

9

Socialization

The influence of society is what had roused in us the sentiments of sympathy and solidarity drawing us toward others; it is society which, fashioning us in its image, fills us with religious, political and moral beliefs that control our actions.

Durkheim Suicide,(1951[1897]) pp. 211–212.

9.1 Introduction

In addition to trial and error experimentation, preferences are acquired by genetic inheritance—a taste for sweets, for example—and by a learning process, termed cultural transmission, from our parents, others elders, and our peers—a taste for rice over potatoes, for instance. As we have seen (Chapter 1), genetic and cultural transmission are in many ways similar, a fact has been exploited by the classic contributions to the modeling of cultural evolution by Cavalli-Sforza and Feldman (1981) and Boyd and Richerson (1985). The main similarity between the biological and cultural processes that is exploited by these models is the fact that both social learning and genetic inheritance from parents can be represented as the replication of traits over time. Two additional similarities may be mentioned.

First, whether of cultural or genetic origin, the taste for sweets or rice activates the same reward-processing regions of the brain. The taste for sweets is certainly more universal among humans than is the taste for rice. But, there is no meaningful sense in which one can say that one is more deeply rooted or fundamental than the other. The genetically transmitted taste for sweets can easily be unlearned (a nauseating experience with sweet food overrides a genetic predisposition to like sweets, for instance). Similarly, culturally learned traits, such as the U.S. Southern “culture of honor” (Nisbett and Cohen 1996), has physiological correlates, such as elevated testosterone when insulted, much as physical danger elevates adrenaline in virtually all humans.

Second, those who are relatively successful in acquiring material resources tend to produce more copies of their traits in the next generation, whether

the process works through their differential success in producing offspring who survive to reproductive age or because of their greater likelihood of being copied as cultural models. In the previous chapters we have specified evolutionary processes in which the frequency of a behavioral type in the population increases if its expected payoff exceeds the average. These so-called “payoff-monotone models” provide a challenging, if highly simplified, way of posing the puzzle we are addressing, namely, the evolution of preferences that induce people to act in ways that reduce their payoffs by comparison to other members of their group. In the models proposed in Chapters 6-8, altruistic traits may overcome their within-group payoff disadvantages first, because of the superior payoffs enjoyed by members of groups in which there are many altruists and second, because groups devise, and culturally transmit over generations, the institutions that mitigate the within-group selection pressures tending to eliminate altruists.

Cultural transmission provides an additional way that the payoff disadvantages of altruistic social preferences might be overcome: the behaviors favored by cultural transmission need not be those that garner higher fitness or other measures of payoffs. Pre-industrial cities provide an example (Knauff 1989). Prior to modern medicine the city was a cultural success, recruiting steady streams of migrants to forsake the countryside for urban living. But, it was a biological disaster, typically not reproducing its own population even among the social elites residing there. A second example is the demographic transition, whereby the culturally-transmitted preference for smaller families proliferated in many populations despite having apparently reduced fitness (Zei and Cavalli-Sforza 1977, Kaplan, Lancaster, Block and Johnson 1995, Ihara and Feldman 2004).

In his book *Sick Societies*, Robert Edgerton (1992) catalogues dozens of examples of culture overriding fitness, all, as the title suggests, with unpleasant consequences. But, if cultural transmission can induce people to limit their fitness or to choose a lethal residential environment, it certainly might also overcome the payoff disadvantages associated with altruistic social preferences (Feldman and Cavalli-Sforza 1976). It is this possibility that we explore here.

9.2 Cultural Transmission and Socialization

Cultural transmission may overrule fitness because it causes people to want to do things that result in their having fewer offspring. Thus our explanation

will involve the proximate causes of behavior, that is to say, preferences. Here, and in the next chapter as well, we depart from the framework of the previous three chapters, which have focused entirely on fitness and behavior without exploring the question of motivation. It is not difficult, of course, to associate proximate motives with the kinds of behaviors that we have shown may evolve. Ethically motivated outrage—moralistic aggression—is a plausible motivation for the strong reciprocators' punishment of defectors in Chapter 6, and group loyalty and outgroup hostility could provide the psychological basis for the behaviors studied in Chapters 7 and 8. Our models show that these and other preferences motivating the behaviors in question could have evolved by a fitness based evolutionary process. Here we seek to understand how altruistic preferences might evolve under the influence of cultural transmission.

We will take account of two facts. First, the phenotypic expression of an individual's genetic inheritance depends on a developmental process that is plastic and open-ended. One expression of this fact is that while human ancestral groups are similar genetically (Feldman, Lewontin and King 2003), they differ in important ways in behaviors. We surveyed some of our experimental evidence for this behavioral variability in Chapter 3. This developmental plasticity explains why humans are among the most ubiquitous of species, capable of making a living and surviving in virtually all of the world's environments.

Second, this developmental process is deliberately structured, by elders, teachers, political leaders, and religious figures, to foster certain kinds of development and to thwart others. In many of Edgerton's "sick" societies, the socialization processes affecting development motivate lethal practices such as cigarette smoking or, as in the highlands of New Guinea, consuming the brains of deceased relatives (Cavalli-Sforza and Feldman 1981, Durham 1991). In both cases individuals contract a terminal illness with high probability. But, in most societies, socialization stresses not only the desirability of behaviors that contribute to one's own well-being, such as moderation, planning ahead, and personal hygiene, but also those that benefit others, such as the altruistic social preferences we have identified as common among humans.

In this chapter we analyze the process by which social norms become internalized, that is, taken on as preferences to be sought in their own right rather than constraints on behavior or instrumental means to other ends. Internalization is thus an aspect of cultural transmission that affects prefer-

ences rather than beliefs and capacities. The idea of internalized norms is captured in a passage attributed to Abraham Lincoln: “when I do good, I feel good. When I do bad, I feel bad. That is my religion.”

Much of the content of cultural transmission can be modeled as information transfer. Members of a group, most often as children, are taught ‘how to’ accomplish particular ends such as acquiring and preparing food or performing music, as in the study of the Central African Aka by Hewlett and Cavalli-Sforza (1986). We focus instead on the process by which a society’s ‘oughts’ become its members ‘wants,’ thereby narrowing the hiatus between what Jeremy Bentham famously termed people’s ‘dutys’ and their ‘interests.’ As a result, we draw upon studies of how values, rather than factual information, are transmitted, such as generosity among the Inuit (Guemple 1988), social solidarity among children on Israeli kibbutzim (Bronfenbrenner 1969), and the control of hostility among children in cross-cultural perspective (Whiting and Whiting 1975). We refer to these ‘ought’ rules of behavior as norms and when they are internalized, as preferences.

Though drawing on distinct neurological substrates in the individual and on a somewhat different mix of social institutions for its accomplishment, the internalization of norms has enough in common with other aspects of cultural transmission that we can draw upon the models of Boyd and Richerson and of Cavalli-Sforza and Feldman. We posit three influences on the cultural transmission of preferences in a population. The first, the vertical transmission of traits from parents to offspring is coupled with the fact that some parental traits will be associated with greater fitness. As a result, these traits will be passed on to larger numbers in the next generation. The second is oblique transmission to the young from non-parental members of the parents’ generation in the myriad of personal interactions with neighbors, teachers, and spiritual leaders by which the young are socialized to internalize particular norms (Cavalli-Sforza and Feldman 1981). Third is payoff-based social learning according to which periodically, over the life course, people compare their behaviors with the behaviors of other individuals, and tend to adopt behaviors of others who appear to be doing relatively well. We take account of the effect of payoffs on the adoption of norms in order to counter the “oversocialized” concept of the individual according to which socialization simply implants norms in a passive and uncritical target (Wrong 1961, Gintis 1975).

Following the lead of Boyd and Richerson (1985), oblique transmission might be *conformist*, the young tending to adopt the behaviors most com-

mon in the parental generation, irrespective of their payoffs. In this case resulting dynamic will not be monotonic in either fitness or well-being. If virtually all of the population is altruistic, conformist transmission might overcome the payoff disadvantage suffered by the altruists and allow their persistence in a population. Conformism may also stabilize payoff-reducing behaviors that yield no benefit to others, such as smoking. Indeed, this is the most parsimonious explanation of the long term persistence of many of the dysfunctional behaviors documented by Edgerton. Conformism may thus contribute to large between-group differences in behavior, with selection against low payoff behaviors within groups being weak or absent. In the presence of strong conformism, weak group selection may be sufficient to stabilize altruistic preferences. Conformist cultural transmission is unquestionably an important influence on the evolution of preferences.

The approach we have adopted here, however, is to represent oblique transmission as purposeful directed socialization rather than simple conformism. We do this for empirical reasons: most societies devote substantial time and resources to deliberately socializing the young to act in ways that are beneficial to others, and an adequate explanation of social preferences needs to take account of this fact.

All three of these learning processes, vertical, oblique, and payoff-based updating, involve the internalization of norms, which is itself a puzzling process. Taking on a general rule of behavior as an objective rather than a constraint or an instrument towards some other end is likely to be costly, as the rule will not be ideally suited to all situations, and its internalization deprives the individual of flexibility in dealing with such situations on a case-by-case basis. The parochial preferences that motivate the exclusion of outsiders is an example of a personally costly general rule of behavior. Why, then, are humans so susceptible to internalizing general rules? If this susceptibility were subject to a payoff-based selection process, whether fitness-based or payoff-sensitive cultural copying, one might expect it to be eliminated from any population in which it appeared. Why then are general rules of behavior common? An answer that we find persuasive (Heiner 1985) is that internalizing general rules of behavior may persist in an evolutionary dynamic because it relieves the individual from calculating the costs and benefits in each situation and reduces the likelihood of making costly errors.

Why should the norms that are internalized be altruistic? Simon (1990) and Caporael, Dawes, Orbell and van de Kragt (1989) proposed that altruism might proliferate in a population because it is an inseparable part of a

ensemble of culturally transmitted norms that is, on balance, individually advantageous. Simon termed the capacity to internalize such an ensemble of social norms *docility* (literally, ‘teachability’) and explained the evolution of altruistic behaviors as a consequence of the fact that the norms motivating them are linked to other norms that benefit the individual sufficiently to offset the individual costs of altruism. Altruism in this case proliferates in the same way that a genetically transmitted disadvantageous trait may evolve if it is pleiotropically linked to other, advantageous traits and thus may hitchhike on their success.

We wish to explore this reasoning and address two aspects in which it is incomplete. First, one needs to address the above puzzle of how the capacity to internalize norms evolves (Section 9.3). Second, we would like to relax the *ad hoc* “pleiotropic analogy” whereby individually costly altruism and individually beneficial other norms are inseparable. In the model we present in Sections 9.4 and 9.5, as in the Simon model, altruism will ‘hitchhike’ on the beneficial aspects of internalizing other norms. But here the hitchhiker is not forced upon the driver, and rather is picked up voluntarily as it were, the endogenous result of a gene-culture evolutionary dynamic. In Section 9.6 we incorporate the insight that the higher the cost of holding an altruistic norm, the more likely are individuals to abandon that norm. We accomplish this by adding a payoff-based updating process to our dynamical system. In Section 9.7 we show that even when individually beneficial and individually costly altruistic behaviors are not linked other than by the fact that both may be internalized, altruism may successfully hitchhike under plausible parameter values. This is why, despite the evidence provided by Edgerton, institutions of socialization tend to favor altruistic social preferences.

In these models the existence of an internalized norm that enhances fitness is critical to the evolution of the capacity for internalization. But developing the capacity to internalize norms is costly to the individual, and sustaining the institutions whereby internalization takes place is costly to society. Why would evolution favor bearing these costs rather than relying on genetic transmission to sustain individually beneficial norms? The answer we propose in Section 9.8 is an application of the reasoning of Boyd and Richerson (2000). The cultural transmission of norms allowed humans uniquely among animals to adapt flexibly to rapidly changing circumstances and to modify the results of individual fitness maximization where these are not beneficial on average to members of a group. Finally, in Section 9.9 we comment

on the novel possibilities unique to humans arising from the fact that our preferences are substantially detached from the dictates of fitness.

9.3 When is Virtue its Own Reward?

Consider a group in which members can either adopt, or not, a certain cultural norm A. We shall call those who adopt norm A *altruists*, and we call those who do not adopt norm A *self-interested* types, or “S-types.” Altruism is costly, in that self-interested types have fitness 1, as compared with altruists, who have fitness $1 - s$, where $0 < s < 1$ is a viability loss. We assume in each generation that individuals pair off randomly, mate, and have offspring in proportion to their fitness, after which they die. Families pass on their cultural norms to their offspring, so offspring of AA parents are altruists, offspring of SS parents are self-interested, and half of the offspring of AS-families (which are the same as SA-families) are altruists, the other half self-interested (we call this *vertical transmission*). We also assume that the self-interested offspring of AS- and SS-families, but not the self-interested offspring of the SS parents, are susceptible to influence by socialization institutions promoting altruistic norms, a fraction of such offspring becoming altruists (*oblique transmission*).

Vertical transmission is as follows. Suppose there are n males and n females at the beginning of the first period. If the fraction of altruists is f_A , there will be nf_A^2 AA-families, who will have $nf_A^2(1 - s)^2\beta_o^2$ offspring, all of whom are altruists, where β_o is baseline fitness. There will also be $2nf_A(1 - f_A)$ AS-families, who will have $2nf_A(1 - f_A)(1 - s)\beta_o^2$ offspring, half of whom are altruists. Finally there will be $n(1 - f_A)^2$ SS-families who will have $n(1 - f_A)^2\beta_o^2$ offspring. Adding up the number of offspring and setting their number equal to $2n$, we see that we must have $\beta_o = 1/(1 - sf_A)$. Thus, the frequencies of AA, AS, and SS offspring are given by

$$f_{AA} = \frac{f_A^2(1 - s)^2}{(1 - f_{AS})^2}, \quad f_{AS} = \frac{2f_A(1 - f_A)(1 - s)}{(1 - f_{AS})^2}, \quad f_{SS} = \frac{(1 - f_A)^2}{(1 - f_{AS})^2}. \quad (9.1)$$

Oblique transmission then occurs. A fraction $f_A\gamma$ of offspring of AS- and SS- families who are self-interested switch to being altruists under the influence of the oblique transmission of cultural norm A, where γ is a measure of the strength of the oblique transmission process. Note that the effectiveness of oblique transmission is proportional to the fraction of A's in the popula-

tion. We find that the change in the fraction of altruists in the next generation is given by the familiar replicator equation (see Appendix A2):

$$(1 - sf_A)\dot{f}_A = f_A(1 - f_A)(\gamma - s) \quad (9.2)$$

or

$$\dot{f}_A = f(f_A) = f_A(1 - f_A)\frac{\gamma - s}{1 - sf_A}, \quad (9.3)$$

where $(1 - sf_A)$ is the average payoff in the population, $f_A(1 - f_A)$ is a measure of the frequency of A-S pairings in the population, and $(\gamma - s)$ gives the selective advantage (or disadvantage) of the A's over the S's when account is taken of both oblique (γ) and vertical transmission. Equation (9.3) illustrates the tension between the differential fertility effects on the evolution of f_A captured by s and working against the evolution of the A's and the effects of oblique transmission captured by γ , which tend to counteract the selection against altruists. Equation (9.3) shows that when $s = \gamma$, these two effects are exactly offsetting.

Payoff-based updating then occurs. 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 individuals' preference functions do not perfectly track fitness, so it is reasonable to assume that the larger the difference in the payoffs, the more likely the individual is to perceive it, and change. Specifically, we assume the probability p that an A individual will shift to S is proportional to the fitness difference of the two types, so $p = \eta s$ for some *imitation rate* $\eta > 0$.

The expected fraction f'_A of the A population after the above shifts is the fraction before updating f_A , minus those A's who switched to S, the latter being the A's who were paired with a S, who constitute a fraction $f_A(1 - f_A)$ of the population, multiplied by the probability of a switch taking place in these cases. Thus we have

$$f'_A = f_A - \eta sf_A(1 - f_A). \quad (9.4)$$

We now combine these three sources of change in the fraction of altruists, adding the changes described in equation 9.4 to those already shown in 9.3, giving

$$\dot{f}_A = f_A(1 - f_A)\frac{\gamma - s}{1 - sf_A} - \eta f_A(1 - f_A)s \quad (9.5)$$

The second term represents the influence of payoff-based updating, reducing the frequency of the altruistic norm, in comparison with the vertical and oblique cultural transmission mechanisms, represented by the first term, which may favor this norm. Simplifying (9.5) we have

$$\dot{f}_A = \frac{f_A(1 - f_A)}{1 - sf_A}(\gamma - s - s\eta(1 - sf_A)). \quad (9.6)$$

Thus whether the share of altruists increases or decreases depends on the sign of $\gamma - s(1 + \eta(1 - sf_A))$, the first term as before representing the effects of oblique transmission and the second term the combined effects of payoff-based updating and differential parental fertility.

We call the situation $\dot{f}_A = 0$, $0 \leq f_A \leq 1$ an *equilibrium* of the dynamical system. Assuming $\gamma \geq 0$ is given, we show in (Gintis 2003b) that if

$$s < s_{\min} = \frac{\gamma}{1 + \eta}, \quad (9.7)$$

$f_A = 1$ is a globally stable altruistic equilibrium. This is because when (9.7) holds, for any interior value of f_A , \dot{f}_A is positive. To see this, note that the selection pressures operating against the A's are greatest when there are few of them in the population (because the force of pro-A oblique transmission is then minimized). When $f_A = 0$, the selection pressure operating against the A's (through the combined effects of differential parental fertility and payoff based updating) is $s(1 + \eta)$ and the sign of \dot{f}_A is given by $\gamma - s(1 + \eta)$. Condition 9.7 assures that \dot{f}_A is positive under these conditions, and by implication for any other value of f_A other than $f_A = 1$. Altruistic norms persist in a equilibrium only if there is a strictly positive rate of cultural transmission of altruism *via* social institutions.

If

$$s_{\min} < s < s_{\max} = \frac{1}{2\eta} \left\{ 1 + \eta - \sqrt{(1 + \eta)^2 - 4\gamma\eta} \right\}, \quad (9.8)$$

both $f_A = 0$ and $f_A = 1$ are locally stable equilibria of the system and there is third unstable equilibrium $0 < f_A^* < 1$ separating the basins of attraction of the two stable equilibria: both self-interested and altruistic equilibria are stable. Multiple equilibria arise in the model because of the positive feedbacks given by our assumptions that the strength of pro-altruist oblique transmission is increasing in the frequency of altruists in the population.

Finally if $s > s_{\max}$, then $f_A = 0$ is the only stable equilibrium of the system. This is the case because if s exceeds s_{\max} , then $\gamma - s(1 + \eta(1 - sf_A))$

is negative for even the most A-favorable case, namely when $f_A = 1$, so \dot{f}_A is then negative for all interior values of f_A , leading to the stability of the all-S equilibrium.

We see, then, that the higher the personal cost of altruistic behavior, the more stringent the conditions under which altruism will emerge. Our result illustrates the tension between socialization institutions and the psychological mechanism of norm internalization on the one hand, and payoff-based updating that induces individuals to shift to higher payoff behaviors, whatever the effect of these behaviors on others, and on society as a whole, on the other hand. This tension is also revealed by noting that the altruistic equilibrium is globally stable if the strength of payoff-based updating η is less than the difference in the size of the oblique transmission and the fitness cost of altruism, normalized by the latter:

$$\eta < \frac{\gamma - s}{s}, \quad (9.9)$$

If

$$\frac{\gamma - s}{s} < \eta < \frac{\gamma - s}{s(1 - s)}, \quad (9.10)$$

both the self-interested and the altruistic equilibria are locally stable, and the basin of attraction of the altruistic equilibrium shrinks as η increases. Finally, if

$$\eta > \frac{\gamma - s}{s(1 - s)}, \quad (9.11)$$

the self-interested equilibrium is globally stable.

9.4 The Genetics of Internalization

Thus if the internalization of norms accomplished by the society's socialization processes is sufficiently strong relative to the strength of payoff-based updating and the cost of altruism, the altruistic equilibrium may be stable. In effect, there is a net flow into altruism at rate γ , the rate of oblique transmission, a net flow out of altruism due to its fitness cost s , and another flow out because individuals switch from altruistic to selfish behavior by copying the more successful selfish individuals, at rate η . When the net balance favors a positive flow into altruism, i.e., when $\gamma > s(1 + \eta(1 - s))$, the altruistic equilibrium is at least locally stable.

But why would people internalize norms at all? To answer this question we present a model of the coevolution of a genetically transmitted capacity to internalize norms and a culturally transmitted norm that can be learned only by those who have the capacity to internalize. The capacity to internalize is costly in fitness terms, but if it is used to learn the culturally transmitted norm, the net effect is fitness enhancing. Our objective is to understand the conditions under which the fitness-enhancing norm and the costly capacity to internalize will co-evolve. Having done this we will investigate if an altruistic norm can proliferate if it is subject to an evolutionary cultural dynamic. Here we assume that cultural traits are acquired from parents or through oblique transmission, but that payoff based switching of traits, as modeled in the previous section, does not occur. We introduce payoff based cultural updating in this gene-culture coevolutionary model in Section 9.5.

To simplify the analysis we assume that there is one genetic locus that controls the capacity to internalize norms, and that norm internalization is the expression of a single allele, which we will call the ‘internalization allele’. This is highly unlikely, but a more complicated treatment would not provide any additional illumination of the questions under investigation. We will assume that each individual has only one copy at this locus (i.e., genetics are haploid), which can be inherited with equal probability from either parent (a diploid model, in which each locus has two alleles at each locus, has almost the same properties as the haploid model, but is much more complicated). Individuals without the proper allele cannot internalize norms, whereas individuals with the proper allele are capable of internalization, but whether or not they internalize a norm depends on costs and benefits, as well as the individual’s personal history. In this section we assume that an internal norm is fitness enhancing and we derive the conditions under which the allele for internalization of norms is globally stable. Suppose the norm in question is *C* (Cleanliness, for instance), which confers fitness $1 + t > 1$, while the normless phenotype, denoted by *D* (Dirty, perhaps), has baseline fitness 1. There is a genetic locus with two alleles, *a* and *b*. Allele *a* permits the internalization of norms, whereas *b* does not. We assume that possessing *a* imposes a fitness cost *u*, with $0 < u < 1$, on the grounds that there are costly physiological and cognitive prerequisites for the capacity to internalize norms. We assume $(1 + t)(1 - u) > 1$, so the cost of the internalization allele is more than offset by the benefit of the norm *C*. An individual is now characterized not only by his genes, but his phenotype

(whether he is a C or a D). There are thus three “phenogenotypes,” whose fitnesses are shown in Table 9.1.

Individual Phenogenotype	Individual Fitness
aC	$(1 - u)(1 + t)$
aD	$1 - u$
bD	1

Table 9.1. Fitnesses of the Three Phenogenotypes. Here u is the fitness cost of possessing the internalization allele, and t is the fitness value of possessing the norm C. Note that bC cannot occur because an individual must have a to be able to internalize C.

The rules of gene-culture transmission are as follows. If familial phenogenotype is $xyXY$, where $\{x, y\} \in \{a, b\}$, $X, Y \in \{C, D\}$, an offspring is equally likely to inherit x or y . An offspring whose genotype is a is equally likely to inherit X or Y. But an offspring of genotype b always has the normless phenotype D. The transition table is shown in Table 9.2.

Familial Type	Offspring Phenogenotypic Frequency		
	aC	aD	bD
$aaCC$	1		
$aaCD$	1/2	1/2	
$aaDD$		1	
$abCD$	1/4	1/4	1/2
$abDD$		1/2	1/2
$bbDD$			1

Table 9.2. Phenotypic Inheritance is Controlled by Genotype. Note that bCC and bCD are not listed. This is because bC cannot occur, because an individual must have the a allele to internalize the C norm.

Families are formed by random pairing, males and females are indistinguishable (i.e., there is recombination but only one sex), and offspring genotypes obey the laws of Mendelian segregation. A family is characterized by its *familial genotype*, which is the pattern of genes of the two members, and its *familial phenotype*, which is the pattern of norms of the two members. Thus there are three familial genotypes, aa , ab , bb . We assume also

that only the phenotypic traits of parents, and not which particular parent expresses them, are relevant to the transmission process. Therefore, there are three familial phenotypes, CC, CD, and DD, and 9 familial phenogenotypes, of which only 6 can occur (because a parent of genotype b must have the D phenotype). The frequencies of familial phenogenotypes are as shown in Table 9.3, where $P(i)$ represents the frequency of phenogenotype $i = aC, aD, bD$. For example, the $aaCD$ phenogenotype can occur in two ways: father aC and mother aD , or *vice-versa*. The probability of each is proportional to $P(aC)P(aD)$ time the product of their fitnesses. One parent has fitness $(1 - u)(1 + t)$ while the other has fitness $1 - u$. This give the entry in the second line of Table 9.3.

Phenogenotype	Frequency
$aaCC$	$P(aC)^2(1 - u)^2(1 + t)^2\beta_o^2$
$aaCD$	$2P(aC)P(aD)(1 - u)^2(1 + t)\beta_o^2$
$aaDD$	$P(aD)^2(1 - u)^2\beta_o^2$
$abCD$	$2P(aC)P(bD)(1 - u)(1 + t)\beta_o^2$
$abDD$	$2P(aD)P(bD)(1 - u)\beta_o^2$
$bbDD$	$P(bD)^2\beta_o^2$

Table 9.3. Frequencies of Phenogenotypes. Here, β_o is baseline fitness, and is chosen so the sum of the frequencies is unity. Note that bCC and bCD are not listed, because bC cannot occur.

Equilibrium occurs when each phenogenotype has constant generation to generation frequency. In this case, we need consider only two of the phenogenotypes, say aC and aD , because bC cannot occur, and $P(bD) = 1 - P(aC) - P(aD)$. This system has three equilibria, in which the whole population bears a single phenogenotype. These are aC , in which all individuals internalize the fitness enhancing norm, aD , in which the internalization allele is present but the phenotype C is absent, and bD , in which neither the internalization allele nor the norm is present.

Gintis (2003b) proves the following assertions concerning the stability of the various equilibria of the diploid version of this system; the results reported are for the haploid version, however. First, the aD equilibrium is unstable, while the aC equilibrium is locally stable. The unnormed equilibrium bD is locally stable if $(1 - u)(1 + t) < 2$, and unstable when the opposite inequality holds. There are two conditions that render the bD equilibrium unstable,

in which case aC will be globally stable. The first is the condition that t is sufficiently large that $(1 - u)(1 + t) > 2$. Second, if parental transmission is sufficiently biased in favor of C, the internalization equilibrium is globally stable. We consider the first of these conditions rather implausible, but the biased transmission condition is quite plausible. Thus, parental transmission biased towards the internal norm are plausible forces fostering the global stability of the aC equilibrium.

9.5 Altruism as Hitchhiker

We now add a second dichotomous phenotypic trait with two variants. Internal norm A is altruistic in the sense that its expression benefits the group, but imposes fitness loss s , with $0 < s < 1$ on those who adopt it. The normless state, S, is neutral, imposing no fitness loss on those who adopt it, but also no gain or loss to other members of the social group. An individual phenotype is then one of SD (internalizes neither norm), SC (internalizes only the fitness-enhancing norm), AD (internalizes only the altruistic norm), and AC (internalizes both the fitness-enhancing and altruistic norm).

We assume A has the same cultural transmission rules as C: a -individuals inherit their phenotypes from their parents, while b -individuals always adopt the normless phenotype SD. In addition, there is oblique transmission, as before. There are now two genotypes and four phenotypes, giving rise to five phenogenotypes that can occur, which we denote by aAC , aAD , aSC , aSD , and bSD , and three that cannot occur, bAC , bAD , and bSC . We represent the frequency of phenogenotype i by $P(i)$, for $i = aAC, \dots, SD$.

We maintain the assumption that families are formed by random pairing and the offspring genotype obeys Mendelian segregation. We assume also that only the phenotypic traits of parents, and not which particular parent expresses them, are relevant to the transmission process. Therefore there are nine family phenotypes, which can be written as AACC, AACD, AADD, ASCC, ASCD, ASDD, SSCC, SSCD, and SSDD. It follows that there are 27 familial phenogenotypes, which we can write as $aaAACC, \dots, bbSSDD$, only 14 of which can occur. We write the frequency of familial phenogenotype j as $P(j)$, and we assume the population is sufficiently large that we can ignore random drift. For instance, $aaAACC$ represents the case where both parents have the internalization allele a , and both parents internalize the altruistic norm A and the fitness-enhancing norms. Similarly, $aaAACD$ represents the case where both parents have the internalization allele a ,

and both parents internalize the altruistic norm A, but only one internalizes the fitness-enhancing norm C. Finally, *abASCD* represents the case one parent carries the internalization norm and the other does not, the former internalizing both the altruistic norm A, and the fitness-enhancing norm C. For illustrative purposes, here are a few of the phenogenotypic frequencies,

$$\begin{aligned} P(aaAACC) &= P(aAC)^2(1-u)^2(1+t)^2(1-s)^2\beta_o^2, \\ P(aaAACD) &= 2P(aAC)P(aAD)(1-u)^2(1-s)^2(1+t)\beta_o^2, \\ P(abASCD) &= 2P(2aAC)P(bSD)(1-u)(1+t)(1-s)\beta_o^2, \\ P(bbSSDD) &= P(bSD)^2\beta_o^2, \end{aligned}$$

and so on for each of the phenogenotypes, where β_o is baseline fitness, and is chosen so the sum of the frequencies is unity. To understand this calculation, consider, for instance the *abASCD* phenogenotype. This can arise in two ways: (1) *aAC* mother and *bSD* father or (2) *bSD* mother and *aAC* father. In both cases, one parent came from a pool with fitness $(1-s)(1+t)(1-u)$ and the other with fitness 1.

The rules of cultural transmission are as before. If familial phenogenotype is $xyXYZW$, where $\{x, y\} \in \{a, b\}$, $\{X, Y\} \in \{A, S\}$, and $Z, W \in \{C, D\}$, an offspring is equally likely to inherit x or y . An a offspring is equally likely to inherit X or Y , and equally likely to inherit Z or W . Offspring of genotype b always have the normless phenotype SD . We assume oblique transmission takes the form of a -individuals with phenotype S switching to phenotype A at a rate equal to $\gamma p_A p_S$, where p_A and p_S are the frequencies of a -individuals of phenotype A and S , respectively.

We assume both genotypic and phenotypic fitness, as well as their interactions, are multiplicative. Thus, for instance, an *aAC* individual incurs a fitness cost u from the capacity to internalize, a fitness gain of t from holding norm C , and a fitness loss s from holding the A norm. The individual's resulting fitness is then $(1-u)(1+t)(1-s)$. Clearly, $(1-u)(1+t)(1-s) > 1$ is a necessary condition for the altruistic trait to evolve, so we assume this inequality holds; i.e., the fitness increase due to having phenotype C must be sufficient to offset both the cost of having the internalization allele and the cost of altruism. The fitness of the phenogenotypes that can appear with positive frequency are as shown in Table 9.4. The resulting system consists of four equations in four of the five offspring phenogenotypes. One offspring phenogenotype is dropped, because the sum of phenogenotype frequencies must be unity.

Individual Phenogenotype	Individual Fitness
aAC	$(1 - u)(1 - s)(1 + t)$
aAD	$(1 - u)(1 - s)$
aSC	$(1 - u)(1 + t)$
aSD	$(1 - u)$
bSD	1

Table 9.4. Payoffs to Five Phenogenotypes.

The population is in equilibrium when the frequency of each phenogenotype is constant from generation to generation. There are thus four equations, one each for the constancy of frequency of aAC , aAD , aSC , and aSD , the frequency of bSD being one minus the sum of the other frequencies. These equations have five equilibria, in which the whole population bears a single phenogenotype. These are aAC , in which all individuals internalize both the altruistic and fitness enhancing norms, aAD , in which only the altruistic norm is internalized, aSC , in which only the fitness-enhancing norm is internalized, aSD , in which individuals carry the allele for internalization of norms, but no norms are in fact internalized, and bSD , in which internalization is absent, and neither altruistic nor fitness-enhancing norms are transmitted from parents to offspring. A check of the eigenvalues of the Jacobian matrix shows that the aAD and the aSD equilibria are unstable, and hence will not survive an evolutionary process.

The analysis of the stability of the aAC equilibrium is given in Gintis (2003a), which shows that stability holds when $s < \gamma$. The inequality $s < \gamma$ is simply (9.7) with $\eta = 0$, which holds because we have not yet introduced payoff-based strategy-shifting. This inequality expresses the key condition that altruism cannot be evolutionarily stable unless the level of oblique transmission is sufficient to overcome the fitness cost of altruism. Groups with high levels of altruism solve the problem of rendering altruism fitness-enhancing by increasing the level of oblique socialization to sufficiently high levels that the new converts to altruistic norms compensate for the lower fitness of altruists.

Gintis (2003a) shows that the aSC is stable when $\gamma < s$, and unstable when the opposite inequality holds. This reinforces the interpretation presented in the previous paragraph. Moreover, as in the single phenotype case, bSD is unstable if $(1 - u)(1 + t) > 2$, and is stable if the opposite inequality holds.

Adding assortative mating leads to the instability of the *bSD* equilibrium under the same conditions as in the single phenotype case.

This analysis shows that if $s < \gamma$, the altruistic phenotype A coexists in a stable equilibrium with the fitness-enhancing phenotype C. We say that A *hitchhikes* on C in the sense that the fitness value of C renders the internalization allele a evolutionarily viable, and once this allele occurs in high frequency, the altruistic phenotype A is evolutionarily viable because its fitness cost s is less than the oblique transmission bias γ , which favors A. Agent-based modeling, presented below, is able to deal with mutations, group interactions, and other aspects of groups as complex dynamical systems. We will see that even under these more demanding, and realistic, conditions, our conclusions continue to hold.

9.6 When can Altruism Survive Payoff-based Updating?

We now add the payoff-based updating dynamic developed in Section 9.3 to our genetic model, thus allowing individuals to shift from lower to higher payoff strategies, and we show that the result is similar to that of the model developed without genetics in Section 9.3. In the current context, there are four phenotypes only a -individuals will copy another phenotype, because only such types are capable of internalizing a norm, and noninternalizers will not desire to mimic internalizers.

We assume an a -individual with the allele and phenotype XY meets an individual of type WZ with probability p_{WZ} , where p_{WZ} is the fraction of the population with phenotype WZ, and in this case switches to WZ with probability η if that type has higher fitness than XY. The parameter η is a measure of the strength of the tendency to shift to high-payoff phenotypes.

Adding payoff-based updating does not change the single phenogenotype equilibria, because all equilibria consist of a single phenogenotype, so in equilibrium, an individual can never meet a distinct phenotype to which he might switch. Checking the eigenvalues of the Jacobian matrix we find that the *aAD* and *aSD* equilibria remain unstable, and payoff-based updating does not affect the conditions for stability of the unnormed equilibrium *bSD*. The condition $\gamma > s$ for stability of the altruism equilibrium *aAC* now becomes

$$\eta < \frac{\gamma - s}{1 - \gamma} \left(\frac{1}{s} - 1 \right). \quad (9.12)$$

Note the similarity to the altruistic equilibrium conditions (9.9), (9.10), and (9.11) in the model without the explicit modeling of genetics. We conclude that a sufficiently strong payoff-based updating process can undermine the stability of the aAC equilibrium. The condition $s > \gamma$ for stability of the nonaltruism internalization equilibrium aSC when payoff-based updating is included now becomes

$$\eta > \frac{\gamma - s}{s(1 + \gamma - s)},$$

and this equilibrium is unstable when the reverse inequality holds. Thus in this case, $s > \gamma$ continues to ensure that aSC is stable, but now for sufficiently large η , this equilibrium is stable even when $\gamma > s$.

In sum, adding payoff-based updating changes the stability properties of the model in only one important way: a sufficiently strong payoff-based updating process can render the nonaltruistic yet internalized equilibrium aSC , rather than the altruistic equilibrium, aAC , stable. The intuition here is that altruism imposes a fitness cost s leading individuals to abandon altruism. The greater the rate at which this occurs, the larger must be the oblique socialization force γ that replenishes the stock of altruists in the group.

In sum, one internalization equilibrium is stable—the altruism equilibrium when $\gamma > s$ and the nonaltruism equilibrium when $s > \gamma$. Either a high return to the internal norm, considerable oblique transmission, or assortative mating of a -individuals, or some combination of the three, ensures that the only stable equilibrium of the system will exhibit a high level of altruism, but a sufficiently high level of imitating successful non-altruists can undermine the altruistic equilibrium.

We may summarize the model as follows. Internalization is individually fitness-enhancing, i.e., $(1-s)(1+t) > 1$, and hence the a allele undergoes positive selection to fixation (all a). Altruism is costly, i.e., $u > 0$, but by assumption, the net fitness of altruists, $(1-s)(1+t)(1-u)$ remains greater than that of non-internalizers, who have fitness unity. Because $u > 0$, internalizers tend to abandon altruism at a positive rate, both because they are at a reproductive disadvantage and they abandon altruism for the selfish alternative at a positive rate, so only an additional force will allow altruism to evolve. The major such force is oblique cultural transmission ($\gamma > s$), although a high level of assortative mating of altruists would work as well.

9.7 Why is Internalization Socially Beneficial?

Internalized norms need not enhance the fitness of group members. Indeed, many social norms are costly, such as those involving invidious displays of physical prowess and promoting costly wars and feuds (Edgerton 1992). The reason for the feasibility of antisocial norms is that once the internalization allele has evolved to fixation, there is nothing to prevent group-harmful phenotypic norms from also emerging, provided they are not excessively costly in comparison with the strength of the payoff-based updating process. The evolution of these harmful norms directly reduces the overall fitness of the population.

Yet, as Brown (1991) and others have shown, there is a tendency in virtually all successful societies for cultural institutions to promote social and eschew anti-social norms, and for altruistic individuals to embrace these social norms. The most reasonable explanation for the predominance of socially beneficial norms is weak group selection: societies that promote social norms have higher survival rates than societies that do not.

Weak group selection is sufficient for the proliferation of socially beneficial norms as long as the conditions for the stability of the altruist equilibrium (9.12) are met. Altruists in groups at or near this equilibrium will be as fit as other members of their groups and will therefore not suffer adverse within-group selection. But, the fitness of all members of groups at or near the altruistic equilibrium will exceed that of members of groups that support group-harmful norms. The evolutionary dynamic is thus an equilibrium selection problem with differential group survival favoring the selection of the altruistic equilibrium.

The question of interest, then, is whether the updating system captured by our vertical, oblique and payoff based transmission is likely to evolve such that the condition for the stability of the altruistic equilibrium (9.12) will be satisfied. If groups with strong systems of oblique transmission (i.e., high levels of γ) were to do poorly for some reason, then 9.12 might not be satisfied in a long-term evolutionary dynamic. It is also of interest to determine if an initially rare altruistic trait can proliferate in a reasonable time frame, and if it is sustained in a stochastic environment.

To explore these questions we created an agent-based model of society with the following characteristics (the specific assumptions made are not critical, unless otherwise noted). The society consists of 2500 groups, each initially comprising 12 members (the typical size of a Pleistocene hunter-gatherer

Simulation Parameter	Value	Description
Initial Frequency of aAC	2%	
Initial Frequency of aAD	1%	
Initial Frequency of aSC	1%	
Initial Frequency of aSD	1%	
Initial Frequency of bSD	95%	
s	0.03	Fitness cost of altruism
t	0.06	Gain from internalizing fitness-enhancing norms
u	0.01	Fitness cost of internal- ization physiology
Initial Range of γ	[0,0.9]	Rate of oblique transmission
Initial Range of η	[0,0.9]	Rate of imitating selfish types
Initial Group Size	12	
Conflict Rate	10%	
Cost of γ	$5s$	
Cost of η	$5s$	
Fitness contribution of altruist to group	0.05	
Mutation Rate	.01%	
Migration Rate	25%	
Number of Groups	2500	

Table 9.5. Parameters for the Simulation of the Spread of Strong Reciprocity Through Weak Group Selection. $[a, b]$ signifies the initial seeding of the groups with values drawn from the uniform distribution on $[a, b]$. The values of s , t , u , as well as the fitness contribution of altruists, the mutation and migration rates are the same and unchanging for all groups and all generations.

group), arranged spatially on a torus (a 50×50 grid with the opposite edges identified). Each group was seeded with 2% aAC types, 1% aAD types, 1% aSC types, 1% aSD types, and 95% bSD types. Table 9.5 summarizes the parameter choices of the simulation. In all groups, $s = 0.03$, $t = 0.06$ and $u = 0.01$, so the net fitness advantage of the internalization allele is about $0.06 - 0.03 - 0.01 = 0.02$. We take t and u to be constant across groups because they represent individual-level costs and benefits unrelated to social structure. We take s as constant because we are not concerned with

the obvious point that groups with higher s will be disadvantaged. We also fixed the benefit of altruism a 0.05 for all groups; i.e., a group of all altruists has a 5% fitness advantage over a group of all non-altruists. By contrast, the extent γ of oblique transmission is clearly a socially-determined variable, societies with higher γ according more social influence to altruistic elders. We set the cost of per altruist of γ to be $s(\gamma)$; i.e. setting $\gamma = 0.80$ in a group is equivalent to raising the fitness cost to altruists by 0.8s. We found that this cost is inversely related to the long-run value of γ , as one might expect. The level η of regression from altruism to self-interest is also socially determined, in particular by the rate at which altruists encounter non-altruists, and the effect of the higher payoff to non-altruists on updating. We also set the cost of per altruist of η to be $s(1 - \eta)$; i.e. setting $\eta = 0.20$ in a group is equivalent to raising the fitness cost to altruists by 0.8s.

Each group was randomly assigned a value of γ (level of oblique transmission), and a value of η (strength of movement from altruism to self-interest). The migration rate was set to 25% per generation, and the mutation rate was set to 0.01% per generation. Mutation allowed η , and γ to increase or decrease by 1% of their values. The migration rate is very high, of course, and it was local migration on the grid (i.e., migration was always to a neighboring group), individuals taking their phenotypic traits with them. A migration rate one-fifth this size would doom genetic group selection, but gene-culture coevolution can handle very high migration rates.

In each generation, for each of the 2500 groups, we simulated the model as described in the previous sections, and update the frequencies of the various types in each group, according to the fitness effect of their A phenotype and the fraction of the group that exhibits this phenotype. Then, 25% of individuals in each group migrated randomly to neighboring groups, bringing their phenogenotype with them. Selection among groups takes two forms in this model. First, group size drops below a minimum (set to one third of initial group size, or four), it is replaced by a copy of the neighboring group that has the highest average fitness of group members. Second, with a 1 probability each generation, a groups enters into conflict with another randomly chosen group. The group with higher fitness prevails, and members of the losing group copy the parameters of members of the winning group (fraction of altruists and rate of oblique transmission).

We ran this model many times with varying numbers of generations, and varying the parameters described above. The system always stabilized rapidly, there is virtually no variation in final values across runs, the specific

assumptions concerning the parameters move in the intuitively expected direction, and were never critical. The parameter values always allow zero altruism to be a stable evolutionary equilibrium, but with as little as 2% initial altruists, altruism always stabilized at a high level. A typical run with the parameters described above is exhibited in Figure 9.1. There is always strong selection on the rate of oblique transmission, unless the cost of maintaining γ at a high level is extremely high (about $10s$). Selection for lower η is also quite strong, so a cost of $5s$ is needed to prevent η from falling to very low levels in the long run.

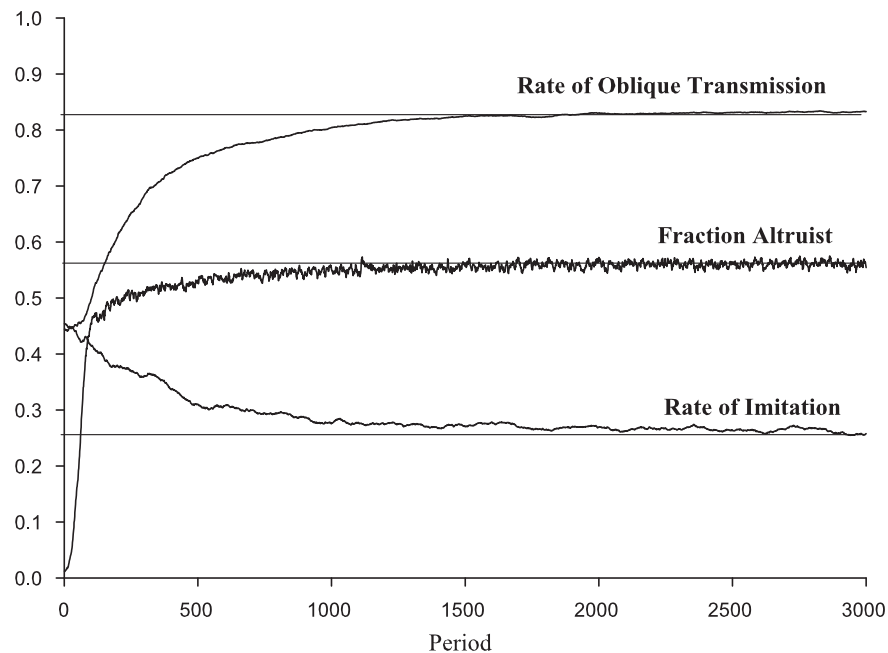


Figure 9.1. The Evolution of Endogenous Parameters. The Rate of Oblique Transmission refers to γ , and the rate of imitation to η in the model.

Figure 9.1 shows the evolution of the endogenous parameters in this simulation. The fraction of altruists then rapidly increases to about 57% by the end of the run. This value varies between 50% and 75%, depending on the costs, borne by altruists alone, for maintaining a high γ and a low η . It is clear that all three parameters of the model undergo strong selection, γ rising to 0.083, and η falling to 0.26.

These simulations thus support the basic arguments of this chapter. In particular, a high level of migration does not undermine the altruistic equilibrium, because most of the effects occur on the cultural rather than the genetic level. Moreover, plausible patterns of population growth and migration account for the sociality of the altruism phenotype A. The critical assumption that drives the model is simply that there is a fitness-enhancing effect of the selfish phenotypic norm sufficiently strong to ensure that C can invade a population of D individuals. The ability of the altruism phenotype A to “hitchhike” on C is quite robust.

9.8 Fitness and Internalization

The internalization of norms is costly ($u > 0$). A considerable fraction of the total available time of the members of most societies is spent teaching the young how to behave, rather than providing for the nutritional and other needs of its members. The capacity to internalize norms is costly to the internalizer as well, given the opportunity cost of the time spent learning norms and the energetic and other trade-offs that constrain the construction of human brains. The above models show that cultural transmission and the capacity to internalize norms may coevolve if some of these norms are fitness-enhancing for the individuals who adopt them. But, if this is the case, what is the evolutionary advantage of taking on the costs of socialization and internalization? Why is genetic transmission not a more economical solution? In many conditions it is, and behaviors that are initially culturally learned may become genetically encoded over long periods of time. But cultural transmission and internalization retain distinct advantages that sometimes justify their costs.

Like other animals, our body produces the sensations of pleasure and pain in response to the things we experience, and this is what induces our behavior. These hedonic responses that constitute the proximate causes of behavior can be represented as what we have in Section 3.1 defined as *preferences*: reasons for behavior, other than beliefs and capacities, that account for the actions an individual takes in a given situation. These preferences are subject to natural selection, and there is some reason to think that, for most animals most of the time, preferences induce behavior approximating that would result if the individual animal were to deliberately maximize its fitness, at least locally.

Cultural transmission and internalization make humans an exception to this general proposition. Cultural transmission and internalization affect our hedonic responses to situations and induce behaviors that may diverge substantially and systematically from what an individual fitness maximizer would do. As we saw in the Introduction to this chapter, individual and even average fitness-reducing behaviors can be successfully promoted by cultural transmission and internalization. But the internalization of culturally transmitted norms can also do better than natural selection in inducing behaviors that enhance fitness. This is true for two reasons.

First, except under special circumstances, individual fitness maximization does not maximize average fitness of the members of a group. The elimination of altruistic individuals from a random mixing population as a result of a fitness-based payoff-based updating dynamic is a pertinent example. Other examples were studied in Chapters 6 to 8. This being the case, groups that override individual fitness-maximizing by means of the cultural transmission of internalized norms may experience higher group average fitness than other groups. These group benefits may offset the costs just mentioned. Indeed, this is one of the key dynamics accounting for the emergence of altruism in the above models, and of social preferences in general.

The second reason why internalization may be fitness enhancing is that, as Boyd and Richerson (2000) have stressed, cultural transmission can respond quickly and flexibly to changed circumstances. By contrast, the response of genetically transmitted traits is much slower, normally requiring hundreds if not thousands of generations or more for any appreciable effect, especially when complex characteristics are involved. It is this aspect of cultural transmission that allowed biologically modern humans originating probably in the Upper Rift Valley in Africa, eventually to occupy virtually all the regions of the world, sustaining life by radically different means in environments varying from arctic desert to tropical rainforest, and to survive the dramatic climactic changes of the Late Pleistocene.

9.9 Conclusion

Social change and the rapid rise in social complexity since the agricultural revolution some 10,000 years ago has been far too swift to permit even the internalization of norms to produce a close fit between preference 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 if either genes or culture, or their interaction, rendered us fitness maximizers, technical advance would have been accompanied by an equivalent increase in the rate of population growth, as Thomas Malthus reasoned. The demographic transition, which has led to dramatically reduced human birth rates throughout most of the world, is a testimonial to our capacity to sever behavior from the logic of fitness maximization (Ihara and Feldman 2004).