistic equilibrium shrinks as u increases. Finally, if the replicator dynamic is strong, then the self-interested cultural equilibrium is globally stable.

5.2 The Evolution of Norm Internalization

Why do we have the generalized capacity to internalize norms? From a biological standpoint, internalization may be an elaboration of imprinting and imitation mechanisms found in several species of birds and mammals, but its highly developed form in humans indicates that it probably had great adaptive value during our evolutionary emergence as a species. Moreover, from an economic standpoint, the everyday observation that people who exhibit a strongly internalized moral codes lead happier and more fulfilled lives than those who subject all actions to a narrow calculation of personal costs and benefits of norm compliance, suggests it might not be 'rational' to be self-interested.

Gintis (2003a) shows that if internalization of some norms is personally fitness enhancing (e.g., preparing for the future, having good personal hygiene, positive work habits, and/or control of emotions), then genes promoting the capacity to internalize can evolve. Given this genetic capacity, as we have seen above, altruistic norms will be internalized as well, provided their fitness costs are not excessive. In effect, altruism 'hitchhikes' on the personal fitness-enhancing capacity of norm internalization.[3] Altruistic behavior, then, is an exaption, in the sense of Gould and Vrba (1981).

Why, however, should the internalization of any norms be individually fitness-enhancing? The following is a possible explanation, based on the observation that internalization alters the agent's goals, whereas instrumental and conventional cultural forms merely aid the individual in attaining *pre-given* goals. In humans, as much as in other species, these goals are related to, but not reducible to, biological fitness.

Biological fitness is a theoretical abstraction unknown to virtually every real-life organism. Organisms therefore do not, in any circumstance, literally maximize fitness. Rather, organisms have a relatively simple state-dependent objective function that is itself subject to selection according to its ability to promote individual fitness (Alcock, 1993). In a slowly-changing environment, this objective function will track fitness closely. In a rapidly changing environment, however, natural selection will be too slow, and the objective function will not track fitness well.

The development of cultural transmission, in the form of instrumental techniques and conventions, and the ensuing increase in social complexity of hominid society, doubtless produced such a rapidly changing environment, thus

[3]This mechanism was asserted by Simon (1990), who instead of 'internalization of norms', used the term 'docility,' in the sense of 'capable of being easily led or influenced.'

conferring high fitness value on the development of a *non-genetic mechanism for altering the agent's objective function.* Internalization is adaptive because it allows the human objective function to shift in directions conducive to higher personal fitness. The internalization of norms is thus adaptive because it facilitates the transformation of drives, needs, desires, and pleasures (arguments in the human objective function) into forms that are more closely aligned with fitness maximization. Internalization is limited to our species, mainly because no other species places such great emphasis on cultural transmission.

We humans thus have a 'primordial' objective function that does not well serve our fitness interests, and which is more or less successfully 'overridden' by our internalized norms. This primordial objective function knows nothing of 'thinking ahead,' but rather satisfies immediate desires. Lying, cheating, killing, stealing, and satisfying short-term bodily needs (wrath, lust, greed, gluttony, sloth) are all actions that produce immediate pleasure and drive-reduction, at the expense of our overall well-being in the long run.

This evolutionary argument is meant to apply to the long period in the Pleistocene during which the human character was formed. Social change since the agricultural revolution of about 10,000 years ago has been far too swift to permit even the internalization of norms to produce a close fit between utility and fitness. Indeed, with the advent of modern societies, the internalization of norms has been systematically diverted from fitness (expected number of offspring) to welfare (net degree of contentment) maximization. This, of course, is precisely what we would expect when humans obtain control over the content of ethical norms. Indeed, this misfit between welfare and fitness is doubtless a necessary precondition for civilization and a high level of *per capita* income. This is true because, were we fitness maximizers, every technical advance would have been accompanied by an equivalent increase in the rate of population growth, thus nullifying its contribution to human welfare, as predicted long ago by Thomas Malthus. The demographic transition, which has led to dramatically reduced human birth rates throughout most of the world, is a testimonial to the gap between welfare and fitness. Perhaps the most important form of prosocial cultural transmission in the world today is the norm of having few, but highly successful offspring.

6 General Social Equilibrium

This section illustrates the power of an analytical formulation of general social equilibrium in a particularly simple case, that of explaining the structure of social classes when all social differences reduce to wealth differences. We avoid having to deal with the structure of social norms by abstracting from situations in which social coordination is problematic and social dilemmas are present.

Consider a society whose members engage in household and market production. There are two social institutions, families and firms, and two types of social roles, family member and worker. Each individual belongs to a family, and may in addition sell labor to (i.e., work for) another family or a firm. In both family and market sectors, labor and capital goods are combined to produce goods and services. Labor and capital goods are owned by individuals and are the only form of social wealth. To produce goods and services, firms must purchase labor, i.e., hire workers, and rent capital goods, i.e. borrow them from their owners. Firms then produce marketable goods and services which they sell to families, whose income derives from the labor services and capital goods they supply to firms. Each family has a number of members, who pool their wealth and apply the labor and capital goods that they do not sell to firms to produce goods and services that are consumed in their household. Families buy the market goods of firms, some of which they consume, and some of which they add to their stock of capital goods. The economy is in equilibrium when the vector of prices p for all market is set so that the supply equals the demand in each sector, while firms maximize profits and families maximize their utility from consumption and wealth creation.

We will model a highly simplified version of this society in which there is a single household good (f) and a single market good (m). These restrictions are easily lifted at the expense of more complex notation, with little insight being thereby gained. We assume there are only two families (x and y). We also assume there is a single type of labor (l). Suppose l_x and l_y are the amounts of labor owned by families x and y, (l_f^x, k_f^x) are the amounts of labor and capital goods inputs used by family $x, (l_m^y, k_m^y)$ are the amounts of labor and capital goods inputs used by family $y, (l_m^x, k_m^x)$ and (l_m^y, k_m^y) are the labor and capital goods supplied to firms by families x and y. Then we have the equations

$$l_f^x + l_m^x = l_x \tag{7}$$

$$l_f^y + l_m^y = l_y \tag{8}$$

$$k_f^x + k_m^x = k_x \tag{9}$$

$$k_f^y + k_m^y = k_y \tag{10}$$

The equations say that the total amount of labor and capital goods demanded by firms to use in production, plus the total amount of labor and capital goods used in family production, equals the total amount of these factors supplied by families. This assumes that families use all the factors they own either by supplying them to firms or applying them to family production.

Now suppose the wage rate is w and the interest rate (which is the price for renting one unit of the capital good for one production period) is r. Also, suppose the price of the market good is p, and family x consumes x_f of family

goods and x_m of market goods, while family y consumes y_f of family goods and y_m of market goods. Finally, we assume family x owns a share α of the net profits of firms, while family y owns a share $1-\alpha$.

Then if m_x and m_y are the incomes of families x and y from supplying labor and capital goods to firms, we have the following two equations.

$$m_x = \alpha\pi + wl_m^x + rk_m^x \tag{11}$$

$$m_y = (1-\alpha)\pi + wl_m^y + rk_m^y, \tag{12}$$

where π is the net profits of firms.

The next equations are production functions for the family goods and market goods firms. They say that each good is produced by using capital goods and labor.

$$f^x(l_f^x, k_f^x) = x_f \tag{13}$$

$$f^y(l_f^y, k_f^y) = y_f \tag{14}$$

$$g(l_m^x + l_m^y, k_m^x + k_m^y) = x_m + y_m, \tag{15}$$

where x_m and y_m are the market goods purchased by families x and y. We assume firms maximize profits, given by

$$\pi = p(x_m + y_m) - (w(l_m^x + l_m^y) + k(k_m^x + k_m^y)). \tag{16}$$

Profit maximization gives two first-order conditions

$$g_l = \frac{w}{p} \tag{17}$$

$$g_k = \frac{r}{p} \tag{18}$$

where subscripts represent partial derivatives.

We assume families have utility function $u^x(x_f, x_m)$ and $u^y(y_f, y_m)$, which they maximize subject to their income constraints (11) and (12). Maximizing utility given these income constraints gives four additional equations

$$\frac{u_f^x f_l^x}{w} = \frac{u_m^x f_m^x}{r} = \frac{u_m^x}{p} \tag{19}$$

$$\frac{u_f^y f_l^y}{w} = \frac{u_m^y f_m^y}{r} = \frac{u_m^y}{p}. \tag{20}$$

Finally, we can normalize the nominal price level p to unity, and we assume that competition among firms reduces excess profits to zero:

$$p = 1, \tag{21}$$

$$\pi = 0. \tag{22}$$

In this system, variables α, l_x, l_y, k_x , and k_y are parameters representing the structure of ownership in the economy. There remain eighteen variables to

be determined: $l_f^x, l_f^y, l_m^x, l_m^y, k_f^x, k_f^y, k_m^x, k_m^y, x_f, y_f, x_m, y_m, \pi, m_x, m_y, p, w$, and r. There are also eighteen equations, expressed in (7)–(22). The equality in the number of equations and unknowns generically determines a unique equilibrium, but there is no general guarantee that prices and quantities will be nonnegative in this solution. However, the appropriate assumptions concerning the shape of the production function and utility functions will guarantee the existence of a social equilibrium, along the lines of Debreu (1952) and Arrow and Debreu (1954). The conditions that make this possible, roughly speaking, are that consumers have concave preferences (declining marginal utility) and firms have convex production functions (declining marginal productivity).

6.1 Class Structure in General Social Equilibrium

An elaboration on the general social equilibrium model of the previous section illustrates how wealth inequality can translate into a stratified distribution of social classes. This model is a variant of Eswaran and Kotwal (1986) and Bowles (2004, Ch. 10), who apply a method initiated by Roemer (1982). Suppose all families face the household production function

$$q = f(k, l) \tag{23}$$

where k is capital and l is labor. We assume $f(k, l)$ is increasing and concave in its arguments; i.e., there is decreasing marginal productivity of both labor and capital in household production. However, there is a startup capital cost $\kappa > 0$ for household production. A family can apply its own labor l_f, it can hire labor l_h, and it can sell labor l_w to other households and firms in the market sector. If the household hires labor l_h, it must supervise this labor, incurring a supervisory cost in personal labor time $s(l_h)$. We assume $s(l_h)$ is increasing and convex in the amount of labor hired, with $s(0) = 0$. With supervision, hired workers are as productive as the household labor, so total effective labor in household production is simply $l = l_h + l_f$.

We assume households are credit constrained, with the maximum amount a household with wealth k_f can borrow is $c(k_f)$, where $c(k_f)$ is increasing in k_f with $c(0) = 0$, meaning that a family with no wealth cannot borrow at all. Let w and r be the wage rate and the rate at which capital can be borrowed or loaned. If a household chooses to produce, the credit rationing constraint requires that

$$c(k_f) \geq w(l_f + l_h) + r(k - k_f) + \kappa, \tag{24}$$

where k is the amount of capital the household uses in production. This inequality assumes that all production costs must be paid at the start of the period.

We assume a simple household payoff $y + u(\rho)$, where y is income and ρ is the amount of leisure consumed, and where $u(\rho)$ is increasing and concave

(decreasing marginal utility of leisure). We also assume $u'(0)$ is sufficiently negative that the household always chooses a positive amount of leisure. Then, an individual who choose to enter into household production has payoff

$$\pi_f = f(k, l_f + l_h) - (1 + r)[w(l_h - l_w) + v(k - k_f) + \kappa] + u(\rho), \tag{25}$$

where the $(1 + r)$ term represents the total amount of the loan that must be paid at the end of the period.

An individual who hires out as a worker rather than engaging in household production will have payoff

$$\pi_w = (1 + r)(wl_w + v\kappa) + u(\rho), \tag{26}$$

assuming wages are paid at the start of the production period.

An individual who undertakes household production, such that (25) holds, must choose k, ρ, l_w, l_h, l_f, and l to maximize (25) subject to the credit constraint (24), the inequality constraints $k, l_h, l_f \geq 0$ and a labor constraint given by

$$l_f = 1 - s(l_h) - l_w - \rho \geq 0, \tag{27}$$

where we have normalized the individual's labor endowment to unity. The Lagrangian for this optimization problem is given by

$$\mathcal{L} = f(k, l_f + l_h) - (1 + r)[wl_h + vk + \kappa) + \pi_w + \tag{28}$$
$$\lambda[c(\kappa) - w(l_f + l_h) + r(k - k_f) + \kappa] + \tag{29}$$
$$\mu[1 - s(l_h) - l_f - \rho]. \tag{30}$$

The first-order conditions for this problem are

$$\mathcal{L} = f_k - (1 + r + \lambda)v = 0 \tag{31}$$
$$\mathcal{L}_{lh} = f_l(1 - s'(l_h)) - (1 + r + \lambda)v - \mu s'(l_h) \leq 0 \tag{32}$$
$$\mathcal{L}_\rho = -f_l + u'(\rho) - \mu = 0 \tag{33}$$
$$\mathcal{L}_{l_f} = -f_l + w(1 + r + \lambda) - \mu \leq 0, \tag{34}$$

where (32) is an equality if any labor is hired ($l_h > 0$) and (34) is an equality if the agent himself works in domestic production ($l_f > 0$). The value of λ determined by these equations is the shadow price of borrowed capital, and is strictly positive if the demand for capital in the household sector is positive, which will be the case when the market wage w is not so high that household production is never superior to working in the market sector. In this case $1 + r + \lambda$ is the real cost of borrowing (note that the capital itself is used up in production), and (31) says that if household production is undertaken, the marginal productivity of capital used by households will equal the marginal cost of capital.

Wealth	Class	Borrows	Activities	μ	λ
$0 \leq k_f < k_1$	pure wage	No	$l_w > 0$	$\lambda = 0$	$\mu > 0$
$k_1 < k_f < k_2$	wage and domestic	Yes	$l_w, l_f, k > 0$	$\lambda > 0$	$\mu = 0$
$k_2 < k_f < k_3$	pure domestic	Yes	$l_f, k < 0$	$\lambda > 0$	$\mu = 0$
$k_3 < k_f < k_4$	small capitalist	Yes	$l_f, l_h, k > 0$	$\lambda > 0$	$\mu = 0$
$k_4 < k_f < k_5$	large capitalist	Yes	$l_h, k > 0$	$\lambda > 0$	$\mu > 0$
$k_5 < k_f$	financial	No	Pure Lender	$\lambda > 0$	$\mu > 0$

Table 1: Class Structure in a Market and Domestic Production System

If the household supplies its own labor, then $l_f > 0$, so the constraint (27) is not binding, and hence $\mu = 0$. In this case, (34) asserts that if the household also works in the market sector, the marginal product of labor will be equal to the cost of labor $w(1+r+\lambda)$. Note that the cost of labor is the wage w, plus the interest that must be paid on this, rw, plus the constraint cost of the wage λw.

In this model, then, there will be six classes of households, a household's status being a function of its wealth k_f. Indeed, there is a sequence of increasing wealth levels $0 < k_1 < k_2 < k_3 < k_4 < k_5$ such that households with wealth $k_f < k_1$ are *pure wage workers*, hiring no labor or capital and working only in the market sector ($l_w > 0$). If these households have any capital ($k_f > 0$), they lend it to others. Households with $k_1 < k_f < k_2$ are *mixed wage workers and domestic producers*, working in the market sector ($l_w > 0$) but also in domestic production ($l_f > 0$) using their own capital ($k > 0$). Households with wealth $k_2 < k_f < k_3$ are *pure domestic producers*, using only their own labor ($l_f > 0$) and capital ($k_f > 0$). Households with $k_3 < k_f < k_4$ are *small capitalist producers*, using their own labor ($l_f > 0$) and supervising hired labor ($l_h > 0$), while borrowing ($k > 0$) to achieve a higher capital input to production than possible with their own wealth. Households with $k_4 < k_f < k_5$ are *large capitalist producers* who hire labor and capital ($l_h, k > 0$), supervise the hired labor, but otherwise do not engage in production ($l_f = 0$) and of course do not work for others ($l_w = 0$).

Finally, households for which $k_5 < k_f$ are *financial capitalists* who do no work themselves and do not engage in production, but rather lend all their capital and live of the proceeds. Table 1 illustrates this social equilibrium.

7 Conclusion

A scientific discipline attains maturity when it has developed a core analytical theory that is taught to all fledgling practitioners, is accepted by a large majority of seasoned practitioners, and is the basis for intradisciplinary communication. Theoretical contributions then consist of additions to and emendations of this core theory. Occasionally the core paradigm may come

under attack and be replaced by a more powerful core theory that includes all of the insights of the older doctrine, and new insights as well (Kuhn, 1962). Physics, chemistry, astronomy, and many of their subfields attained this status by the last quarter of the nineteenth century, biology developed a core theory with the synthesis of Mendelian and population genetics in the first half of the twentieth century, and economics followed in the last half of the twentieth century with the general equilibrium model (Arrow and Hahn, 1971) and neoclassical microeconomic theory (Samuelson, 1947; Mas-Colell *et al.*, 1995).

Sociology, anthropology, and social psychology have never developed core analytical theories, and indeed it is not clear why they have not coalesced into a single discipline. Sociology and anthropology have the same object of study—human society. There is no plausible justification for considering the focus of sociology on highly institutional societies and of anthropology on small-scale societies a good reason for maintaining contrasting and barely overlapping theoretical and empirical literatures. Moreover, the practice in social psychology of treating individual social behavior as capable of explanation independent of general social theory is not defensible. All these fields have suffered by separating themselves from sociobiology, which is the study of social life in general (Maynard Smith, 1982; Wilson, 1975; Alcock, 1993; Krebs and Davies, 1997).

Sociology moved haltingly towards a general analytical core with the early work of Talcott Parsons, but Parsons himself strayed into relatively tangential territory in his later work, and no one came along to pick up where Parsons left off in creating an analytical basis for sociology. Moreover, there developed a strong antagonism between economists and sociologists, which prevented sociologists from developing an analytical core that is synergistic with economic theory, while economic theory accepted unrealistic assumptions that allowed economists to model social behavior without the need for sociological notions (Gintis, 2009a). Both fields are worse for their studied mutual antipathies, but sociology has fared worse, because sociological theory since Parsons has become unacceptably fragmented (Turner, 2006).

We have presented here a suggested analytical core for sociology. Our hope is not that this view will be accepted whole cloth, but that sociologists will scrutinize and modify our offering, always with the goal of creating a body of theory that is generally acceptable. Of course, there is much that we simply do not know, especially in the area of social dynamics. The point is to settle upon what we do know and built from a common starting point.

A1 Rational Choice with Moral Values and Character Virtues

The word *rational* has many meanings in different fields. Critics of the rational actor model almost invariably attach meanings to the term that lie quite outside the narrow boundaries of rationality as used in decision theory, and

incorrectly reject the theory by referring to these extraneous meanings. We present here a set of axioms, inspired by Savage (1954), that are sufficient to derive the major tools of rational decision theory, the so-called expected utility theorem.[4]

A preference function $\succeq$ on a choice set Y is a binary relation, where $\{x \succeq y|Y\}$ is interpreted as the decision-maker preferring x to y when the choice set is Y and $x, y \in Y$.[5] We assume this binary relation has the following three properties, which must hold for any choice set Y, for all $x, y, z \in Y$ and for any set $Z \subset Y$:

1. **Completeness**: $\{x \succeq y|Y\}$ or $\{y \succeq x|Y\}$;
2. **Transitivity**: $\{x \succeq y|Y\}$ and $\{y \succeq z|Y\}$ imply $\{x \succeq z|Y\}$;
3. **Independence from irrelevant alternatives**: For $x, y \in Z$, $\{x \succeq y|Z\}$ if and only if $\{x \succeq y|Y\}$.

Because of the third property, we need not specify the choice set and can simply write $x \succeq y$. We also make the rationality assumption that the actor chooses his most preferred alternative. Formally, this means that given any choice set A, the individual chooses an element $x \in A$ such that for all $y \in A, x \succeq y$. When $x \succeq y$, we say "x is weakly preferred to y."

One can imagine cases where completeness would fail. For instance, an individual may find all alternatives so distasteful that he prefers to choose none of them. However, if "prefer not to choose" is an option then this option can be added to the choice set with an appropriate payoff. For instance, in the movie *Sophie's Choice*, a woman is asked to choose one of her two children to save from a concentration camp. The cost of the option "prefer not to choose" in this case was having both children sent to the camp.

Note that the decision-maker may have absolutely no grounds to choose x over y, knowing full well more information might show one to be preferred over the other. In this case we have both $x \succeq y$ and $y \succeq x$. In this case we say that the individual is *indifferent* between x and y. This notion of indifference leads to a well-known philosophical problem. If the preferences are transitive, then it is easy to see that indifference is also transitive. However it is easy to see that because humans have positive sensory thresholds, indifference cannot be transitive over many iterations. For instance, I may prefer more milk to less in my tea up to a certain point, but I am indifferent to amounts of milk that

[4] We regret using the term "utility" which suggests incorrectly that the theorem is related to philosophical utilitarianism or that it presupposes that all human motivation is aimed at maximizing pleasure or happiness. The weight of tradition bids us to retain the venerable name of the theorem, despite its connotational baggage.

[5] With reference to the previous section, the preference function may be seen as a consequence of the possibility of ordering decision probabilities according to size (Helbing, 1995).

differ by one molecule. Yet starting with one teaspoon of milk and adding one molecule of milk at a time, eventually I will experience an amount of milk that I prefer to one teaspoon. To address this problem, we suggest that the analysis of human behavior should avoid iterating indifference more than a few times.

The transitivity axiom is implicit in the very notion of rational choice. Nevertheless, it is often asserted that intransitive choice behavior is quite commonly observed (Grether and Plott, 1979; Ariely, 2010). In fact, most of such observations exhibit transitivity when the state-dependence (see Gintis 2007b and Section A1.1 below), time dependence (Ahlbrecht and Weber, 1995; Ok and Masatlioglu, 2003), social context dependence (Brewer and Kramer, 1986; Andreoni, 1995; Cookson, 2000; Carpenter *et al.*, 2005) and dependence of preferences are taken into account.

Independence from Irrelevant Alternatives fails when the relative value of two alternatives depends on other elements of the choice set Y, but the axiom can usually be restored by suitably redefining the choice situation (Gintis 2009a, Ch. 1). For example, suppose the relative quality of two dishes at a restaurant can be inferred from the menu of choices available to the diner. For example, a diner may prefer fish to steak when the fish is very fresh, but not otherwise. A restaurant that serves many types of fish is likely to have very fresh fish, so the diner's preference for fish vs. steak depends on the choice set available to him. This violation of Axiom A1 can be corrected by differentiating between "very fresh fish" and "fish of unknown freshness" in choice space.

The most general situation in which the Independence of Irrelevant Alternatives fails is when the choice set supplies independent information concerning the *social frame* in which the decision-maker is embedded. This aspect of choice is analyzed in Section A1.3, where we deal with the fact that the preferences are generally state-dependent; when the individual's social or personal situation changes, his preferences will change as well. Unless this factor is taken into account, rational choices may superficially appear inconsistent.

When the preference relation $\succeq$ is complete, transitive, and independent from irrelevant alternatives, we term it *consistent*. It should be clear from the above that preference consistency is an extremely weak condition that is violated only when the decision-maker is quite lacking in reasonable principles of choice.

If $\succeq$ is a consistent preference relation, then there will always exist a preference function such that individuals behave as if maximizing their preference functions over the sets Y from which they are constrained to choose. Formally, we say that a preference function $u : Y \to \mathbf{R}$ *represents* a binary relation $\succeq$ if, for all $x, y \in Y, u(x) \geq u(y)$ if and only if $x \succeq y$. We have the following theorem, whose simple proof we leave to the reader.

Theorem 2. *A binary relation $\succeq$ on the finite set Y can be represented by a preference function $u : Y \to \mathbf{R}$ if and only if $\succeq$ is consistent.*

A1.1 Rational Choice under Uncertainty

Let X be a finite set of outcomes and let $\mathcal{A}$ be a finite set of actions. We write the set of pairs (x, a) where x is an outcome and a is an action as $X \times A$. We now assume that any action $a \in A$ determines a statistical distribution of possible outcomes rather than a particular outcome. Let $\succeq$ be a consistent preference relation on $X \times \mathcal{A}$. By Theorem 2 we can associate $\succeq$ with a preference function $u : X \times \mathcal{A} \to \mathbf{R}$.

Let Ω be a finite set of *states of nature*. For instance, Ω could consist of the days of the week, so a particular state $\omega \in \Omega$ can take on the values Monday,..., Sunday, or Ω could be the set of permutations (about 8×10^{67} in number) of the 52 cards in a deck of cards, so each $\omega \in \Omega$ would be a particular shuffle of the deck. We call any $A \subseteq \Omega$ an *event*. For instance, if Ω is the days of the week, the event "weekend" would equal the set {Saturday, Sunday}, and if Ω is set of card deck permutations, the event "the top card is a queen" would be the set of permutations (about 6×10^{66} in number) in which the top card is a queen.

Following Savage (1954) we show that if the individual has a preference relation over lotteries (functions that associate states of nature $\omega \in \Omega$ with outcomes $x \in X$) that has some plausible properties, then not only can the individual's preferences be represented by a preference function, but also we can infer the probabilities the individual implicitly places on various events (his so-called subjective priors), and the expected utility principle holds for these probabilities.

Let $\mathcal{L}$ be a set of lotteries, where a *lottery* is a function $\pi : \Omega \to X$ that associates with each state of nature $\omega \in \Omega$ a outcome $\pi(\omega) \in X$. We suppose that the individual chooses among lotteries without knowing the state of nature, after which the state $\omega \in \Omega$ that obtains is revealed, so that if the individual choses action $a \in A$ that entails lottery $\pi \in \mathcal{L}$, his outcome is $\pi(\omega)$, which has payoff $u(\pi(\omega), a)$.

Now suppose the individual has a preference relation $\succ$ over $\mathcal{L} \times \mathcal{A}$. We seek a set of plausible properties of $\succ$ that together allow us to deduce (a) a preference function $u : X \times \mathcal{A} \to \mathbf{R}$ corresponding to the preference relation $\succ$ over $X \times \mathcal{A}$; (b) there is a probability distribution $p : \Omega \to \mathbf{R}$ such that the expected utility principle holds with respect to the preference relation $\succ$ over $\mathcal{L}$ and the utility function $u(\cdot, \cdot)$; i.e., if we define

$$E_\pi[u|a; p] = \sum_{\omega \in \Omega} p(\omega) u(\pi(\omega), a), \tag{A1}$$

then for any $\pi, \rho \in \mathcal{L}$ and any $a, b \in A$,

$$(\pi, a) \succ (\rho, b) \Longleftrightarrow \mathbf{E}_\pi[u|a; p] > \mathbf{E}_\rho[u|b; p]. \tag{A2}$$

We present a set of axioms that ensure (A2) formally in Gintis (2009a). Here we present these axioms more descriptively and omit a few uninteresting mathematical details. Our first condition is the rather trivial assumption that

A1. If π and ρ are two lotteries, then whether $(\pi, a) \succ (\rho, b)$ is true or false depends only on states of nature where π and ρ have different outcomes.

This axiom allows us to define a *conditional preference* $\pi \succ_A \rho$, where $A \subseteq \Omega$, which we interpret as "π is strictly preferred to ρ, conditional on event A." We define the conditional preference by revising the lotteries so that they have the same outcomes when $\omega \notin A$. Because of axiom **A1**, it does not matter what we assign to the lottery outcomes when $\omega \notin A$. This procedure also allows us to define $\succeq_A$ and $\sim_A$ in a similar manner. We say $\pi \succeq_A \rho$ if it is false that $\rho \succeq_A \pi$, and we say $\pi \sim_A \rho$ if $\pi \succeq_A \rho$ and $\rho \succeq_A \pi$.

Our second condition is equally trivial, and says:

A2. if π pays x given event A and action a, and ρ pays y given event A and action b, and if $(x, a) \succ (y, b)$, then $\pi \succ_A \rho$, and conversely.

Our third condition asserts that the decision-maker's subjective prior concerning likelihood that an event A occurs is *independent* from the payoff one receives when A occurs. More precisely, let A and B be two events, let (x, a) and (y, a) be two available choices, and suppose $(x, a) \succ (y, a)$. Let π be a lottery that pays x when action a is taken and $\omega \in A$, and pays some z when $\omega \notin A$. Let ρ be a lottery that pays y when action a is taken and $\omega \in B$, and pays z when $\omega \notin B$. We say event A is *more probable than* event B, given x, y and a if $\pi \succ \rho$. Clearly this criterion does not depend on the choice of z, by **A1**. We assume a rather strong condition:

A3. If A is more probable than B for some x, y, and a, then A is more probable than B for any other choice of x, y, and a.

This axiom, which we might term the *no wishful thinking condition*, is often violated when individuals assume that states of nature tend to conform to their ideological preconceptions, and where they reject new information to the contrary rather than update their subjective priors. Such individuals may have consistent preferences, which is sufficient to model their behavior, but their wishful thinking often entails pathological behavior. For instance, a healthy individual may understand that a certain unapproved medical treatment is a scam, but change his mind when he acquires a disease that has no conventional treatment. Similarly, an individual may attribute his child's autism to a vaccination and continue to believe this in the face of extensive evidence concerning the safety of the treatment.

The fourth condition is another trivial assumption:

A4. Suppose the decision maker prefers outcome x to any outcome that results from lottery ρ. Then the decision maker prefers a lottery n that pays x with probability one to ρ

in all states $\omega \in \Omega$, lottery π has payoff x whenever action a is taken. Suppose lottery ρ has a payoff $y = \rho(\omega)$ when action a is taken such that $x \succ y$ for all $\omega \in \Omega$. Then $\pi \succ \rho$. Conversely, if $y \succ x$ for all $\omega \in \Omega$, then $\rho \succ \pi$.

This says that if x is preferred to any outcome that may occur when lottery ρ is chosen, the decision-maker prefers the lottery π that pays x for sure to the lottery ρ.

We then have the following theorem.

Theorem 3. *Suppose A1–A4 hold. Then there is a probability function p on the state space Ω and a utility function $u\colon X \to \mathbf{R}$ such that for any $\pi, \rho \in \mathcal{L}$ and any $a, b \in \mathcal{A}$, $(\pi, a) \succ (\rho, b)$ if and only if $\mathbf{E}_\pi[u|a; p] > \mathbf{E}_\rho[u|b; p]$.*

We call the probability p the individual's *subjective prior* and say that A1–A4 imply *Bayesian rationality*, because they together imply Bayesian probability updating. Because only A3 is problematic, it is plausible to accept Bayesian rationality except in cases where some form of wishful thinking occurs, although there are other, rather exceptional, circumstances in which the expected utility theorem fails (Machina, 1989; Starmer, 2000).

A1.2 Bayesian Updating with Radical Uncertainty

We have stressed that the only problematic axiom among those needed to demonstrate the expected utility principle is the *wishful thinking* axiom A3. While there are many cases in which at least a substantial minority of social actors engage in wishful thinking, there is considerable evidence that Bayesian updating is a key neural mechanism permitting humans to acquire complex understandings of the world given severely underdetermining data. For instance, the spectrum of light waves received in the eye depends both on the color spectrum of the object being observed and the way the object is illuminated. Therefore, inferring the object's color is severely underdetermined, yet we manage to consider most objects to have constant color even as the background illumination changes. Brainard and Freeman (1997) show that a Bayesian model solves this problem fairly well, given reasonable subjective priors as to the object's color and the effects of the illuminating spectra on the object's surface.

Several students of developmental learning have stressed that children's learning is similar to scientific hypothesis testing (Carey, 1985; Gopnik and Meltzoff, 1997), but without offering specific suggestions as to the calculation

mechanisms involved. Recent studies suggest that these mechanisms include causal Bayesian networks (Glymour, 2001; Gopnik and Schultz, 2007; Gopnik and Tenenbaum, 2007). One schema, known as constraint-based learning, uses observed patterns of independence and dependence among a set of observational variables experienced under different conditions to work backward in determining the set of causal structures compatible with the set of observations (Pearl, 2000; Spirtes *et al.*, 2001). Eight-month old babies can calculate elementary conditional independence relations well enough to make accurate predictions (Sobel and Kirkham, 2007). Two-year-olds can combine conditional independence and hands-on information to isolate causes of an effect, and four-year-olds can design purposive interventions to gain relevant information (Glymour *et al.*, 2001; Schultz and Gopnik, 2004). "By age four," observe Gopnik and Tenenbaum (2007), "children appear able to combine prior knowledge about hypotheses and new evidence in a Bayesian fashion." (p. 284). Moreover, neuroscientists have begun studying how Bayesian updating is implemented in neural circuitry (Knill and Pouget, 2004).

For instance, suppose an individual wishes to evaluate an hypothesis h about the natural world given observed data x and under the constraints of a background repertoire T . The value of h may be measured by the Bayesian formula

$$P_T(h|x) = \frac{P_T(x|h)P_T(h)}{\sum_{h' \in T} P_T(x|h')P_T(h')}. \tag{A3}$$

Here, $P_T(x|h)$ is the likelihood of the observed data x, given h and the background theory T, and $P_T(h)$ gives the likelihood of h in the agent's repertoire T. The constitution of T is an area of active research. In language acquisition, it will include predispositions to recognize certain forms as grammatical and not others. In other cases, T might include physical, biological, or even theological heuristics and beliefs.

A1.3 Preferences are State-Dependent

Preferences are obviously state-dependent. For instance, Bob's preference for aspirin may depend on whether or not he has a headache. Similarly, Bob may prefer salad to steak, but having eaten the salad, he may then prefer steak to salad. These state-dependent aspects of preferences render the empirical estimation of preferences somewhat delicate, but they present no theoretical or conceptual problems.

We often observe that an individual makes a variety of distinct choices under what appear to be identical circumstances. For instance, an individual may vary his breakfast choice among several alternatives each morning without any apparent pattern to his choices.

Following Luce and Suppes (1965) and McFadden (1973), we represent this situation by assuming the individual has a utility function over bundles $x \in X$ of the form

$$u(x) = v(x) + \epsilon(x) \tag{A4}$$

where ϵ is a random error term representing the individual's current idiosyncratic taste for bundle x. This utility function induces a probability distribution π on X such that the probability that the individual chooses x is given by

$$p_x = \pi\{x \in X | \forall y \in X, v(x) + \epsilon(x) > v(y) + \epsilon(y)\}.$$

We assume $\sum_x p_x = 1$, so the probability that the individual is indifferent between choosing two bundles is zero. Now let $B = \{x \in X | p_x > 0\}$, so B is the set of bundles chosen with positive probability, and suppose B has at least three elements. express the Independence of Irrelevant Alternatives in this context by the assumption (Luce, 2005) that for all $x, y \in B$,

$$\frac{p_{yx}}{p_{xy}} = \frac{P[y|\{x,y\}]}{P[x|\{x,y\}]} = \frac{p_y}{p_x}.$$

This means that the relative probability of choosing x vs. y does not depend on whatever other bundles are in the choice set. Note that $p_{xy} \neq 0$ for $x, y \in B$. We then have

$$p_y = \frac{p_{yz}}{p_{zy}} p_z \tag{A5}$$

$$p_x = \frac{p_{xz}}{p_{zx}} p_z, \tag{A6}$$

where $x, y, z \in B$ are distinct, by the Independence of Irrelevant Alternatives. Dividing the first equation by the second in (A5), and noting that $p_y/p_x = p_{yx}/p_{xy}$, we have

$$\frac{p_{yx}}{p_{xy}} = \frac{p_{yz}/p_{zy}}{p_{xz}/p_{zx}}. \tag{A7}$$

We can write

$$1 = \sum_{y \in B} p_y = \sum_{y \in B} \frac{p_{yx}}{p_{xy}} p_x,$$

so

$$p_x = \frac{1}{\sum_{y \in B} p_{yx}/p_{xy}} = \frac{p_{xz}/p_{zx}}{\sum_{y \in B} p_{yx}/p_{zy}}, \tag{A8}$$

where the second equality comes from (A7).

Let us write

$$w(x,z) = \beta \ln p_{xz}/p_{zx},$$

so (A8) becomes

$$p_{x,B} = \frac{e^{\beta w(x,z)}}{\sum_{y\in B} e^{\beta w(x,z)}}. \tag{A9}$$

However, by the Independence of Irrelevant Alternatives, this expression must be independent of our choice of z, so if we write $w(x) = \ln p_{xz}$ for a arbitrary $z \in B$, we have

$$p_x = \frac{e^{\beta w(x)}}{\sum_{y\in B} e^{\beta w(y)}}. \tag{A10}$$

Note that there is one free variable, β, in (A10). This represents the degree to which the individuals are relatively indifferent among the alternatives. As $\beta \to \infty$, the individual chooses his most preferred alternative with increasing probability, and with probability one in the limit. As $\beta \to 0$, the individual becomes more indifferent to the alternative choices.

This model helps to explain the compatibility of the *preference reversal* phenomenon (Lichtenstein and Slovic, 1971; Grether and Plott, 1979; Tversky and Kahneman, 1990; Kirby and Herrnstein, 1995; Berg *et al.*, 2005) with the rationality postulate. As we explain in Gintis (2007b), in the cases discussed in the experimental literature, the experimenters offer only alternative lotteries with expected values that are very close to being equal to one another. Thus decision-makers are virtually indifferent among the choices based on the expected return criterion, so even a small influence of the social frame in which the experimenters embed the choice situation on the subjects' preference state may strongly affect their choices. For experimental support for this interpretation, see Sopher and Gigliotti (1993).

B1 Epistemic Games and Correlated Equilibria

A game G consists of a set of n players, where each player i can choose a move s_i from a set S_i of available moves. A *strategy profile* $s = \{s_1, \ldots, s_n\}$ is a choice of a move for each player. The game has a payoff $\pi_i(s)$ for each player i and each strategy profile s. Thus in general the payoff to a player depends not only on the player's own behavior, but on the behavior of the other players as well. If s is a strategy profile, we write s_i for i's move in this strategy profile, and s_{-i} for the moves of all the other players. We can then write $s = (s_i, s_{-i})$ and $\pi_i(s) = \pi_i(s_i, s_{-i})$.

We say G is an *epistemic game* if each player i has a *conjecture* ϕ_i , which is a probability distribution over the strategies $\{s_{-i}\}$ that the other players are using. A player i is called *rational* in the epistemic game G if i's move s_i

maximizes his expected payoff with respect to the conjecture ϕ_i. A rational player i thus maximizes the expression

$$\pi_i(s_i) = \sum_{s_{-i}} \phi_i(s_{-i})\pi(s_i, s_{-i}). \tag{B1}$$

Equation (B1) follows from the following reasoning. For every possible profile of moves s_{-i} of the other players, the probability that the other players actually make these moves if $\phi_i(s_{-i})$, and the payoff to i with these moves is $\pi_i(s_i, s_{-i})$. The contribution of (s_i, s_{-i}) to the expected payoff in this case is the product of these three numbers, and his expected payoff is the sum of this expression over all possible strategy profiles s_{-i} of the other players. Rational players, in short, choose a move that maximizes their expected payoff, given their conjectures concerning the moves of the other players.

A Nash equilibrium of a game G is a strategy profile s such that the move s_i by each player i maximizes i's payoff, given the strategy profile s_{-i} of the other players. If all players are rational and their conjectures $\{\phi_1, \ldots, \phi_n\}$ are correct, then each player i will necessarily play a *best response* s_i to the other players expected strategy profile, meaning choosing a move that maximizes (B1).

It may seem that rational players must play a Nash equilibrium, but this is wrong for two reasons. First, for any player i, there may be many best responses s_i, only one of which is part of a Nash equilibrium. For instance, consider the two-player game, Throwing Fingers, depicted in Figure B1. In this game, each player has two strategies, throw one finger (c_1) or throw two fingers (c_2). If the number of fingers thrown is even, the first player wins \$1 from the second player, and if the number of fingers is odd, the second player wins \$1 from the first. There is a unique Nash equilibrium to this game, in which each player throws one finger with the probability 1/2. However, if one player chooses this strategy, then *all* strategies of the second player are best responses, and vice-versa. The rationality assumption therefore does not give either player a reason for playing his part in the Nash equilibrium. Note that even if this is an epistemic game, and each player conjectures that the other will play the Nash equilibrium strategy, either player, however rational, still can choose any strategy at all to play.

The second reason rational players need not choose best responses that form a Nash equilibrium is that their conjectures may not be correct. For a concrete example, consider a society in which men and women prefer each other's company, but when a couple, say Bob and Alice, goes out for the evening, Bob prefers one form of entertainment, m, and Alice prefers another, f . This is thus a two player game in which each player has two moves, m and f, so there are four possible strategy profiles, (m, m), (m, f), (f, m), and (f, f), where the first entry is Bob's move and the second is Alice's move.

	c_1	c_2
c_1	$1, -1$	$-1, 1$
c_2	$-1, 1$	$1, -1$

Figure B1: Throwing Fingers

		Alice
Bob	m	f
m	$2, 1$	$0, 0$
f	$0, 0$	$1, 2$

Figure B2: The Battle of the Sexes Game

The game is called the Battle of the Sexes, and the payoffs are described in Figure B2. There are two pure strategy Nash equilibria, the strategy profiles, (f, f) with payoffs $(1, 2)$ and (m, m) with payoffs $(2, 1)$, as well as a mixed strategy equilibrium, in which both players choose their favorite entertainment with probability $1/3$, resulting in the payoff $2/3$ to each.

Suppose Bob conjectures that Alice will play f with probability $1/2$, and Alice conjectures that Bob will play m with probability $1/2$. Then both players will choose their preferred form of entertainment and their payoffs will both be zero! Similarly, if both players conjecture that their partner will play his or her preferred entertainment with probability one, then again each will choose the other's preferred entertainment, and both will still have payoff zero. Even if both players choose the mixed strategy Nash equilibrium strategy, they will coordinate on m with probability $1/3 \times 2/3 = 2/9$, on f with the same probability, and they will choose different forms of entertainment, with zero payoff, with probability $5/9$. Rationality clearly gives no satisfactory solution to playing this game efficiently. A correlated equilibrium, we will see, does the job nicely.

A correlated equilibrium of an epistemic game G is a Nash equilibrium of a game $\mathcal{G}^+$, which is G augmented by a non-player, the choreographer (Aumann, 1974; Aumann, 1987a). Rather than give a general definition of $\mathcal{G}^+$, we will develop the notion in the context of the Battle of the Sexes. The choreographer chooses m with some probability p and sends a message to both players saying "Play m," and with probability $1 - p$ the choreographer sends both players the message "Play f." The two players know that the choreographer will

send both players the same message. Therefore obeying the choreographer's command is a best response for each. The payoffs for the two players is now $(p + 2(1 - p), 2(1 - p) + p)$. For instance, if $p = 1/2$, we get the efficient and egalitarian payoff $(3/2, 3/2)$.

Note that the choreographer does not have to be another player, or even a person. For example, a social norm that says "Play m on even-numbered days and f on odd-numbered days" would do the job perfectly well.

In the general case, the choreographer observes a random variable γ and issues a directive to player i to choose the pure strategy $s_i = f_i(\gamma)$. The choreographer's directives must be chosen so that if the players know the probability distribution of the random variable y and the choreographer's behavior $\{f_1(\gamma), \ldots, f_n(\gamma)\}$, then it is a best response for each to follow the choreographer's directive. The resulting equilibrium is called a correlated equilibrium.

It is easy to see that every Nash equilibrium is also a correlated equilibrium in which the choreographer issues the directive $s = (s_1, \ldots, s_n)$ with the same probability with which this pure strategy profile is played in the Nash equilibrium.

In addition, it is easy to show that *any weighted sum of correlated equilibria is itself a correlated equilibrium.* In this case the choreographer observes a random variable $\gamma = (\gamma_1, \gamma_2)$ where γ_1 tells the choreographer which Nash equilibrium to implement, and γ_2 is the random variable for the chosen game. Note that every weighted sum of Nash equilibria is a correlated equilibrium, although it is not necessarily itself a Nash equilibrium. For instance, the egalitarian payoff $(3/2, 3/2)$ in the Battle of the sexes is the weighted sum of the two pure strategy Nash equilibria, with equal weights $(1/2, 1/2)$. There is no Nash equilibrium with these payoffs.

The key reason that the correlated equilibrium concept is so powerful is that it can also be shown that if the players in epistemic game $\mathcal{G}$ are rational and if there is a random variable γ that all players use formulate their conjectures, so that the strategy choice for each player i can be written in the form $s_i = s_i(\gamma)$, then the strategy profile chosen by the players is a correlated equilibrium. The choreographer in this epistemic game uses the random variable γ and simply sets $f_i(\gamma) = s_i(\gamma)$ for each player i.

B1.1 Common Priors and Social Norm Equilibria

The identity between correlated equilibria and rationality developed above highlights an assumption that lies at the heart of a game-theoretic concept of social norms. This is the requirement that the players have a *common prior* concerning the moves of the choreographer. If the choreographer assigns a strict best response to each player, it is clear that some amount of heterogeneity in priors will not destroy the equilibrium. Moreover, if there are known "types" of

players (e.g., Optimists and Pessimists) whose priors are distinct but commonly known, and the population composition is commonly known, it is usually possible to represent this situation in terms of common priors with respect to a more complex random variable, for which a more complex correlated equilibrium exists.

However, when common priors are lacking and the actual composition and frequency distribution of priors are not held in common for some suitably enlarged state space, the social norm analysis will fail to apply. Rational agents with fundamental disagreements as to the actual structure of their social life do not dance to any choreographer's tune.

B1.2 The Omniscient Choreographer and Moral Preferences

We want to stress that the assertion that rational players always choose to implement some correlated equilibrium requires that the choreographer be *omniscient* in the sense of knowing how the rationality of the players leads them to particular choices $\{s_i(\gamma)\}$ when they commonly observe a value of the random variable γ. For an example of how this can fail, suppose that a game has both *honest* and *dishonest* players, and there is some aspect the players' behavior that cannot be observed by the choreographers. For instance, a dishonest policemen might take a bribe, while an honest policeman will not. Because the choreographer cannot tell the difference between the two types of players, he must issue them the same directive, and the choreographer will receive the same information as to the social actor's behavior whether he is honest or dishonest, so the dishonest player cannot be induced to behave honestly.

We can summarize this problem by saying that the applicability of the correlated equilibrium concept requires either that the choreographer be omniscient, so there are no possibilities for dishonest behavior, or the players must have a moral commitment to honesty. Indeed, many social norms modelers, including Bicchieri (2006), predicate their analysis on the fact that rational individuals may have other-regarding preferences and/or may value certain moral virtues so that they voluntarily conform to a social norm in a situation where as perfectly self-regarding and amoral agent would not. In such cases, the choreographer may be obeyed even at a cost to the players.

For instance, each agent's payoff might consist of a *public component* that is known to the choreographer and a *private component* that reflects the idiosyncrasies of the agent and is unknown to the choreographer. Suppose the maximum size of the private component in any state for an agent is α, but the agent's inclination to follow the choreographer has strength greater than α. Then, the agent continues to follow the choreographer's directions whatever the state of his private information. Formally, we say an individual has an α-*normative predisposition* towards conforming to the social norm if he

strictly prefers to play his assigned strategy so long as all his pure strategies have payoffs no more than α greater than when following the choreographer. We call an α-normative predisposition a *social preference* because it facilitates social coordination but violates self-regarding preferences for $\alpha > 0$. There are evolutionary reasons for believing that humans have evolved such social preferences for fairly high levels of α through gene-culture coevolution (Bowles and Gintis, 2011; Grund *et al.*, 2013).

Suppose, for example, that police in a certain town are supposed to apprehend criminals, where it costs police officer i a variable amount f_i to file a criminal report. For instance, if the identified perpetrator is in the same ethnic group as i, or if the perpetrator offers a bribe to be released, f_i might be very high, whereas an offender from a different ethnic group, or one who does not offer a bribe, might entail a low value of f_i . How can this society erect incentives to induce the police to act in a non-corrupt manner?

Assuming police officer i is self-regarding and amoral, i will report a crime only if $f_i \leq w$, where w is the reward for filing an accurate criminal report (accuracy can be guaranteed by fact-checking). A social norm equilibrium requires that all apprehended criminals be prosecuted cannot then be sustained, because all officers for whom $f_i < w$ with positive probability will at least at times behave corruptly. Suppose however that officers have a normative predisposition to behave honestly, in the form of a police culture favoring honesty that is internalized by all officers. If $f_i < w + \alpha$ with probability one for all officers i, where α is the strength of police culture, the social norm equilibrium can be sustained, despite the fact that the choreographer has incomplete information concerning events in which criminal behavior is detected.

For a more realistic example, consider a town with a North-South/East-West array of streets. In the absence of a social norm, whenever two cars find themselves in a condition of possible collision, both stop and each waits for the other go first. Obviously not a lot of driving will get done. So, consider a social norm in which (a) all cars drive on the right, (b) at an intersection both cars stop and the car that arrived first proceeds forward, and (c) if both cars arrive at an intersection at the same time, the car that sees the other car on its left proceeds forward. This is one of several social norms that will lead to an efficient use of the system of streets, provided there is not too much traffic. The social norm serves as a choreographer giving rise to a self-enforcing correlated equilibrium.

Suppose, however, that there is so much traffic that cars spend much of their time stopping at crossings. We might then prefer the social norm in which we amend the above social norm to say that cars traveling North-South always have the right of way and need not stop at intersections. However, if there is really heavy traffic, East-West drivers may never get a chance to move forward at all using this social norm.

Suppose, then, we erect a set of signals at each intersection that indicate "Go" or "Stop" to drivers moving in one direction and another set of "Go" or "Stop" signals for drivers moving in the crossing direction. We can then correlate the signals so that when one set of drivers see "Go", the other set of drivers see "Stop." The social norm then says that "if you see Go, do not stop at the intersection, but if you see Stop, then stop and wait for the signal to change to Go." We add to the social norm that the system of signals alternates sufficiently rapidly and there is a sufficiently effective surveillance system that no driver has an incentive to disobey the social norm.

This would appear to be a perfect example of a social norm, indeed a convention. However, the original game does not have a system of signals, and the proposed social norm does not single out a Nash equilibrium of the original game. Indeed, it is easy to see that there is a wide array of payoffs in the original game in which the only Nash equilibrium is for both cars to stop when an encounter occurs.

B1.3 Why Alternative Social Norm May Proliferate

There are important implications of the fact that a social norm is the choreographer of a correlated equilibrium rather than a Nash equilibrium. A simple game G may have many qualitatively distinct correlated extensions $\mathcal{G}^+$, which implies that life based on social norms can be significantly qualitatively richer than the simple underlying games that they choreograph. The correlated equilibrium concept thus indicates that social theory goes beyond game theory to the extent that it supplies dynamical and equilibrium mechanisms for the constitution and transformation of social norms. At the same time, the power of the correlated equilibrium interpretation of social norms indicates that social theory that rejects game theory is likely to be significantly handicapped.

C1 The Evolution of Social Conventions

A *Markov process* M consists of a finite number of states $S = \{1, \ldots, n\}$, and an n-dimensional square matrix $P = \{p_{ij}\}$ such that p_{ij} represents the probability of making a transition from state i to state j. A *path* $\{i_1, i_2, \ldots\}$ determined by Markov process M consists of the choice of an initial state $i_1 \in S$, and if the process is in state i in period $t = 1, 2, \ldots$, then it is in state j in period $t + 1$ with probability p_{ij}. Despite the simplicity of this definition, finite Markov processes are remarkably flexible in modeling dynamical systems, although characterizing their long-run properties becomes challenging for systems with many states.

We will use the Markov process as a tool to model the evolution of money as a convention in trade among many individuals. Consider a rudimentary

economy in which there are g goods, and each social actor produces one unit of one of these goods in each period. After production takes place, individuals encounter one another randomly and they trade equal amounts of their wares if each wants what the other is offering. However, it often happens that one of the pair does not consume what the other produces, so no direct trade is possible. However, suppose each social actor is willing to accept one of the g goods not for consumption, but rather to use as money in trading with other producers. The use of money increases the efficiency of the economy because the frequency of welfare-increasing trades is higher with the use of money. Moreover, it is clear that the highest efficiency would be attained if all social actors were willing to accept the *same* good as money. Under what conditions might this occur without a central government or other macrosocial institution bringing this about?

To pose the question more formally, what is the long run distribution of the fraction of the population accepting each of the g goods as money? To answer this question, we must make some assumption concerning how individual traders decide to change the good they are willing to accept as money. We simply assume that one of the n traders in the economy in each period switches to the money type of a randomly chosen trading partner. We represent the state of the economy as $(w_1 \cdots w_g)$, where w_i is the number of agents who accept good i as money. The total number of states in the economy is thus the number of different ways to distribute n indistinguishable balls (the n agents) into g distinguishable boxes (the g goods), which is $C(n+g-1, g-1)$, where

$$C(n,g) = \frac{n!}{(n-g)!g!}$$

is the number of ways to choose g objects from a set of n objects. For instance, if there are 100 social actors ($n = 100$) and ten goods ($g = 10$), then the number of states S in the system is $S = C(109, 9) = 4,263,421,511,271$.

To verify this formula, write a particular state in the form

$$s = x \ldots xAx \ldots xAx \ldots xAx \ldots x$$

where the number of x's before the first A is the number of agents choosing type 1 as money, the number of x's between the $(i-1)^{\text{th}} A$ and the $i^{\text{th}} A$ is the number of agents choosing type i as money, and the number of x's after the final A is the number agents choosing type g as money. The total number of x's is equal to n, and the total number of A's is $g-1$, so the length of s is $n+g-1$. Every placement of the $g-1$ A's represents particular state of the system, so there are $C(n+g-1, g-1)$ states of the system.

Suppose that in each period two agents are randomly chosen and the first agent switches to using the second agent's money type as his own money. This gives a determinate probability p_{ij} of shifting from one state i of the system

to any other state j . The matrix $P = \{p_{ij}\}$ is called a *transition probability matrix*, and the whole stochastic system is clearly a finite Markov process.

What is the long-run behavior of this Markov process? Note first that if we start in state i at time $t = 1$, the probability $p_{ij}^{(2)}$ of being in state j in period $t = 2$ is simply

$$p_{ij}^{(2)} = \sum_{k=1}^{S} p_{ik} p_{kj} = (P^2)_{ij}. \tag{C1}$$

This is true because to be in state j at $t = 2$ the system must have been in some state k at $t = 1$ with probability p_{ik}, and the probability of moving from k to j is just p_{kj}. This means that the two period transition probability matrix for the Markov process is just P^2, the matrix product of P with itself. By similar reasoning, the probability of moving from state i to state j in exactly r periods, is P^r. Therefore, the time path followed by the system starting in state $s^0 = i$ at time $t = 0$ is the sequence $s^0, s^1, \ldots$, where

$$P[s^t = j | s^0 = i] = (P^t)_{ij} = P_{ij}^{(t)}.$$

The matrix P in our example has $S^2 \approx 1.818 \times 10^{15}$ entries. The notion of calculation P^t for even small t is quite infeasible. There are ways to reduce the calculations by many orders of magnitude (Gintis, 2009b, Ch. 13), but these methods are completely impractical with so large a Markov process.

Nevertheless, we can easily understand the dynamics of this Markov process. We first observe as that if the Markov process is ever in the state

$$s_*^r = (0_1, \ldots, 0_{r-1}, n_r, 0_{r+1} \ldots 0_k),$$

where all n agents choose type r money, then s_*^r will be the state of the system in all future periods. We call such a state *absorbing*. There are clearly only g absorbing states for this Markov process.

We next observe that from any non-absorbing state s, there is a strictly positive probability that the system moves to an absorbing state before returning to state s. For instance, suppose $w_i = 1$ in state s. There is then a positive probability that w_i increases by 1 in each of the next $n - 1$ periods, so the system is absorbed into state s_*^i without ever returning to state s. Now let $p_s > 0$ be the probability that Markov process never returns to state s. The probability that the system returns to state s at least q times is thus at most $(1 - p_s)^q$. Since this expression goes to zero as $q \to \infty$, it follows that the state s appears only a finite number of times with probability one. We call s a *transient* state.

We can often calculate the probability that a system starting out with of w_r agents choosing type r as money, $r = 1, \ldots, g$ is absorbed by state r. Let us think of the Markov process as that of g gamblers, each of whom starts out

with an integral number of coins, there being n coins in total. The gamblers represent the types and their coins are the agents who choose that type for money, there being n agents in total. We have shown that in the long run, one of the gamblers with have all the coins, with probability one. Suppose that the game is fair in the sense that in any period a gambler with a positive number of coins has an equal chance to increase or decrease his wealth by one coin. Then the expected wealth of a gambler in period $t+1$ is just his wealth in period t. Similarly, the expected wealth $E[w^{t'}|w^t]$ in period $t'>t$ of a gambler whose wealth in period t is w^t is $E[w^{t'}|w^t]=w^t$. This means that if a gambler starts out with wealth $w>0$ and he wins all the coins with probability q_w, then $w=q_w n$, so the probability of being the winner is just $q_w=w/n$.

We now can say that this Markov process, despite its enormous size, can be easily described as follows. Suppose the process starts with w_r agents holding good r. Then in a finite number of time periods, the process will be absorbed into one of the states $1,\ldots,g$, and the probability of being absorbed into state r is w_r/n. In all cases, a single good will eventually evolve as the universal medium of exchange.

Of course the assumption that all traders are willing to adopt any good as money may be unrealistic. For instance, the producers of a particular good i can benefit from having good i as money because it increases their demand. If exactly one of the producer types simply refused to accept any good but their own as money, while all other groups were unbiased in their choice of money, eventually good i will be the universal money good. However, if more than one type of producer adopts this intransigent strategy, an irreducible conflict must obtain.

References

Abbott, R. J., J. K. James, R. I. Milne, and A. C. M. Gillies. 2003. "Plant Introductions, Hybridization and Gene Flow". *Philosophical Transactions of the Royal Society of London B.* 358: 1123–1132.

Ahlbrecht, M. and M. Weber. 1995. "Hyperbolic Discounting Models in Prescriptive Theory of Intertemporal Choice". *Zeitschrift für Wirtschaftsund Sozialwissenschaften.* 115: 535–568.

Alcock, J. 1993. *Animal Behavior: An Evolutionary Approach.* Sunderland, MA: Sinauer.

Alexander, R. D. 1987. *The Biology of Moral Systems.* New York: Aldine.

Allman, J., A. Hakeem, and K. Watson. 2002. "Two Phylogenetic Specializations in the Human Brain". *Neuroscientist.* 8: 335–346.

Andreghetto, G., J. Brandts, R. Conte, J. Sabater-Mir, and H. Solaz. 2013. "Punish and Voice: Punishment Enhances Cooperation when Combined with Norm-Signaling". *PLOSOne.* 8(6): 1–7.

Andreoni, J. 1995. "Warm-Glow versus Cold-Prickle: The Effects of Positive and Negative Framing on Cooperation in Experiments". *Quarterly Journal of Economics*. 110(1): 1–21.

Andreoni, J. and J. H. Miller. 2002. "Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism". *Econometrica*. 70(2): 737–753.

Ariely, D. 2010. *Predictibly Irrational:The Hidden Forces That Shape Our Decisions*. New York: Harper.

Arrow, K. J. and G. Debreu. 1954. "Existence of an Equilibrium for a Competitive Economy". *Econometrica*. 22: 265–290.

Arrow, K. J. and F. Hahn. 1971. *General Competitive Analysis*. San Francisco: Holden-Day.

Aumann, R. J. 1974. "Subjectivity and Correlation in Randomizing Strategies". *Journal of Mathematical Economics*. 1: 67–96.

Aumann, R. J. 1987a. "Correlated Equilibrium and an Expression of Bayesian Rationality". *Econometrica*. 55: 1–18.

Aumann, R. J. 1987b. "Game Theory". In: *The New Palgrave: A Dictionary of Economics*. Ed. by J. Eatwell, M. Milgate, and P. Newman. Vol. 2. London: Macmillan.

Beer, J. S., E. A. Heerey, D. Keltner, D. Skabini, and R. T. Knight. 2003. "The Regulatory Function of Self-conscious Emotion: Insights from Patients with Orbitofrontal Damage". *Journal of Personality and Social Psychology*. 65: 594–604.

Belin, P., R. J. Zatorre, P. Lafaille, P. Ahad, and B. Pike. 2000. "Voice-selective Areas in Human Auditory Cortex". *Nature*. 403: 309–312.

Benabou, R. and J. Tirole. 2002. "Self Confidence and Personal Motivation". *Quarterly Journal of Economics*. 117: 871–915.

Berg, J. E., J. W. Dickhaut, and T. A. Rietz. 2005. *Preference Reversals: The Impact of Truth-Revealing Incentives*. College of Business: University of Iowa.

Bicchieri, C. 1993. *Rationality and Coordination*.

Bicchieri, C. 2006. *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge: Cambridge University Press.

Binder, J. R., J. A. Frost, T. A. Hammeke, R. W. Cox, S. M. Rao, and T. Prieto. 1997. "Human Brain Language Areas Identified by Functional Magnetic Resonance Imaging". *Journal of Neuroscience*. 17: 353–362.

Binmore, K. G. 1993. *Game Theory and the Social Contract: Playing Fair*. Cambridge, MA: MIT Press.

Binmore, K. G. 1998. *Game Theory and the Social Contract: Just Playing*. Cambridge, MA: MIT Press.

Binmore, K. G. 2005. *Natural Justice*. Oxford: Oxford University Press.

Boehm, C. 1999. *Hierarchy in the Forest: The Evolution of Egalitarian Behavior*. Cambridge, MA: Harvard University Press.

Bonner, J. T. 1984. *The Evolution of Culture in Animals*. Princeton: Princeton University Press.

Bowles, S. 2004. *Microeconomics: Behavior, Institutions, and Evolution*. Princeton: Princeton University Press.

Bowles, S. and H. Gintis. 1993. "The Revenge of Homo economicus: Contested Exchange and the Revival of Political Economy". *Journal of Economic Perspectives*. 7(1): 83–102.

Bowles, S. and H. Gintis. 2004. "Persistent Parochialism: Trust and Exclusion in Ethnic Networks". *Journal of Economic Behavior and Organization*. 55(1): 1–23.

Bowles, S. and H. Gintis. 2011. *A Cooperative Species: Human Reciprocity and its Evolution*. Princeton: Princeton University Press.

Boyd, R. and P. J. Richerson. 1985. *Culture and the Evolutionary Process*. Chicago: University of Chicago Press.

Boyd, R. and P. J. Richerson. 1988. "An Evolutionary Model of Social Learning: the Effects of Spatial and Temporal Variation". In: *Social Learning: Psychological and Biological Perspectives*. Ed. by T. R. Zentall and G. G. Jr. Hillsdale, NY: Erlbaum.

Boyd, R. and P. J. Richerson. 2004. *The Nature of Cultures*. Chicago, IL: University of Chicago Press.

Brainard, D. H. and W. T. Freeman. 1997. "Bayesian Color Constancy". *Journal of the Optical Society of America*. A 14: 1393–1411.

Brewer, M. B. and R. M. Kramer. 1986. "Choice Behavior in Social Dilemmas: Effects of Social Identity, Group Size, and Decision Framing". *Journal of Personality and Social Psychology*. 50(543): 543–549.

Brown, D. E. 1991. *Human Universals*. New York: McGraw-Hill.

Brown, J. H. and M. V. Lomolino. 1998. *Biogeography*. Sunderland, MA: Sinauer.

Brown, M., A. Falk, and E. Fehr. 2004. "Relational Contracts and the Nature of Market Interactions". *Econometrica*. 72: 747–780.

Burrows, A. M. 2008. "The Facial Expression Musculature in Primates and Its Evolutionary Significance". *BioEssays*. 30: 212–225.

Camerer, C. 2003. *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton: Princeton University Press.

Camerer, C. F. and E. Fehr. 2004. "Measuring Social Norms and Preferences Using Experimental Games: A Guide for Social Scientists". In: *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies*. Ed. by J. Henrich, R. Boyd, S. Bowles, C. F. Camerer, E. Fehr, and H. Gintis. Oxford: Oxford University Press.

Camille, N. 2004. "The Involvement of the Orbitofrontal Cortex in the Experience of Regret". *Science*. 304: 1167–1170.

Carey, S. 1985. *Conceptual Change in Childhood*. Cambridge: MIT Press.

Carpenter, J. P., S. V. Burks, and E. Verhoogen. 2005. "Comparing Student Workers: The Effects of Social Framing on Behavior in Distribution Games". *Research in Experimental Economics*. 1: 261–290.

Cavalli-Sforza, L. L. 1986. "Cultural Evolution". *American Zoologist*. 26: 845–855.

Cavalli-Sforza, L. L. and M. W. Feldman. 1981. *Cultural Transmission and Evolution*. Princeton: Princeton University Press.

Cavalli-Sforza, L. L. and M. W. Feldman. 1982. "Theory and Observation in Cultural Transmission". *Science*. 218: 19–27.

Coleman, J. S. 1988. "Free Riders and Zealots: The Role of Social Networks". *Sociological Theory*. 6: 52–57.

Conte, R. and C. Castelfranchi. 1999. "From Conventions to Prescriptions. Towards an Integrated View of Norms". *Artificial Intelligence and Law*. 7: 323–340.

Cookson, R. 2000. "Framing Effects in Public Goods Experiments". *Experimental Economics*. 3: 55–79.

Cosmides, L. and J. Tooby. 1992. "Cognitive Adaptations for Social Exchange". In: *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. Ed. by J. H. Barkow, L. Cosmides, and J. Tooby. New York: Oxford University Press. 163–228.

Cosmides, L., J. Tooby, and J. H. Barkow. 1992. "Introduction: Evolutionary Psychology and Conceptual Integration". In: *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. Ed. by J. H. Barkow, L. Cosmides, and J. Tooby. New York: Oxford University Press. 3–15.

Dawkins, R. 1976. *The Selfish Gene*. Oxford: Oxford University Press.

Dawkins, R. 1982. *The Extended Phenotype: The Gene as the Unit of Selection*. Oxford: Freeman.

Dawkins, R. 2004. "Extended Phenotype-but not too Extended. A reply to Laland, Turner and Jablonka". *Biology and Philosophy*. 19: 377–396.

Debreu, G. 1952. "A Social Equilibrium Existence Theorem". *Proceedings of the National Academy of Sciences*. 38: 886–893.

Di Guilmi, C., M. Gallegati, S. Landini, and J. E. Stiglitz. 2012. "Towards an Analytical Solution for Agent Based Models: An Application to a Credit Network Economy". In: *Approaches to the Evolving World Economy: Complex Dynamics, Norms, and Organizations*. Ed. by H. Gintis. London: Palgrave.

DiMaggio, P. 1994. "Culture and Economy". In: *The Handbook of Economic Sociology*. Ed. by N. Smelser and R. Swedberg. Princeton: Princeton University Press. 27–57.

DiMaggio, P. 1998. "The New Institutionalisms : Avenues of Collaboration". *Journal of Institutional and Theoretical Economics*. 154(4): 696–705.

Dugatkin, L. A. and H. K. Reeve. 1998. *Game Theory and Animal Behavior*. Oxford: Oxford University Press.

Dunbar, R. M. 1993. "Coevolution of Neocortical Size, Group Size and Language in Humans". *Behavioral and Brain Sciences*. 16(4): 681–735.

Dunbar, R. M. 2005. *The Human Story*. New York: Faber & Faber.

Durkheim, E. 1902. *The Division of Labor in Society*. New York: The Free Press.

Durrett, R. and S. A. Levin. 2005. "Can Stable Social Groups be Maintained by Homophilous Imitation Alone?" *Journal of Economic Behavior and Organization*. 57: 267–286.

Eswaran, M. and A. Kotwal. 1986. "Access to Capital and Agrarian Production Organization". *Economic Journal*. 96: 482–498.

Fehr, E. and U. Fischbacher. 2004. "Third Party Punishment and Social Norms". *Evolution & Human Behavior*. 25: 63–87.

Fehr, E. and H. Gintis. 2007. "Human Motivation and Social Cooperation: Experimental and Analytical Foundations". *Annual Review of Sociology*. 33: 43–64.

Feldman, M. W. and L. L. Cavalli-Sforza. 1976. "Cultural and Biological Evolutionary Processes, Selection for a Trait under Complex Transmission". *Theoretical Population Biology*. 9(2): 238–259.

Feldman, M. W. and L. A. Zhivotovsky. 1992. "Gene-Culture Coevolution: Toward a General Theory of Vertical Transmission". *Proceedings of the National Academy of Sciences*. 89: 11935–11938.

Fischer, I., A. Frid, S. J. Goerg, S. A. Levin, D. I. Rubenstein, and R. Selten. 2013. "Fusing Enacted and Expected Mimicry Generates a Winning Strategy that Promotes the Evolution of Cooperation". *Procedings of the National Academy of Sciences*. 110 (25): 10229–10233.

Fishburn, P. C. 1970. *Utility Theory for Decision Making*. New York: John Wiley & Sons.

Fudenberg, D., D. K. Levine, and E. Maskin. 1994. "The Folk Theorem with Imperfect Public Information". *Econometrica*. 62: 997–1039.

Gadagkar, R. 1991. "On Testing the Role of Genetic Asymmetries Created by Haplodiploidy in the Evolution of Eusociality in the Hymenoptera". *Journal of Genetics*. 70(1): 1–31.

Gauthier, D. 1986. *Morals by Agreement*. Oxford: Clarendon Press.

Gigerenzer, G. and P. M. Todd. 1999. *Simple Heuristics That Make Us Smart*. New York: Oxford University Press.

Gintis, H. 1975. "Welfare Economics and Individual Development: A Reply to Talcott Parsons". *Quarterly Journal of Economics*. 89(2): 291–302.

Gintis, H. 2000. "Strong Reciprocity and Human Sociality". *Journal of Theoretical Biology*. 206: 169–179.

Gintis, H. 2003a. "The Hitchhiker's Guide to Altruism: Genes, Culture, and the Internalization of Norms". *Journal of Theoretical Biology*. 220(4): 407–418.

Gintis, H. 2003b. "Solving the Puzzle of Human Prosociality". *Rationality and Society*. 15(2): 155–187.

Gintis, H. 2007a. "The Dynamics of General Equilibrium". *Economic Journal*. 117: 1289–1309.
Gintis, H. 2007b. "A Framework for the Unification of the Behavioral Sciences". *Behavioral and Brain Sciences*. 30, no. 1: 1–61.
Gintis, H. 2009a. *The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences*. Princeton: Princeton University Press.
Gintis, H. 2009b. *Game Theory Evolving, Second Edition*. Princeton: Princeton University Press.
Gintis, H. 2011. "Gene-culture Coevolution and the Nature of Human Sociality". *Proceedings of the Royal Society*. B 366: 878–888.
Gintis, H. 2013. "Markov Models of Social Exchange: Theory and Applications". *ACM Transactions in Intelligent Systems and Technology*. 4: 53.
Gintis, H., S. Bowles, R. Boyd, and E. Fehr. 2005. *Moral Sentiments and Material Interests: On the Foundations of Cooperation in Economic Life*. Cambridge, MA: MIT Press.
Gintis, H. and A. Mandel. 2012. "The Stability of Walrasian General Equilibrium". (in press).
Glymour, A., D. M. Sobel, L. Schultz, and C. Glymour. 2001. "Causal Learning Mechanism in Very Young Children: TwoThree- and four-year-olds Infer Causal Relations from Patterns of Variation and Covariation". *Developmental Psychology*. 37(50): 620–629.
Glymour, C. 2001. *The Mind's Arrows: Bayes Nets and Graphical Causal Models in Psychology*. Cambridge: MIT Press.
Gneezy, U. and A. Rustichini. 2000. "A Fine Is a Price". *Journal of Legal Studies*. 29: 1–17.
Gopnik, A. and A. Meltzoff. 1997. *Words, Thoughts, and Theories*. Cambridge: MIT Press.
Gopnik, A. and L. Schultz. 2007. *Causal Learning, Psychology, Philosophy, and Computation*. Oxford: Oxford University Press.
Gopnik, A. and J. B. Tenenbaum. 2007. "Bayesian Networks, Bayesian Learning and Cognitive Development". *Developmental Studies*. 10: 281–287.
Gould, S. J. and E. Vrba. 1981. "Exaption: A Missing Term in the Science of Form". *Paleobiology*. 8: 4–15.
Granovetter, M. 1985. "Economic Action and Social Structure: The Problem of Embeddedness". *American Journal of Sociology*. 91: 481–510.
Granovetter, M. 1995. "The Economic Sociology of Firms and Entrepreneurs". In: *The Economic Sociology of Immigration: Essays on Networks, Ethnicity, and Entrepreneurship*. Ed. by A. Portes. New York: Russell Sage. 128–165.
Grether, D. and C. Plott. 1979. "Economic Theory of Choice and the Preference Reversal Phenomenon". *American Economic Review*. 69 (4): 623–638.
Grund, C., W. Thomas, and D. Helbing. 2013. "How Natural Selection Can Create Both Selfand Other-Regarding Preferences, and Networked Minds". *Scientific Reports*. 3: 1480.

Grusec, J. E. and L. Kuczynski. 1997. *Parenting and Children's Internalization of Values: A Handbook of Contemporary Theory.* New York: John Wiley & Sons.

Güth, W., R. Schmittberger, and B. Schwarze. 1982. "An Experimental Analysis of Ultimatum Bargaining". *Journal of Economic Behavior and Organization.* 3: 367–388.

Haidt, J. 2001. "The Emotional Dog and its Rational Tail: A Social Intuitionist Approach to Moral Judgment". *Psychological Review.* 108: 814–834.

Haidt, J. 2012. *The Righteous Mind: Why Good People Are Divided by Politics and Religion.* New York: Pantheon.

Hechter, M. and S. Kanazawa. 1997. "Sociological Rational Choice". *Annual Review of Sociology.* 23: 199–214.

Hedström, P. and P. Bearman. 2009. *The Oxford Handbook of Analytical Sociolog.* Oxford: Oxford University Press.

Helbing, D. 1995. *Quantitative Sociodynamics.* Dordrecht: Kluwer Academic.

Helbing, D. 1996. "A Stochastic Behavioral Model and a 'Microscopic' Foundation of Evolutionary Game Theory". *Theory and Decision.* 40: 149–179.

Helbing, D. 2010. *Quantitative Sociodynamics, 2nd Edition.* Berlin: Springer-Verlag.

Helbing, D. 2012. *Social Self-organization: Agent-Based Simulations and Experiments to Study Emergent Social Behavior.* Springer.

Helbing, D., M. Schönhof, H.-U. Stark, and J. A. Holyst. 2005. "How Individuals Learn to Take Turns: Emergence of Alternating Cooperation in a Congestion Game and the Prisoner's Dilemma". *Advances in Complex Systems.* 8: 87–116.

Helbing, D., A. Szolnoki, M. Perc, and G. Szabó. 2010. "Evolutionary Establishment of Moral and Double Moral Standards through Spatial Interactions". *PLoS Computational Biology.* 6: e10000758.

Henrich, J. and F. Gil-White. 2001. "The Evolution of Prestige: Freely Conferred Status as a Mechanism for Enhancing the Benefits of Cultural Transmission". *Evolution and Human Behavior.* 22: 165–196.

Holden, C. J. 2002. "Bantu Language Trees Reflect the Spread of Farming across Sub-Saharan Africa: A Maximum-parsimony Analysis". *Proceedings of the Royal Society of London.* B 269: 793–799.

Holden, C. J. and R. Mace. 2003. "Spread of Cattle Led to the Loss of Matrilineal Descent in Africa: A Coevolutionary Analysis". *Proceedings of the Royal Society of London.* B 270: 2425–2433.

Huxley, J. S. 1955. "Evolution, Cultural and Biological". *Yearbook of Anthropology*: 2–25.

Ihara, Y. 2011. "Evolution of Culture-dependent Discriminate Sociality: a GeneCulture Coevolutionary Model". *Proceedings of the Royal Society of London.* B 366.

Jablonka, E. and M. J. Lamb. 1995. *Epigenetic Inheritance and Evolution: The Lamarckian Case*. Oxford: Oxford University Press.

James, W. 1880. "Great Men, Great Thoughts, and the Environment". *Atlantic Monthly*. 46: 441–459.

Jurmain, R., H. Nelson, L. Kilgore, and W. Travathan. 1997. *Introduction to Physical Anthropology*. Cincinatti: Wadsworth Publishing Company.

Kandel, D. 1978. "Homophily, Selection and Socialization in Adolescent Friendships". *American Journal of Sociology*. 84(2): 427–436.

Kandori, M. G., G. Mailath, and R. Rob. 1993. "Learning, Mutation, and Long Run Equilibria in Games". *Econometrica*. 61: 29–56.

Kirby, K. N. and R. J. Herrnstein. 1995. "Preference Reversals Due to Myopic Discounting of Delayed Reward". *Psychological Science*. 6(2): 83–89.

Knill, D. and A. Pouget. 2004. "The Bayesian Brain: the Role of Uncertainty in Neural Coding and Computation". *Trends in Cognitive Psychology*. 27(12): 712–719.

Krebs, J. R. and N. B. Davies. 1997. *Behavioral Ecology: An Evolutionary Approach. Fourth edition*. Oxford: Blackwell Science.

Kuhn, T. 1962. *The Nature of Scientific Revolutions*. Chicago: University of Chicago Press.

Laland, K. N., F. J. Odling-Smee, and M. W. Feldman. 2000. "Group Selection: A Niche Construction Perspective". *Journal of Consciousness Studies*. 7(1/2): 221–224.

Lewis, D. 1969. *Conventions: A Philosophical Study*. Cambridge, MA: Harvard University Press.

Lewontin, R. C. 1981. "Sleight of Hand". *The Sciences*: 23–26.

Lichtenstein, S. and P. Slovic. 1971. "Reversals of Preferences between Bids and Choices in Gambling Decisions". *Journal of Experimental Psychology*. 89: 46–55.

Lindenberg, S. 1983. "Utility and Morality". *Kyklos*. 36: 450–468.

Lindenberg, S. 2004. "Social Rationality". In: *Encyclopedia of Social Theory*. Ed. by G. Ritzer. Vol. II. Thousand Oaks: Sage.

Linton, R. 1936. *The Study of Man*. New York: Appleton-Century-Crofts.

Luce, R. D. 2005. *Individual Choice Behavior*. New York: Dover.

Luce, R. D. and P. Suppes. 1965. "Preference, Utility, and Subjective Probability". In: *Handbook of Mathematical Psychology*. Ed. by R. D. Luce, R. R. Bush, and E. Galanter. Vol. III. New York: Wiley.

Lumsden, C. J. and E. O. Wilson. 1981. *Genes, Mind, and Culture: The Coevolutionary Process*. Cambridge, MA: Harvard University Press.

Mace, R. and M. Pagel. 1994. "The Comparative Method in Anthropology". *Current Anthropology*. 35: 549–564.

Machina, M. J. 1989. "Dynamic Consistency and Non-Expected Utility Models of Choice under Uncertainty". *Journal of Economic Literature*. 27: 1622–1668.

Masclet, D., C. Noussair, S. Tucker, and M.-C. Villeval. 2003. "Monetary and Nonmonetary Punishment in the Voluntary Contributions Mechanism". *American Economic Review*. 93(1): 366–380.

Mas-Colell, A., M. D. Whinston, and J. R. Green. 1995. *Microeconomic Theory*. New York: Oxford University Press.

Maynard, S. J. 1982. *Evolution and the Theory of Games*. Cambridge: Cambridge University Press.

Maynard, S. J. and N. Warren. 1982. "Models of Cultural and Genetic Change". *Evolution*. 36: 620–627.

McFadden, D. 1973. "Conditional Logit Analysis of Qualitative Choice Behavior". In: *Frontiers in Econometrics*. Ed. by P. Zarembka. New York: Academic Press. 105–142.

McPherson, M., L. Smith-Lovin, and J. Cook. 2001. "Birds of a Feather: Homophily in Social Networks". *Annual Review of Sociology*. 27: 415–444.

Mead, G. H. 1934. *Mind, Self, and Society*. Chicago: University of Chicago Press.

Mednick, S. A., L. Kirkegaard-Sorenson, B. Hutchings, J. Knop, R. Rosenberg, and F. Schulsinger. 1977. "An Example of Bio-social Interaction Research: The Interplay of Socio-environmental and Individual Factors in the Etiology of Criminal Behavior". In: *Biosocial Bases of Criminal Behavior*. Ed. by S. A. Mednick and K. O. Christiansen. New York: Gardner Press. 9–24.

Mesoudi, A., A. Whiten, and K. N. Laland. 2006. "Towards a Unified Science of Cultural Evolution". *Behavioral and Brain Sciences*. 29: 329–383.

Miller, B. L., A. Darby, D. F. Benson, J. L. Cummings, and M. H. Miller. 1997. "Aggressive, Socially Disruptive and Antisocial Behaviour Associated with Fronto-temporal Dementia". *British Journal of Psychiatry*. 170: 150–154.

Miller, J. H. and S. E. Page. 2007. *Complex Adaptive Systems: An Introduction to Computational Models of Social Life*. Princeton: Princeton University Press.

Moll, J., R. Zahn, R. di Oliveira-Souza, F. Krueger, and J. Grafman. 2005. "The Neural Basis of Human Moral Cognition". *Nature Neuroscience*. 6: 799–809.

Morowitz, H. 2002. *The Emergence of Everything: How the World Became Complex*. Oxford: Oxford University Press.

Nash, J. F. 1950. "Equilibrium Points in n-Person Games". *Proceedings of the National Academy of Sciences*. 36: 48–49.

Newman, M., A.-L. Barabasi, and D. J. Watts. 2006. *The Structure and Dynamics of Networks*. Princeton: Princeton University Press.

Nisbett, R. E. and D. Cohen. 1996. *Culture of Honor: The Psychology of Violence in the South. Boulder*. CO: Westview Press.

O'Brien, M. J. and R. L. Lyman. 2000. *Applying Evolutionary Archaeology*. New York: Kluwer Academic.

Odling-Smee, F. J., K. N. Laland, and M. W. Feldman. 2003. *Niche Construction: The Neglected Process in Evolution.* Princeton: Princeton University Press.

Ok, E. A. and Y. Masatlioglu. 2003. "A General Theory of Time Preference". Economics Department, New York University.

Parsons, T. 1937. *The Structure of Social Action.* New York: McGraw-Hill.

Parsons, T. 1964. "Evolutionary Universals in Society". *American Sociological Review.* 29: 339–357.

Parsons, T. 1967. *Sociological Theory and Modern Society.* New York: Free Press.

Parsons, T. and E. Shils. 1951. *Toward a General Theory of Action.* Cambridge, MA: Harvard University Press.

Pearl, J. 2000. *Causality.* New York: Oxford University Press.

Pinker, S. 2002. *The Blank Slate: The Modern Denial of Human Nature.* New York: Viking.

Popper, K. 1979. *Objective Knowledge: An Evolutionary Approach.* Oxford: Clarendon Press.

Rauch, J. E. June 1996. "Trade and Networks: An Application to Minority Retail Entrepreneurship". Russell Sage Working Paper.

Relethford, J. H. 2007. *The Human Species: An Introduction to Biological Anthropology.* New York: McGraw-Hill.

Rivera, M. C. and J. A. Lake. 2004. "The Ring of Life Provides Evidence for a Genome Fusion Origin of Eukaryotes". *Nature.* 431: 152–155.

Roemer, J. 1982. *A General Theory of Exploitation and Class.* Cambridge: Harvard University Press.

Rozin, P., L. Lowery, S. Imada, and J. Haidt. 1999. "The CAD Triad Hypothesis: A Mapping between Three Moral Emotions (Contempt, Anger, Disgust) and Three Moral Codes (Community, Autonomy, Divinity)". *Journal of Personality & Social Psychology.* 76: 574–586.

Samuelson, P. 1947. *The Foundations of Economic Analysis.* Cambridge: Harvard University Press.

Savage, L. J. 1954. *The Foundations of Statistics.* New York: John Wiley & Sons.

Schelling, T. C. 1960. *The Strategy of Conflict.* Cambridge, MA: Harvard University Press.

Schultz, L. and A. Gopnik. 2004. "Causal Learning across Domains". *Developmental Psychology.* 40: 162–176.

Seeley, T. D. 1997. "Honey Bee Colonies are Group-Level Adaptive Units". *American Naturalist.* 150: S22–S41.

Shennan, S. 1997. *Quantifying Archaeology.* Edinburgh: Edinburgh University Press.

Simon, H. 1990. "A Mechanism for Social Selection and Successful Altruism". *Science.* 250: 1665–1668.

Skibo, J. M. and R. A. Bentley. 2003. *Complex Systems and Archaeology.* Salt Lake City: University of Utah Press.

Sobel, D. M. and N. Z. Kirkham. 2007. "Bayes Nets and Babies: Infants' Developing Statistical Reasoning Abilities and their Representations of Causal Knowledge". *Developmental Science.* 10(3): 298–306.

Sopher, B. and G. Gigliotti. 1993. "Intransitive Cycles: Rational Choice or Random Error: An Answer Based on Estimation of Error Rates with Experimental Data". *Theory and Decision.* 35: 311–336.

Spirtes, P., C. Glymour, and R. Scheines. 2001. *Causation, Prediction, and Search.* Cambridge: MIT Press.

Stalnaker, R. 1968. "A Theory of Conditionals". In: *Studies in Logical Theory.* Ed. by N. Rescher. London: Blackwell.

Starmer, C. 2000. "Developments in Non-Expected Utility Theory: The Hunt for a Descriptive Theory of Choice under Risk". *Journal of Economic Literature.* 38: 332–382.

Sterelny, K. 2011. "From Hominins to Humans: How *Sapiens* Became Behaviourally Modern". *Proceedings of the Royal Society of London B.*

Sugden, R. 1989. "Spontaneous Order". *Journal of Economic Perspectives.* 3(4): 85–97.

Sugden, R. 1986. *The Economics of Rights, Co-operation and Welfare.* Oxford: Basil Blackwell.

Taylor, M. 1976. *Anarchy and Cooperation.* London: John Wiley & Sons.

Taylor, M. 1982. *Community, Anarchy, and Liberty.* Cambridge: Cambridge University Press.

Taylor, M. 1987. *The Possibility of Cooperation.* Cambridge: Cambridge University Press.

Tomasello, M., M. Carpenter, J. Call, T. Behne, and H. Moll. 2005. "Understanding and Sharing Intentions: The Origins of Cultural Cognition". *Behavioral and Brain Sciences.* 28(5): 675–691.

Tooby, J. and L. Cosmides. 1992. "The Psychological Foundations of Culture". In: *The Adapted Mind: Evolutionary Psychology and the Generation of Culture.* Ed. by J. H. Barkow, L. Cosmides, and J. Tooby. New York: Oxford University Press.

Trivers, R. L. 1971. "The Evolution of Reciprocal Altruism". *Quarterly Review of Biology.* 46: 35–57.

Turner, J. H. 2006. *Handbook of Sociological Theory.* Berlin: Springer.

Tversky Amos, P. S. and D. Kahneman. 1990. "The Causes of Preference Reversal". *American Economic Review.* 80(1): 204–217.

Ullman-Margalit, E. 1977. *The Emergence of Norms.* Oxford: Clarendon Press.

Walras, L. 1874. *Elements of Pure Economics.* London: George Allen and Unwin.

Weibull, J. W. 1995. *Evolutionary Game Theory.* Cambridge, MA: MIT Press.

Wilson, E. O. 1975. *Sociobiology: The New Synthesis*. Cambridge, MA: Harvard University Press.
Wilson, E. O. 2012. *The Social Conquest of Earth*. New York: W. W. Norton.
Wilson, E. O. and B. Hölldobler. 2005. "Eusociality: Origin and Consequences". *Proceedings of the National Academy of Sciences*. 102(38): 13367–71.
Winter, F., H. Rauhut, and D. Helbing. 2012. "How Norms Can Generate Conflict: An Experiment on the Failure of Cooperative Micro-motives on the Macro-level". *Social Forces*. 90(3): 919–948.
Winterhalder, B. and E. A. Smith. 1992. *Evolutionary Ecology and Human Behavior*. New York: Aldine de Gruyter.
Wrong, D. H. 1961. "The Oversocialized Conception of Man in Modern Sociology". *American Sociological Review*. 26: 183–193.
Young, H. P. 1993. "The Evolution of Conventions". *Econometrica*. 61(1): 57–84.
Young, H. P. 1998. *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions*. Princeton: Princeton University Press.
Zajonc, R. B. 1980. "Feeling and Thinking: Preferences Need No Inferences". *American Psychologist*. 35(2): 151–175.
Zajonc, R. B. 1984. "On the Primacy of Affect". *American Psychologist*. 39: 117–123.