PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

Phil. Trans. R. Soc. B (2011) **366,** 878–888
doi:10.1098/rstb.2010.0310

Review

Gene–culture coevolution and the nature of human sociality

Herbert Gintis[1,2,*]

[1]*Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA*
[2]*Central European University, Nador u. 9, 1051 Budapest, Hungary*

Human characteristics are the product of gene–culture coevolution, which is an evolutionary dynamic involving the interaction of genes and culture over long time periods. Gene–culture coevolution is a special case of niche construction. Gene–culture coevolution is responsible for human other-regarding preferences, a taste for fairness, the capacity to empathize and salience of morality and character virtues.

Keywords: gene–culture coevolution; sociobiology; epistatic information transfer

1. GENE–CULTURE COEVOLUTION

Because of the importance of culture and complex social organization to the evolutionary success of *Homo sapiens*, 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* [1–4]. This coevolutionary process has endowed us with preferences that go beyond the self-regarding concerns emphasized in traditional economic and biological theory, and 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.

Gene–culture coevolution is the application of *sociobiology*, 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 behaviour [5].

The 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 [6], but *cultural transmission* in humans and to a lesser extent in other animals [7,8] 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 & Feldman [9], prestige (higher influencing lower status), as in Henrich & Gil-White [10], popularity-related as in Newman *et al.* [11] and even random population-dynamic transmission, as in Shennan [12] and Skibo & Bentley [13].

The parallel between cultural and biological evolution goes back to Huxley [14], Popper [15] and James [16]—see Mesoudi *et al.* [17] for details. The idea of treating culture as a form of epigenetic

*hgintis@comcast.net

One contribution of 13 to a Theme Issue 'Human niche construction'.

transmission was pioneered by Dawkins [18], 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 [2,20]. 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 [21–23].

Because of their common informational and evolutionary character, there are strong parallels between models of genetic and cultural evolution [17]. 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 [6,24,25]. Moreover, anthropologists reconstruct the history of social groups by analysing homologous and analogous cultural traits, much as biologists reconstruct the evolution of species by the analysis of shared characters and homologous DNA [26]. Indeed, the same computer programs developed by biological systematists are used by cultural anthropologists [27,28]. In addition, archeologists who study cultural evolution have a similar *modus operandi* as palaeobiologists who study genetic evolution [17]. Both attempt to reconstruct lineages of artifacts and their carriers. Like palaeobiology, archaeology assumes that when analogy can be ruled out, similarity implies causal connection by inheritance [29]. Like biogeography's study of the spatial distribution of organisms [30], behavioural ecology studies the interaction of ecological, historical and geographical factors that determine distribution of cultural forms across space and time [31].

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

Dawkins [32] 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 [5]. 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 the high degree of relatedness fostered by haplodiploidy, 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 [33–35]. The social structure of the hive 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 [36]. 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 defence [4,37]. While some contemporary biological and economic theorists have attempted to show that such cooperation can be supported by self-regarding rational agents [38–40], the conditions under which their models work are implausible even for small groups [41,42]. 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 [43].

Neuroscientific studies exhibit clearly the genetic basis for moral behaviour. Brain regions involved in moral judgements and behaviour include the prefrontal cortex, the orbitalfrontal cortex and the superior temporal sulcus [44]. These brain structures are virtually unique to or most highly developed in humans and are doubtless evolutionary adaptations [45]. The evolution of the human prefrontal cortex is closely tied to the emergence of human morality [46]. Patients with focal damage to one or more of these areas exhibit a variety of antisocial behaviours, including the absence of embarrassment, pride and regret [47,48], and sociopathic behaviour [49]. 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 [50].

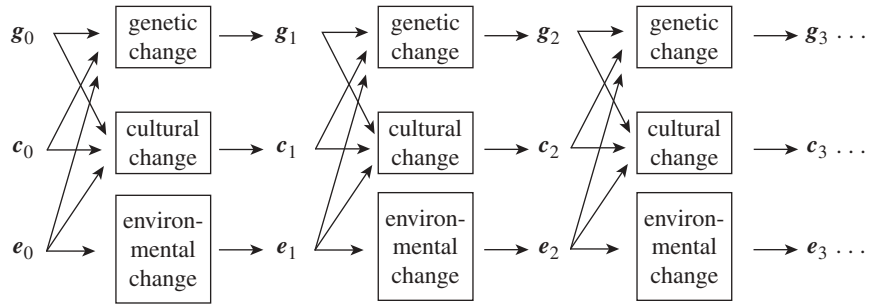

Figure 1. The dynamics of gene–culture coevolution.

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.

2. CULTURE IS NOT A BY-PRODUCT OF GENETIC EVOLUTION

It might be thought that the complex and intimate interaction of genes and culture outlined above is overdrawn, and that human genetic evolution is the effect of genetic inclusive fitness maximization, culture being an effect of genes that can be factored out in the long run. For instance, the eminent evolutionary psychologist David Buss holds that 'Culture is not an autonomous casual process in competition with biology for explanatory power' [51, p. 407]. This denial of gene–culture coevolution can be shown to be *prima facie* untenable. To see this, suppose we have a vector $\boldsymbol{g}$ of genetic variables, a vector $\boldsymbol{c}$ of cultural variables, and a vector $\boldsymbol{e}$ of environmental variables, including the prevalence of predators and prey, weather and the like. In an evolutionary model, the rate of change of variables is a function of the variables, so we have

$$\dot{\boldsymbol{g}} = F(\boldsymbol{g}, \boldsymbol{c}, \boldsymbol{e}), \tag{2.1}$$

$$\dot{\boldsymbol{c}} = G(\boldsymbol{g}, \boldsymbol{c}, \boldsymbol{e}) \tag{2.2}$$

$$\dot{\boldsymbol{e}} = H(\boldsymbol{e}). \tag{2.3}$$

Note that it is plausible for $\boldsymbol{c}$ to affect the nature and pace of environmental change, in which case it should be included in equation (2.3). We abstract from this causal path in order to strengthen the case for Buss' argument. The contention that culture is an effect of genetic fitness maximization in this framework is the assertion that $\boldsymbol{c}$ can be eliminated from these equations. Under what conditions can this occur? Taking the derivative of equation (2.1), and substituting equations (2.2) and (2.3) into equation (2.1), we get

$$\begin{aligned}\ddot{\boldsymbol{g}} = {} & F_g(\boldsymbol{g}, \boldsymbol{c}, \boldsymbol{e})F(\boldsymbol{g}, \boldsymbol{c}, \boldsymbol{e}) + F_c(\boldsymbol{g}, \boldsymbol{c}, \boldsymbol{e})G(\boldsymbol{g}, \boldsymbol{c}, \boldsymbol{e}) \\ & + F_e(\boldsymbol{g}, \boldsymbol{c}, \boldsymbol{e})H(\boldsymbol{e}).\end{aligned} \tag{2.4}$$

If $\boldsymbol{c}$ is to be absent from this second order differential equation, the derivative of the right-hand side of equation (2.4) with respect to $\boldsymbol{c}$ must be identically zero. Thus, we have

$$0 \equiv F_{\boldsymbol{gc}}F + F_{\boldsymbol{g}}F_{\boldsymbol{c}} + F_{\boldsymbol{cc}}G + F_{\boldsymbol{c}}G_{\boldsymbol{c}} + F_{\boldsymbol{ec}}H. \tag{2.5}$$

All five of the above terms must then be identically zero, so $F_{\boldsymbol{c}} \equiv 0$, implying that $\boldsymbol{c}$ does not enter on the right-hand side of the defining equations (2.1)–(2.3); i.e. genes are not a function of culture. This is obviously not appropriate for humans, since both genes and culture are functions of culture.

Figure 1 illustrates this dynamical process. Note that as long as there is high fidelity cultural transmission over multiple generations (signified by the middle row of horizontal arrows), genetic and cultural evolution are inextricably intertwined. By contrast, for species that do not have cumulative learning, these arrows are absent, and despite the fact that genes affect culture in every period, there is no cumulative interrelatedness of genes and culture.

We will give two examples of understanding human evolution using gene–culture evolution, the repositioning of the larynx and other physiological changes facilitating linguistic communication [52], and the role of culture in creating a genetic predisposition for cooperative activity in humans [53].

3. GENE–CULTURE COEVOLUTION AND THE PHYSIOLOGY OF COMMUNICATION

The evolution of the physiology of speech and facial communication is a dramatic 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 [54]. 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 [55,56].

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 [57]. 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, 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 in 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 [58]. 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 [59]. In mammals, this mimetic musculature attaches to the skin of the face, thus permitting the facial communication of such emotions as fear, surprise, disgust and anger. In most mammals, however, a few wide sheet-like 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. SOCIALIZATION AND THE INTERNALIZATION OF NORMS

Human 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* [60–63], 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 to the duties and obligations of the role-positions they expect to occupy. The internalization of norms, of course, presupposes a genetic predisposition to moral cognition that can be explained only by gene–culture coevolution.

The human openness to socialization is perhaps the most powerful form of epigenetic transmission found in nature. This epigenetic flexibility in considerable part accounts for the stunning success of the species *Homo sapiens*, because when individuals internalize a norm, the frequency of the desired behaviour will be higher than if people follow the norm only instrumentally—i.e. when they perceive it to be in their interest to do so on other grounds. The increased incidence of prosocial behaviours is precisely what permits humans to cooperate effectively in groups [64].

There are, of course, limits to socialization [65,66], 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 [53,67]. Moreover, there are often swift society-wide value changes that cannot be accounted for by socialization theory [68,69]. However, socialization theory has an important place in the general theory of culture, strategic learning and moral development.

The susceptibility to socialization is controlled by neuronal structures and is hence the product of genetic evolution. The socialized individual is highly sensitized to the particular rewards offered for prosocial behaviour and penalties imposed for antisocial behaviour. But this sensitivity is characteristic of all creatures who live in social settings. The distinguishing characteristic of the internalization of norms is that individuals behave prosocially even when there is no possibility of being rewarded for prosocial or penalized for antisocial behaviour. Such altruistic behaviour has been confirmed in scores of laboratory and field studies across a wide variety of societies [42,70,71].

Gintis [53] provides a plausible evolutionary scenario in which the genetic predisposition to internalize norms may have developed. The prerequisite is a cultural system sufficiently complex that the learning process for youth in acquiring facility with this system extends throughout childhood, and hence takes the form of an authoritarian imposition carried out by elders. Because the skills acquired in this manner (e.g. hunting, recognizing and preparing nutritious foodstuffs) do not have immediate intrinsic payoffs for the learner, those who respond to the rewards and sanctions of teachers will reproduce at the expense of those who do not. Internalizing the norms associated with instrumental skills will then be directly fitness-enhancing, and hence the neural structures that support internalization will be privileged in human evolution. Once these neural structures are in place, they can be deployed for more general purposes, including internalizing moral values and deriving pleasure from helping others and punishing those who act contrary to social norms.

5. ALTRUISM IS AN EMERGENT PROPERTY OF HUMAN GENE–CULTURE EVOLUTION

Many empirical findings from behavioural game theory [70] show that human subjects regularly exhibit altruistic behaviours towards enhancing cooperative payoffs [42, ch. 4]. Indeed, it is likely that such altruistic predispositions account for the remarkable evolutionary success of our species [64]. Among such predispositions are the character virtues (honesty, courage, trustworthiness, considerateness and the like) and strong reciprocity, which is a predisposition to cooperate with others in a collective task, and to

punish those who fail to cooperate, even when a self-regarding individual would simply free-ride on the effort of others [72]. These behaviours are *altruistic* in the sense that they enhance the payoffs to other group members at a cost to the cooperator. Of course, in a framework of biological evolution, where payoffs are fitnesses, altruism could not evolve unless the fitness cost of altruism were somehow recouped in the long run. Some question calling the behaviour altruistic in this case, but requiring altruism to be fitness-reducing in the whole population in the long run amounts to excluding the possibility of altruism by definition [73]. I shall stick to a more fruitful definition of altruism, according to which the behaviour of an individual is altruistic if it benefits other members of the group and the individual would increase his own payoff by switching to another behaviour [74].

Those who deny the causal importance of culture in human evolution generally argue that the altruistic behaviour exhibited in modern societies is simply a maladaptation to current environmental conditions. Thus, altruistic cooperation and punishment, they argue, stem from mental confusion due to the difficulty in avoiding detection when behaving anti-socially in our evolutionary past. Throughout most of the history of our species, they argue, hunter–gatherer societies offered little room for the sorts of anonymous interaction and covert behaviour found in modern society [75,76]. Because of our evolutionary past, they argue, modern humans are hyper-sensitive to even remote possibilities that their actions may be observed and their reputations sullied. For this reason, those who hold the maladaptation position suggest that modern humans tend to behave as though every social interaction were publicly observable.

However, our understanding of contemporary hunter–gatherer societies indicates the lack of credibility of this argument, as do analyses of trade and migration patterns in Pleistocene [77,78].

If the altruism-as-maladaptation view were correct, we should expect similar behaviour from our closest primate relatives. Many non-human primates live in hunter–gatherer type groups and there is constant migration among groups for reasons of exogamous mating. Primates are also able to distinguish kin from non-kin, and engage in repeated interactions, much as humans do. Nevertheless, there is no evidence of behaviors akin to altruistic cooperation and punishment in any primate group [79,80]. Even in such 'unnatural' situations as living in large groups in zoos and protective environments, such primates do not exhibit the 'confusions' that the mistake hypothesis attributes to humans.

The maladaptation explanation of altruistic cooperation suggests that humans find it difficult to distinguish between one-shot and repeated interactions because humans experienced only repeated interaction prior to the appearance of settled communities some 10 000 years before the present. However, humans are perfectly capable of distinguishing short- from long-term interactions, and they cooperate much more in the latter case than in the former [81–83].

Moreover, if altruism results from confusion, we would not expect individuals to adjust their level of altruistic contribution rationally according to the costs and benefits. The evidence is that they do. Preferences for altruistic acts entail transitive preferences as required by the notion of rationality in decision theory [84]. In the Dictator Game, the experimenter gives a subject, called the dictator, a certain amount of money and instructs him to give any portion of it he desires to a second, anonymous, subject, called the receiver. The dictator keeps whatever he does not choose to give to the receiver. Obviously, a self-regarding dictator will give nothing to the receiver. Suppose the experimenter gives the dictator m points (exchangeable at the end of the session for real money) and tells him that the price of giving some of these points to the receiver is p, meaning that each point the receiver gets costs the giver p points. For instance, if $p = 4$, then it costs the dictator 4 points for each point that he transfers to the receiver. The dictator's choices must then satisfy the budget constraint $\pi_s + p\pi_o = m$, where π_s is the amount the dictator keeps and π_o is the amount the receiver gets. The question, then, is simply, is there a preference function $u(\pi_s, \pi_o)$ that the dictator maximizes subject to the budget constraint $\pi_s + p\pi_o = m$? If so, then it is just as rational, from a behavioural standpoint, to care about giving to the receiver as to care about consuming marketed commodities.

Varian [85] developed a generalized axiom of revealed preference (GARP) that ensures that individuals are rational as in the sense of traditional consumer demand theory. Andreoni & Miller [84] worked with 176 students in an elementary economics class and had them play the Dictator Game multiple times each, with the price p taking on the values $p = 0.25$, 0.33, 0.5, 1, 2, 3 and 4, with amounts of tokens equaling $m = 40$, 60, 75, 80 and 100. They found that only 18 of the 176 subjects violated GARP at least once and that of these violations, only four were at all significant. By contrast, if choices were randomly generated, we would expect that between 78 and 95 per cent of subjects would have violated GARP.

As to the degree of altruistic giving in this experiment, Andreoni and Miller found that 22.7 per cent of subjects were perfectly selfish, 14.2 per cent were perfectly egalitarian at all prices, and 6.2 per cent always allocated all the money so as to maximize the total amount won (i.e. when $p > 1$, they kept all the money, and when $p < 1$, they gave all the money to the receiver).

Fischbacher *et al.* [86] also found that subjects adjust their altruistic behaviour strategically when strategic parameters change. They staged an Ultimatum Game with one proposer and several responders, who had to respond simultaneously to the proposal.[2] If no responder accepted the offer, both responders and the proposer received zero payoff. If more than one responder accepted the offer, one of the accepting responders was chosen randomly to receive the proposed amount, and the proposer received the remainder, rejecting responders receiving zero. If rejecting a positive offer is simply muddle-headed or blindly emotional, this new setting should not change responder behaviour. If responders reject offers

with the goal of punishing proposers, then the frequency of rejection should be very low in the new situation, because no single responder could ensure that punishment would take place. If, however, there are adaptive reasons for altruistic punishment, then we would expect the punishing behaviour to occur only when it can be instrumentally effective, which it cannot in the multiple-recipient version of the Ultimatum Game. The experimenters found that with multiple responders, the rejection rate fell significantly in the two-responder case, and even more in the five-responder case. For instance, whereas in the traditional single-responder case, offers of 20 per cent of the pie are rejected with 80 per cent probability, such offers are rejected with 15 per cent probability when there are five responders. Moreover, the average share of the pie accruing to the responders falls from about 40 per cent with one responder to 20 per cent with two responders, and to about 15 per cent with five responders.

Many additional examples can be given suggesting that other-regarding and moral behaviour in humans is part of our adaptive repertoire, rather than being the results of misdirected attempts at maximizing long-term self-interest.

6. THE RATIONALITY OF ALTRUISTIC BEHAVIOUR: THEORY AND EXPERIMENTAL EVIDENCE

Morality is an emergent property of the gene–culture evolutionary dynamic that gave rise to our species. We can frame and test propositions concerning moral behaviour using the methods of game theory, involving subjects from a variety of social backgrounds and cultures. Moral behaviour is often held to be incompatible with rational choice. This is incorrect. The rational actor model of economic theory presupposes that people have consistent preferences, but does not require that preferences be self-regarding or materialistic. We can just as easily measure how much people value honesty or loyalty as we can chart how much they value fried chicken or cashmere sweaters.

Because the use of the word 'rational' in the rational actor model is so circumscribed compared with the general usage of the word, we often call the rational actor model the *beliefs, preferences and constraints* model (BPC), because this captures the notion of consistent preference, the centrality of beliefs and the notion of making trade-offs subject to informational and material constraints.

In the BPC model, choices give rise to probability distributions over outcomes, the expected values of which are the payoffs to the choice from which they arose. Game theory extends this analysis to cases where there are multiple decision makers. In the language of game theory, players are endowed with strategies, and have certain information, and for each array of choices by the players, the game specifies a distribution of payoffs to the players. Game theory predicts the behaviour of the players by assuming each is rational; in other words, each maximizes a preference function subject to beliefs as well as informational and material constraints.

The experiments described below are all based on using game theory to set up the choices available to subjects, the knowledge they have on which their choices are based and the payoffs to each subject as a function of their joint strategy choices. We assume the subjects are rational (i.e. consistent) decision-makers, so that their choices reflect their subjective trade-offs among heterogeneous payoffs—some material and some moral and/or other-regarding.

(a) *Conditional altruistic cooperation*

A *social dilemma* is a situation in which members of a group can gain by cooperating, but cooperation is costly, so each individual does better personally by not cooperating, no matter what the others do. For instance, suppose if a member of a group of size $n \geq 2$ pays the cost $c > 0$, he benefits the others by a total amount $b > c$. We then have a social dilemma: each member can enhance the net gain of the group by cooperating, but a selfish individual will not do so. If all cooperate, each will earn $b - c > 0$, but in a group of self-regarding individuals, each will earn zero.

Conditional altruistic cooperation is a predisposition to cooperate in a social dilemma as long as the other players also cooperate. Consider the above social dilemma, with $n = 2$, called the Prisoner's Dilemma. In this game, let *CC* stand for 'both players cooperate', let *DD* stand for 'both players defect', let *CD* stand for 'player 1 cooperates but his partner defects', and let *DC* stand for 'player 1 defects and his partner cooperates'. A self-regarding player 1 will prefer *DC* to *CC*, *CC* to *DD* and *DD* to *CD*, while an altruistic cooperator will prefer *CC* to *DC*, *DC* to *DD* and *DD* to *CD*; i.e. the self-regarding individual prefers to defect no matter what his partner does, whereas the conditional altruistic cooperator prefers to cooperate so long as his partner cooperates.

Kiyonari *et al.* [87] ran an experiment based on this game with real monetary payoffs using 149 Japanese university students. The experimenters ran three distinct treatments, with about equal numbers of subjects in each treatment. The first treatment was a standard 'simultaneous' Prisoner's Dilemma, the second was a 'second-player' situation in which the subject was told that the first player in the Prisoner's Dilemma had already chosen to cooperate, and the third was a 'first-player' treatment in which the subject was told that his decision to cooperate or defect would be made known to the second player before the latter made his own choice. The experimenters found that 38 per cent of the subjects cooperated in the simultaneous treatment, 62 per cent cooperated in the second-player treatment and 59 per cent cooperated in the first-player treatment. The decision to cooperate in each treatment cost the subject about $5 (600 yen). This shows unambiguously that a majority of subjects were conditional altruistic cooperators (62%). Almost as many were not only cooperators, but were also willing to bet that their partners would be (59%), provided the latter were assured of not being defected upon, although under standard conditions, without this assurance, only 38 per cent would in fact cooperate.

(b) *Altruism and cooperation in groups*

The *Public Goods Game*, an *n*-person social dilemma, captures many areas of altruistic cooperation in social life, including voluntary contribution to team and community goals. Researchers [88–91] uniformly find that groups exhibit a much higher rate of cooperation than can be expected assuming the standard model of the self-regarding actor.

A typical Public Goods Game consists of a number of rounds, say 10. In each round, each subject is grouped with several other subjects—say 3 others. Each subject is then given a certain number of points, say 20, redeemable at the end of the experimental session for real money. Each subject then places some fraction of his points in a 'common account' and the remainder in the subject's 'private account'. The experimenter then tells the subjects how many points were contributed to the common account and adds to the private account of *each* subject some fraction, say 40 per cent, of the total amount in the common account. So if a subject contributes his whole 20 points to the common account, each of the four group members will receive 8 points at the end of the round. In effect, by putting the whole endowment into the common account, a player loses 12 points but the other three group members gain in total 24 (8 times 3) points. The players keep whatever is in their private accounts at the end of the round.

A self-regarding player contributes nothing to the common account. However, most of the subjects do not in fact conform to the self-regarding model. Subjects begin by contributing on average about half of their endowments to the public account. The level of contributions decays over the course of the 10 rounds until in the final rounds most players are behaving in a self-regarding manner. This is, of course, exactly what is predicted by the strong reciprocity model. Because they are altruistic contributors, strong reciprocators start out by contributing to the common pool, but in response to the norm violation of the self-regarding types, they begin to refrain from contributing themselves.

How do we know that the decay of cooperation in the Public Goods Game is due to cooperators punishing free riders by refusing to contribute themselves? Subjects often report this behaviour retrospectively. More compelling, however, is the fact that when subjects are given a more constructive way of punishing defectors, they use it in a way that helps sustain cooperation [92–96].

Fehr & Gächter [82], for instance, used 6- and 10-round Public Goods Games with group sizes of 4, and with costly punishment allowed at the end of each round, employing three different methods of assigning members to groups. There were sufficient subjects to run between 10 and 18 groups simultaneously. Under the partner treatment, the four subjects remained in the same group for all 10 periods. Under the stranger treatment, the subjects were randomly reassigned after each round. Finally, under the perfect stranger treatment, the subjects were randomly reassigned but assured that they would never meet the same subject more than once.

Fehr & Gächter [82] performed their experiment for 10 rounds with punishment and 10 rounds without. Their results are illustrated in figure 2. We see that when costly punishment is permitted, cooperation does not deteriorate, and in the partner game, despite strict anonymity, cooperation increases almost to full cooperation even in the final round. When punishment is not permitted, however, the same subjects experienced the deterioration of cooperation found in previous Public Goods Games. The contrast in cooperation rates between the partner treatment and the two stranger treatments is worth noting because the strength of punishment is roughly the same across all treatments. This suggests that the credibility of the punishment threat is greater in the partner treatment because in this treatment the punished subjects are certain that, once they have been punished in previous rounds, the punishing subjects are in their group. The prosociality impact of strong reciprocity on cooperation is thus more strongly manifested the more coherent and permanent the group in question.

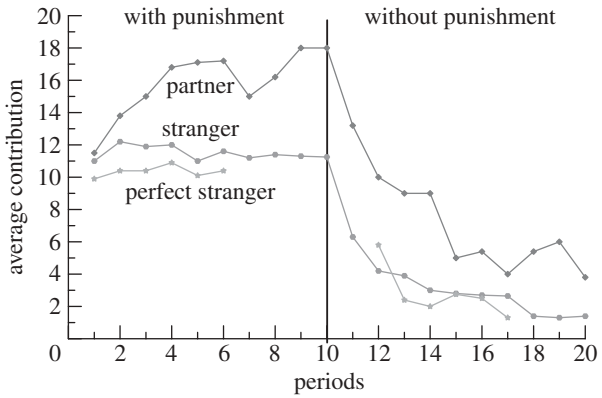

Figure 2. Average contributions over time in the partner, stranger and perfect stranger treatments when the punishment condition is played first [82].

(c) *Character virtues*

Character virtues are ethically desirable behavioural regularities that individuals value for their own sake, while having the property of facilitating cooperation and enhancing social efficiency. Character virtues include *honesty*, *loyalty*, *trustworthiness*, *promise-keeping* and *fairness*. Unlike such other-regarding preferences as strong reciprocity and empathy, these character virtues operate without concern for the individuals with whom one interacts. An individual is honest in his transactions because this is a desired state of being, not because he has any particular regard for those with whom he transacts. Of course, the sociopath *Homo economicus* is honest only when it serves his material interests to be so, whereas the rest of us are at times honest even when it is costly to be so and even when no one but us could possibly detect a breach.

Common sense, as well as the experiments described below, indicates that honesty, fairness and promise-keeping are not absolutes. If the cost of virtue is sufficiently high, and the probability of detection of a breach of virtue is sufficiently small, many individuals will behave dishonestly. When one is

aware that others are unvirtuous in a particular region of their lives (e.g. marriage, tax paying, obeying traffic rules, accepting bribes), one is more likely to allow one's own virtue to lapse. Finally, the more easily one can delude oneself into inaccurately classifying an unvirtuous act as virtuous, the more likely one is to allow oneself to carry out such an act.

One might be tempted to model honesty and other character virtues as *self-constituted constraints* on one's set of available actions in a game, but a more fruitful approach is to include the state of being virtuous in a certain way as an argument in one's preference function, to be traded off against other valuable objects of desire and personal goals. In this respect, character virtues are in the same category as ethical and religious preferences and are often considered subcategories of the latter.

Numerous experiments indicate that most subjects are willing to sacrifice material rewards to maintain a virtuous character even under conditions of anonymity. Sally [97] undertook a meta-analysis of 137 experimental treatments, finding that face-to-face communication, in which subjects are capable of making verbal agreements and promises, was the strongest predictor of cooperation. Of course, face-to-face interaction violates anonymity and has other effects besides the ability to make promises. However, both Bochet *et al.* [98] and Brosig *et al.* [99] report that only the ability to exchange verbal information accounts for the increased cooperation.

A particularly clear example of such behaviour is reported by Gneezy [100], who studied 450 undergraduate participants paired off to play three games of the following form, all payoffs to which were of the form (b,a), where player 1, Bob, receives b and player 2, Alice, receives a. In all games, Bob was shown two pairs of payoffs, $A:(x,y)$ and $B:(z,w)$ where x, y, z and w are amounts of money with $x<z$ and $y>w$, so in all cases B is better for Bob and A is better for Alice. Bob could then say to Alice, who could not see the amounts of money, either 'option A will earn you more money than option B', or 'option B will earn you more money than option A'. The first game was $A:(5,6)$ versus $B:(6,5)$ so Bob could gain 1 by lying and being believed while imposing a cost of 1 on Alice. The second game was $A:(5,15)$ versus $B:(6,5)$, so Bob could gain 1 by lying and being believed, while still imposing a cost of 10 on Alice. The third game was $A:(5,15)$ versus $B:(15,5)$, so Bob could gain 10 by lying and being believed, while imposing a cost of 10 on Alice.

Before starting play, Gneezy asked the various Bobs whether they expected their advice to be followed. He induced honest responses by promising to reward subjects whose guesses were correct. He found that 82 per cent of Bobs expected their advice to be followed (the actual number was 78%). It follows from the Bobs' expectations that if they were self-regarding, they would always lie and recommend B to Alice.

The experimenters found that, in game 2, where lying was very costly to Alice and the gain from lying was small for Bob, only 17 per cent of Bobs lied. In game 1, where the cost of lying to Alice was only 1 but the gain to Bob was the same as in game 2, 36 per cent of Bobs lied. In other words, Bobs were loath to lie but considerably more so when it was costly to Alices. In game 3, where the gain from lying was large for Bob and equal to the loss to Alice, fully 52 per cent of Bobs lied. This shows that many subjects are willing to sacrifice material gain to avoid lying in a one-shot anonymous interaction, their willingness to lie increasing with an increased cost to them of truth telling, and decreasing with an increased cost to their partners of being deceived. Similar results were found by Boles *et al.* [101] and Charness & Dufwenberg [102]. Gunnthorsdottir *et al.* [103] and Burks *et al.* [104] have shown that a socio-psychological measure of 'Machiavellianism' predicts which subjects are likely to be trustworthy and trusting.

7. CONCLUSION

Population biology traditionally takes the environment as exogenous. However, we know that life-forms affect their own environment and the environments they produce change the pattern of genetic evolution they undergo. Niche construction augments population biology by rendering environmental change itself part of the evolutionary dynamic. Gene–culture coevolution is the application of niche-construction reasoning to the human species, recognizing that both genes and culture are subject to similar dynamics, and human society is a cultural construction that provides the environment for fitness-enhancing genetic changes in individuals. The resulting social system is a complex dynamic nonlinear system. Such systems have *emergent properties*, some of which have been analysed in this paper: social norms, morality, other-regarding preferences and the internalization of norms.

I would like to thank the editors of this volume for suggesting improvements in my analysis, and the European Science Foundation and the Hungarian Scientific Research Fund (OTKA) for financial support.

ENDNOTES

[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 niche-constructive effect of modifying the selective environment for other genotypes (see Dawkins [19]).

[2]In the standard Ultimatum Game, the proposer is given some money (the 'pie'), and is instructed to offer any fraction he desires to an anonymous second player, the responder. If the responder accepts, they divide the money accordingly. If the responder rejects the offer, neither player receives anything.

REFERENCES

1 Boyd, R. & Richerson, P. J. 1985 *Culture and the evolutionary process*. Chicago, IL: University of Chicago Press.
2 Cavalli-Sforza, L. L. & Feldman, M. W. 1982 Theory and observation in cultural transmission. *Science* **218**, 19–27. (doi:10.1126/science.7123211)
3 Dunbar, R. M. 1993 Coevolution of neocortical size, group size and language in humans. *Behav. Brain Sci.* **16**, 681–735. (doi:10.1017/S0140525X00032325)
4 Richerson, P. J. & Boyd, R. 2004 *Not by genes alone*. Chicago, IL: University of Chicago Press.

5 Odling-Smee, F. J., Laland, K. N. & Feldman, M. W. 2003 *Niche construction: the neglected process in evolution*. Princeton, NJ: Princeton University Press.
6 Jablonka, E. & Lamb, M. J. 1995 *Epigenetic inheritance and evolution: the lamarckian case*. Oxford, UK: Oxford University Press.
7 Bonner, J. T. 1984 *The evolution of culture in animals*. Princeton, NJ: Princeton University Press.
8 Richerson, P. J. & Boyd, R. 1998 The evolution of ultrasociality. In *Indoctrinability, ideology and warfare* (eds I. Eibl-Eibesfeldt & F. K. Salter), pp. 71–96. New York, NY: Berghahn Books.
9 Cavalli-Sforza, L. L. & Feldman, M. W. 1981 *Cultural transmission and evolution*. Princeton, NJ: Princeton University Press.
10 Henrich, J. & Gil-White, F. 2001 The evolution of prestige: freely conferred status as a mechanism for enhancing the benefits of cultural transmission. *Evol. Hum. Behav.* **22**, 165–196. (doi:10.1016/S1090-5138(00)00071-4)
11 Newman, M., Barabasi, A.-L. & Watts, D. J. 2006 *The structure and dynamics of networks*. Princeton, NJ: Princeton University Press.
12 Shennan, S. 1997 *Quantifying archaeology*. Edinburgh: Edinburgh University Press.
13 Skibo, J. M. & Bentley, R. A. 2003 *Complex systems and archaeology*. Salt Lake City: University of Utah Press.
14 Huxley, J. S. 1955 Evolution, cultural and biological. *Yearbook of Anthropology* 2–25.
15 Popper, K. 1979 *Objective knowledge: an evolutionary approach*. Oxford, UK: Clarendon Press.
16 James, W. 1880 Great men, great thoughts, and the environment. *Atlantic Mon.* **46**, 441–459.
17 Mesoudi, A., Whiten, A. & Laland, K. N. 2006 Towards a unified science of cultural evolution. *Behav. Brain Sci.* **29**, 329–383. (doi:10.1017/S0140525X06009083)
18 Dawkins, R. 1976 *The selfish gene*. Oxford, UK: Oxford University Press.
19 Dawkins, R. 2004 Extended phenotype—but not *too* extended. A reply to Laland, Turner and Jablonka. *Biol. Phil.* **19**, 377–396. (doi:10.1023/B:BIPH.0000036180.14904.96)
20 Parsons, T. 1964 Evolutionary universals in society. *Am. Sociol. Rev.* **29**, 339–357. (doi:10.2307/2091479)
21 Ihara, Y. 2011 Evolution of culture-dependent discriminate sociality: a gene–culture coevolutionary model. *Phil Trans. R. Soc. B* **366**, 889–900. (doi:10.1098/rstb.2010.0247)
22 Zajonc, R. B. 1980 Feeling and thinking: preferences need no inferences. *Am. Psychol.* **35**, 151–175. (doi:10.1037/0003-066X.35.2.151)
23 Zajonc, R. B. 1984 On the primacy of affect. *Am. Psychol.* **39**, 117–123. (doi:10.1037/0003-066X.39.2.117)
24 Abbott, R. J., James, J. K., Milne, R. I. & Gillies, A. C. M. 2003 Plant introductions, hybridization and gene flow. *Phil. Trans. R. Soc. Lond. B* **358**, 1123–1132. (doi:10.1098/rstb.2003.1289)
25 Rivera, M. C. & Lake, J. A. 2004 The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature* **431**, 152–155. (doi:10.1038/nature02848)
26 Mace, R. & Pagel, M. 1994 The comparative method in anthropology. *Curr. Anthropol.* **35**, 549–564. (doi:10.1086/204317)
27 Holden, C. J. 2002 Bantu language trees reflect the spread of farming across sub-Saharan Africa: a maximum-parsimony analysis. *Proc. R. Soc. Lond. B* **269**, 793–799. (doi:10.1098/rspb.2002.1955)
28 Holden, C. J. & Mace, R. 2003 Spread of cattle led to the loss of matrilineal descent in Africa: a coevolutionary analysis. *Proc. R. Soc. Lond. B* **270**, 2425–2433. (doi:10.1098/rspb.2003.2535)
29 O'Brien, M. J. & Lyman, R. L. 2000 *Applying evolutionary archaeology*. New York, NY: Kluwer Academic.
30 Brown, J. H. & Lomolino, M. V. 1998 *Biogeography*. Sunderland, MA: Sinauer.
31 Smith, E. A. & Winterhalder, B. 1992 *Evolutionary ecology and human behavior*. New York, NY: Aldine de Gruyter.
32 Dawkins, R. 1982 *The extended phenotype: the gene as the unit of selection*. Oxford, UK: Freeman.
33 Gadagkar, R. 1991 On testing the role of genetic asymmetries created by haplodiploidy in the evolution of eusociality in the hymenoptera. *J. Genet.* **70**, 1–31. (doi:10.1007/BF02923575)
34 Seeley, T. D. 1997 Honey bee colonies are group-level adaptive units. *Am. Nat.* **150**, S22–S41. (doi:10.1086/286048)
35 Wilson, E. O. & Holldobler, B. 2005 Eusociality: origin and consequences. *Proc. Natl Acad. Sci. USA* **102**, 13 367–13 371. (doi:10.1073/pnas.0505858102)
36 Feldman, M. W. & Zhivotovsky, L. A. 1992 Gene–culture coevolution: toward a general theory of vertical transmission. *Proc. Natl Acad. Sci. USA* **89**, 11 935–11 938. (doi:10.1073/pnas.89.24.11935)
37 Boehm, C. 2000 *Hierarchy in the forest: the evolution of egalitarian behavior*. Cambridge, MA: Harvard University Press.
38 Alexander, R. D. 1987 *The biology of moral systems*. New York, NY: Aldine.
39 Fudenberg, D., Levine, D. K. & Maskin, E. 1994 The folk theorem with imperfect public information. *Econometrica* **62**, 997–1039. (doi:10.2307/2951505)
40 Trivers, R. L. 1971 The evolution of reciprocal altruism. *Q. Rev. Biol.* **46**, 35–57. (doi:10.1086/406755)
41 Boyd, R. & Richerson, P. J. 1988 The evolution of reciprocity in sizable groups. *J. Theor. Biol.* **132**, 337–356. (doi:10.1016/S0022-5193(88)80219-4)
42 Gintis, H. 2009 *The bounds of reason: game theory and the unification of the behavioral sciences*. Princeton, NJ: Princeton University Press.
43 Sterelny, K. 2011 From hominins to humans: how *sapiens* became behaviourally modern. *Phil. Trans. R. Soc. B* **366**, 809–822. (doi:10.1098/rstb.2010.0301)
44 Moll, J., Zahn, R., di Oliveira-Souza, R., Krueger, F. & Grafman, J. 2005 The neural basis of human moral cognition. *Nat. Neurosci.* **6**, 799–809. (doi:10.1038/nrn1768)
45 Schulkin, J. 2000 *Roots of social sensitivity and neural function*. Cambridge, MA: MIT Press.
46 Allman, J., Hakeem, A. & Watson, K. 2002 Two phylogenetic specializations in the human brain. *Neuroscientist* **8**, 335–346. (doi:10.1177/107385840200800409)
47 Beer, J. S., Heerey, E. A., Keltner, D., Skabini, D. & Knight, R. T. 2003 The regulatory function of self-conscious emotion: insights from patients with orbitofrontal damage. *J. Pers. Soc. Psychol.* **65**, 594–604. (doi:10.1037/0022-3514.85.4.594)
48 Camille, N. 2004 The involvement of the orbitofrontal cortex in the experience of regret. *Science* **304**, 1167–1170. (doi:10.1126/science.1094550)
49 Miller, B. L., Darby, A., Benson, D. F., Cummings, J. L. & Miller, M. H. 1997 Aggressive, socially disruptive and antisocial behaviour associated with fronto-temporal dementia. *Br. J. Psychiatry* **170**, 150–154. (doi:10.1192/bjp.170.2.150)

50 Mednick, S. A., Kirkegaard-Sorenson, L., Hutchings, B., Knop, J., Rosenberg, R. & Schulsinger, F. 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* (eds S. A. Mednick & K. O. Christiansen), pp. 9–24. New York, NY: Gardner Press.

51 Buss, D. M. 1999 *Evolutionary psychology: the new science of the mind*. Boston, MA: Allyn Bacon.

52 Deacon, T. W. 1998 *Symbolic species: the co-evolution of language and the brain*. New York, NY: Norton.

53 Gintis, H. 2003 The Hitchhiker's guide to altruism: genes, culture, and the internalization of norms. *J. Theor. Biol.* **220**, 407–418. (doi:10.1006/jtbi.2003.3104)

54 Jurmain, R., Nelson, H., Kilgore, L. & Travathan, W. 1997 *Introduction to physical anthropology*. Cincinatti: Wadsworth Publishing Company.

55 Belin, P. R., Zatorre, J., Lafaille, P., Ahad, P. & Pike, B. 2000 Voice-selective areas in human auditory cortex. *Nature* **403**, 309–312. (doi:10.1038/35002078)

56 Binder, J. R., Frost, J. A., Hammeke, T. A., Cox, R. W., Rao, S. M. & Prieto, T. 1997 Human brain language areas identified by functional magnetic resonance imaging. *J. Neurosci.* **17**, 353–362.

57 Relethford, J. H. 2007 *The human species: an introduction to biological anthropology*. New York, NY: McGraw-Hill.

58 Dunbar, R. M. 2005 *The human story*. New York, NY: Faber & Faber.

59 Burrows, A. M. 2008 The facial expression musculature in primates and its evolutionary significance. *BioEssays* **30**, 212–225. (doi:10.1002/bies.20719)

60 Grusec, J. E. & Kuczynski, L. 1997 *Parenting and children's internalization of values: a handbook of contemporary theory*. New York, NY: John Wiley & Sons.

61 Nisbett, R. E. & Cohen, D. 1996 *Culture of honor: the psychology of violence in the south*. Boulder, CO: Westview Press.

62 Parsons, T. 1967 *Sociological theory and modern society*. New York, NY: Free Press.

63 Rozin, P., Lowery, L., Imada, S. & Haidt, J. 1999 The CAD triad hypothesis: a mapping between three moral emotions (contempt, anger, disgust) and three moral codes (community, autonomy, divinity). *J. Personality Soc. Psychol.* **76**, 574–586. (doi:10.1037/0022-3514.76.4.574)

64 Gintis, H., Bowles, S., Boyd, R. & Fehr, E. 2005 *Moral sentiments and material interests: on the foundations of cooperation in economic life*. Cambridge, UK: MIT Press.

65 Pinker, S. 2002 *The blank slate: the modern denial of human nature*. New York, NY: Viking.

66 Tooby, J. & Cosmides, L. 1992 The psychological foundations of culture. In *The adapted mind: evolutionary psychology and the generation of culture* (eds J. H. Barkow, L. Cosmides & J. Tooby), pp. 19–136. New York, NY: Oxford University Press.

67 Gintis, H. 2003 Solving the puzzle of human prosociality. *Rationality Soc.* **15**, 155–187. (doi:10.1177/1043463103015002001)

68 Gintis, H. 1975 Welfare economics and individual development: a reply to Talcott Parsons. *Q. J. Econ.* **89**, 291–302. (doi:10.2307/1884434)

69 Wrong, D. H. 1961 The oversocialized conception of man in modern sociology. *Am. Soc. Rev.* **26**, 183–193. (doi:10.2307/2089854)

70 Camerer, C. 2003 *Behavioral game theory: experiments in strategic interaction*. Princeton, NJ: Princeton University Press.

71 Gintis, H., Bowles, S., Boyd, B. & Fehr, E. 2003 Explaining altruistic behavior in humans. *Evol. Hum. Behav.* **24**, 153–172. (doi:10.1016/S1090-5138(02)00157-5)

72 Gintis, H. 2000 Strong reciprocity and human sociality. *J. Theor. Biol.* **206**, 169–179. (doi:10.1006/jtbi.2000.2111)

73 West, S., Griffin, A. S. & Gardner, A. 2007 Social semantics: altruism, cooperation, mutualism, strong reciprocity and group selection. *J. Evol. Biol.* **20**, 415–432. (doi:10.1111/j.1420-9101.2006.01258.x)

74 Kerr, B., Godfrey-Smith, P. & Feldman, M. W. 2004 What is altruism? *Trends Ecol. Evol.* **19**, 135–140. (doi:10.1016/j.tree.2003.10.004)

75 Haley, K. J. & Fessler, D. M. T. 2005 Nobody's watching? subtle cues affect generosity in an anonymous economic game. *Evol. Hum. Behav.* **26**, 245–256. (doi:10.1016/j.evolhumbehav.2005.01.002)

76 Trivers, R. L. 2007 Reciprocal altruism: 30 years later. In *Cooperation in primates and humans: mechanisms and evolution* (eds P. K. Kappeler & C. P. van Schaik), pp. 67–85. Berlin, Germany: Springer.

77 Bowles, S. & Gintis, H. 2009 A cooperative species: human reciprocity and its evolution. *J. Econ. Theory* **17**, 1–14.

78 Fehr, E. & Joseph, H. 2004 Is strong reciprocity a maladaption? on the evolutionary foundations of human altruism. In *Genetic and cultural origins of cooperation* (ed. P. Hammerstein). Cambridge, MA: The MIT Press.

79 Silk, J. B. 2006 The evolution of cooperation in primate groups. In *Moral sentiments and material interests: on the foundations of cooperation in economic life* (eds H. Gintis, S. Bowles, R. Boyd & E. Fehr), pp. 43–73. Cambridge, UK: The MIT Press.

80 Silk, J. B., Brosnan, S. F., Vonk, J., Henrich, J., Povinelli, D. J., Richardson, A. S., Lambeth, S. P., Mascaro, J. & Schapiro, S. J. 2005 Chimpanzees are indifferent to the welfare of unrelated group members. *Nature* **437**, 1357–1359. (doi:10.1038/nature04243)

81 Keser, C. & van Winden, F. 2000 Conditional cooperation and voluntary contributions to public goods. *Scand. J. Econ.* **102**, 23–39. (doi:10.1111/1467-9442.00182)

82 Fehr, E. & Gächter, S. 2000 Cooperation and punishment. *Am. Econ. Rev.* **90**, 980–994. (doi:10.1257/aer.90.4.980)

83 Gächter, S. & Falk, A. 2002 Reputation and reciprocity: consequences for the labour relation. *Scand. J. Econ.* **104**, 1–26. (doi:10.1111/1467-9442.00269)

84 Andreoni, J. & Miller, J. H. 2002 Giving according to GARP: an experimental test of the consistency of preferences for altruism. *Econometrica* **70**, 737–753. (doi:10.1111/1468-0262.00302)

85 Varian, H. R. 1982 The nonparametric approach to demand analysis. *Econometrica* **50**, 945–972. (doi:10.2307/1912771)

86 Fischbacher, U., Fong, C. M. & Fehr, E. 2009 Fairness, errors and the power of competition. *J. Econ. Behav. Organ* **72**, 527–545. (doi:10.1016/j.jebo.2009.05.021)

87 Kiyonari, T., Tanida, S. & Yamagishi, T. 2000 Social exchange and reciprocity: confusion or a heuristic? *Evol. Hum. Behav.* **21**, 411–427. (doi:10.1016/S1090-5138(00)00055-6)

88 Ledyard, J. O. 1995 Public goods: a survey of experimental research. In *The handbook of experimental economics* (eds J. H. Kagel & A. E. Roth), pp. 111–194. Princeton, NJ: Princeton University Press.

89 Ostrom, E., Walker, J. M. & Gardner, R. 1992 Covenants with and without a sword: self-governance is possible. *Am. Political Sci. Rev.* **86**, 404–417. (doi:10.2307/1964229)

90 Yamagishi, T. 1986 The provision of a sanctioning system as a public good. *J. Personality Soc. Psychol.* **51**, 110–116. (doi:10.1037/0022-3514.51.1.110)

91 Gächter, S. & Fehr, E. 1999 Collective action as a social exchange. *J. Econ. Behav. Organ.* **39**, 341–369. (doi:10.1016/S0167-2681(99)00045-1)
92 Orbell, J. M., Dawes, R. M. & van de Kragt, J. C. 1986 Organizing groups for collective action. *Am. Political Sci. Rev.* **80**, 1171–1185. (doi:10.2307/1960862)
93 Sato, K. 1987 Distribution and the cost of maintaining common property resources. *J. Exp. Soc. Psychol.* **23**, 19–31. (doi:10.1016/0022-1031(87)90023-0)
94 Yamagishi, T. 1988 The provision of a sanctioning system in the United States and Japan. *Soc. Psychol. Q.* **51**, 265–271. (doi:10.2307/2786924)
95 Yamagishi, T. 1988 Seriousness of social dilemmas and the provision of a sanctioning system. *Soc. Psychol. Q.* **51**, 32–42. (doi:10.2307/2786982)
96 Yamagishi, T. 1992 Group size and the provision of a sanctioning system in a social dilemma. In *Social dilemmas: theoretical issues and research findings* (eds W. B. G. Liebrand, D. M. Messick & H. A. M. Wilke), pp. 267–287. Oxford, UK: Pergamon Press.
97 Sally, D. 1995 Conversation and cooperation in social dilemmas. *Rationality Soc.* **7**, 58–92. (doi:10.1177/1043463195007001004)
98 Bochet, O., Talbot, P. & Louis, P. 2006 Communication and punishment in voluntary contribution experiments. *J. Eco. Behav. Organ.* **60**, 11–26. (doi:10.1016/j.jebo.2003.06.006)
99 Brosig, J., Ockenfels, A. & Weimann, J. 2003 The effect of communication media on cooperation. *German Econ. Rev.* **4**, 217–242. (doi:10.1111/1468-0475.00080)
100 Gneezy, U. 2005 Deception: the role of consequences. *Am. Econ. Rev.* **95**, 384–394. (doi:10.1257/0002828053828662)
101 Boles, T. L., Rachel, T., Croson, A. & Murnighan, J. K. 2000 Deception and retribution in repeated ultimatum bargaining. *Organ. Behav. Hum. Decis. Processes* **83**, 235–259. (doi:10.1006/obhd.2000.2908)
102 Charness, G. & Dufwenberg, M. 2004 *Promises and partnership*. Santa Barbara: University of California.
103 Gunnthorsdottir, A., McCabe, K. & Smith, V. L. 2002 Using the Machiavellianism instrument to predict trustworthiness in a bargaining game. *J. Econ. Psychol.* **23**, 49–66. (doi:10.1016/S0167-4870(01) 00067-8)
104 Burks, S. V., Carpenter, J. P. & Verhoogen, E. 2003 Playing both roles in the trust game. *J. Econ. Behav. Organ.* **51**, 195–216. (doi:10.1016/S0167-2681(02)00093-8)