10

Socialization

> It is society which, fashioning us in its image, fills us with religious, political and moral beliefs that control our actions.
>
> Émile Durkheim, *Suicide* Ch. 3 (1951[1897]) pp. 211–212

> Leges Sine Moribus Vanae
> (Laws without morals are empty)
>
> Horace, *Odes* (c. 24–25 BC) *III.24*

In addition to trial and error experimentation, preferences are acquired by genetic predisposition (e.g., a taste for sweets) and by a social learning process termed cultural transmission from our parents, others elders, and our peers (e.g., a taste for rice over potatoes). As we saw in §2.3, genetic and cultural transmission are in many ways similar, a fact that 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 genetic 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), have physiological correlates, such as elevated testosterone when insulted among males of European origin from the U. S. South (but not the North), 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 command of resources, more elevated social status, or other reasons for 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 what they would get if they acted in some other manner. In the models proposed in Chapters 7, 8, and 9, 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 fitness disadvantages of particular preferences might be overcome. In his book *Sick Societies*, Robert Edgerton (1992) catalogues dozens of examples of culture overriding fitness, all, as the title suggests, with unpleasant consequences. Pre-industrial cities provide an example (Knauft 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 failure, typically not reproducing its own population even among the urban social elites. 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 et al. 1995, Ihara and Feldman 2004).

But if cultural transmission can induce people to limit their fitness by having small families, or to choose a lethal residential environment, it certainly might also overcome the payoff disadvantages associated with altruistic social preferences. It is this possibility that we explore here. The puzzle, of course, is to explain why humans or any other animal would ever develop the capacity to override fitness concerns, for that capacity itself would seem to be doomed by natural selection.

10.1 Cultural Transmission

Cultural transmission overrides fitness when it causes people to want or feel obliged to do things that result in their having fewer surviving offspring or to reduce their inclusive fitness in other ways. 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 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, what Robert Trivers called "moralistic aggression," is a plausible motivation for the strong reciprocators' punishment of defectors in Chapter 9, and group loyalty and out-group 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 et al. 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 result in proximate motives leading people to engage in such lethal practices as cigarette smoking or, in the highlands of New Guinea, consuming the brains of deceased relatives (Cavalli-Sforza and Feldman 1981, Durham 1991, Edgerton 1992). 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 and character virtues 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 preferences 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 Barry Hewlett (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 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 and model how they may interact so as to favor the evolution of other-regarding and ethical preferences. First, we model the *vertical transmission* of traits from parents to offspring. Parental traits that are associated with greater fitness will evolve for the same reason that genes that confer greater fitness enjoy supra-average survival rates. 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 Boyd and Richerson (1985), oblique transmission may be conformist, the young tending to adopt the behaviors most common in the parental generation, independently of their payoffs. In this case the resulting dynamic will not be monotonic in either fitness or well-being. If virtually all of the population is altruistic, conformist cultural 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 (§4.2) may be sufficient to stabilize altruistic preferences.

Conformist cultural transmission may arise for a variety of reasons, ranging from an evolved social learning strategy in which individuals regard the population frequency of a trait as a measure of its desirability, all the way to population-level institutional arrangements for the deliberate socialization of the young, in which the content reflects which types are prevalent in the population. We stress the latter 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.

Why should the norms that are internalized be altruistic? Linnda Caporeal and her coauthors (Caporael et al. 1989) and Herbert Simon (1990) proposed that altruism might proliferate in a population because it is an inseparable part of an 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 puzzle of how the capacity to internalize norms evolves. Second, we would like to explain the "pleiotropic analogy" whereby individually costly altruism and individually beneficial other norms are inseparable, with a model in which norms are explicitly cultural, phenotypic expressions of behavior.

These two challenges lead us to model explicitly the interplay between the genetic predisposition to internalize norms and the nature of the norms thereby engendered. In §10.2, we develop a purely phenotypic model in which, as a result of the effectiveness of socialization, a fitness-reducing norm, whether it be smoking or contributing to the public good, may be maintained in a population. Critical to this result is the effec-

tiveness of schools, story telling, and other socialization agents, which in turn depends upon our capacity to internalize norms, and our receptivity to socialization. In §10.3, we therefore model the genetic basis of internalization and give the conditions under which there population equilibrium in which individuals have an "internalization allele" and acquire a fitness-enhancing norm. In §10.4, and §10.5 we reintroduce the fitness-reducing norm of §10.2 into the model of §10.3, and study the conditions under which it can "hitchhike" on the internalization allele to form a stable population equilibrium in which all individuals express both a fitness-enhancing and a fitness-reducing norm. This will turn out to depend critically on the effectiveness of the socialization process compared to the strength of selection against the fitness-reducing norm, much as in the model of §10.2.

Finally, in §10.6 we explain why the individually fitness-reducing norm will generally tend to enhance the average fitness of group members, and hence will be altruistic. This occurs because groups with social norms that enhance the fitness of their members will outcompete groups that foster norms that are both costly to their bearers and of zero or negative fitness benefit to the group. Because the fitness-reducing norm can be maintained in a population (§10.4) weak multi-level selection (§4.2) is sufficient to guarantee this result. This is why, notwithstanding the evidence provided by Edgerton, institutions of socialization tend to favor prosocial preferences.

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 §10.7 is an application of the reasoning of Boyd and Richerson (2000), extending the explanation given in Chapter 2: the cultural transmission of preferences allowed humans, exceptionally 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.

10.2 Socialization and the Survival of Fitness-Reducing Norms

Consider a group in which members can either adopt, or not, a certain cultural norm A. We shall call those who adopt the norm A-types, and we call those who do not adopt the norm S-types. Adopting A is costly, in that S-types have fitness 1, as compared with A-types, who have fitness $1-s$, where $0 < s < 1$ is a fitness loss. The A norm, despite its notation, need not be altruistic; we will later investigate the conditions under which this could be the case. What matters for now is that a person switching from S to A incurs a fitness loss. We assume that in each generation 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 (we call this *vertical transmission*). Oblique cultural transmission also takes place because the S-type offspring of AS- and SS-families are susceptible to influence by socialization institutions promoting the A norm. As a result, offspring of AA parents are A-types, offspring of SS parents are S-types, and half of the offspring of AS-families (which are the same as SA-families) are A-types, the other half S-types. Table 10.7 summarizes the mathematical symbols used in this chapter.

Socialization occurs when an S offspring encounters a "cultural model" randomly drawn from the population, which occurs once for every member of each generation. If the model is an A, which occurs with probability p, the offspring switches to being an A with a probability $\gamma > 0$.

Combining oblique and vertical transmission, we find that the change in the fraction of A-types in the next generation is given by the familiar replicator equation (see §A5):

$$\Delta p = p(1-p)\frac{\gamma - s}{1 - sp}, \tag{10.1}$$

where p is the frequency of A's in the population and Δp is its change over some discrete time period. The term $1 - sp$ is the average payoff in the population, γ is the rate of oblique transmission, and $\gamma - s$ is the selective advantage (disadvantage if negative) of the A's over the S's when account is taken of both oblique and vertical transmission. Equation 10.1 illustrates the tension between the differential fitness effects on the evolution of p captured by s that work against the evolution of the A's and the effects of oblique transmission captured by γ, which tend to counteract the selection against A-types. Equation 10.1 shows that when $s = \gamma$, these two effects are exactly offsetting, and the population frequency of A-types will be stationary ($\Delta p = 0$).

Payoff-based updating then occurs. Each group member i observes the fitness and the type of a randomly chosen other member j, and may change 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 that an A individual will shift to S is η times the fitness difference of the two types, where $\eta > 0$. The term η represents the power of payoff differences to induce changes in type, and this, naturally will play a big role in our account.

The expected fraction p' of the population that are A's after the above payoff-based updating is the fraction before updating p, minus those A's who switched to S, the latter being the A's who observed an S (a fraction $p(1-p)$ of the population), multiplied by the probability of a switch taking place in these cases. Thus we have

$$p' = p - \eta s p(1-p). \tag{10.2}$$

We now combine these three sources of change in the fraction of A-types, adding the changes described in equation 10.2 to those already shown in 10.1, giving

$$\Delta p = p(1-p)\frac{\gamma - s}{1 - sp} - \eta p(1-p)s \tag{10.3}$$

The second term on the right hand side 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 or not, depending on whether $\gamma > s$ or $\gamma < s$.

Not surprisingly, the higher the personal cost of altruistic behavior, the more stringent the conditions under which the A norm will emerge, illustrating 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 evident from the conditions under which the all-A equilibrium is *globally stable*, meaning that starting from any of the possible states of the population, the population dynamic will move to the all-A equilibrium. In order for this to be the case, the strength of payoff-based updating η must be less than the difference in the size of the oblique transmission and the fitness cost of the A norm, normalized by the latter:

$$\eta < \frac{\gamma - s}{s}. \tag{10.4}$$

However, if

$$\frac{\gamma - s}{s} < \eta < \frac{\gamma - s}{s(1-s)}, \tag{10.5}$$

both the all S-type equilibrium and the A-type equilibria are locally stable, meaning that there exists a neighborhood of states around these two equilibrium states such that if the equilibrium state is displaced to some state in this neighborhood, the population dynamic will return to the equilibrium. The basin of attraction of the A-type equilibrium, that is, the neighborhood of states from which the dynamic will converge to the all-A equilibrium, shrinks as η increases. Finally, if

$$\eta > \frac{\gamma - s}{s(1-s)}, \tag{10.6}$$

the all-S equilibrium is globally stable.

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

10.3 Genes, Culture, and the Internalization of Norms

But why would people, or any animal, internalize norms if taking a norm on board leads one to act in ways that reduce fitness? We will answer this in two steps. Here we explain why the capacity to internalize fitness-enhancing norms, those that correct for human impatience or weakness of will, for example, might evolve even if the capacity to internalize is costly. In the next section we will show that when the capacity to internalize a norm has evolved and societies have developed socialization practices to do this, people will be susceptible to internalizing norms that also reduce fitness, such as the A norm of the previous section. This is what we mean when we say that a fitness-reducing norm can *hitchhike* on a process of norm internalization that has evolved due to the existence of an individually fitness-enhancing norm.

Here we assume that cultural traits are acquired through vertical transmission alone. Oblique transmission and the payoff-based switching of traits, as modeled in §10.2, will be reintroduced in §10.4 and §10.5.

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" with, as usual, the quotation marks serving as a reminder that this simple genotype-phenotype mapping is a considerable simplification. We will assume that each individual has only one copy at this locus (i.e., genetics are haploid), which is inherited with equal probability from either parent. An alternative diploid model, in which each locus has two alleles, has almost the same properties as the haploid model, but is much more complicated, and is developed in full in Gintis (2003a). Individuals without the allele cannot internalize norms, whereas individuals with the 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, including which cultural models he has encountered. 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, and hence can proliferate when rare.

Suppose the norm in question is C (Cleanliness, for instance), which confers fitness $1 + f > 1$, 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 $(1 + f)(1 - u) > 1$, 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 by his phenotype (whether he is a C or a D). There are thus three "phenogenotypes," whose fitnesses are shown in Table 10.1.

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

Table 10.1. Fitnesses of the three phenogenotypes. Note: Here u is the fitness cost of possessing the internalization allele, and f is the fitness value of possessing the norm C; 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 a familial phenogenotype is xyXY, where x and y can be either a or b, and X and Y can be either C or 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 10.2, where $\beta \in [0, 1]$ measures the strength of the cultural transmission of C. We assume unbiased cultural transmission ($\beta = 1/2$) unless otherwise stated.

	Offspring Phenogenotypic Frequency		
Familial Type	aC	aD	bD
aaCC	1		
aaCD	β	$1-\beta$	
aaDD		1	
abCD	$\beta/2$	$(1-\beta)/2$	1/2
abDD		1/2	1/2
bbDD			1

Table 10.2. Phenotypic inheritance is controlled by genotype. Note: abCC and bbCC, and bbCD are not listed because bC cannot occur, as an individual must have the a allele to internalize the C norm. Note that $\beta \in [0, 1]$ measures the strength of the cultural transmission of C.

Families are formed, as before, 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 (i.e., an offspring is equally likely to inherit a gene from either parent). 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 nine familial phenogenotypes, of which only six can occur (because a parent of genotype b must have the D phenotype). The frequencies of the offspring of different familial phenogenotypes are as shown in Table 10.3, where $P(i)$ represents the frequency of parental phenogenotype $i = a\text{C}, a\text{D}, b\text{D}$. For example, the aaCD phenogenotype can occur in two ways: father aC and mother aD, or vice-versa. The probability of each occurrence is $P(a\text{C})P(a\text{D})$. The fitness of this phenogenotype is $(1-u)^2(1+f)$ because both parents have the a allele at fitness cost u, and one has the C trait, at fitness gain f. The share of the next generation total population constituted by the offspring of this phenogenotype is thus as given in the second row of Table 10.3.

Phenogenotype	*Frequency*
aaCC	$P(a\text{C})^2(1-u)^2(1+f)^2\beta_o^2$
aaCD	$2P(a\text{C})P(a\text{D})(1-u)^2(1+f)\beta_o^2$
aaDD	$P(a\text{D})^2(1-u)^2\beta_o^2$
abCD	$2P(a\text{C})P(bD)(1-u)(1+f)\beta_o^2$
abDD	$2P(a\text{D})P(b\text{D})(1-u)\beta_o^2$
bbDD	$P(b\text{D})^2\beta_o^2$

Table 10.3. Frequencies of phenogenotypes. Note: β_o is baseline fitness, chosen so the sum of the frequencies is unity; bCC and bCD are not listed, because bC cannot occur.

Equilibrium occurs when the frequency of each phenogenotype is constant from generation to generation. In this case, we need consider only two of the phenogenotypes, say aC and aD, because bC cannot occur, and since the probabilities must add up to unity, we have $P(b\mathrm{D}) = 1 - P(a\mathrm{C}) - P(a\mathrm{D})$. 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 internalization allele is present but the phenotype C is absent, and bD, in which neither the internalization allele nor the norm is present.

Elsewhere (Gintis 2003b) we have proven the following assertions concerning the stability of the various equilibria of this system. The aD equilibrium is unstable, while the aC equilibrium is locally stable, meaning the system will return to this equilibrium starting from nearby states (§A4). The unnormed equilibrium bD is locally stable if $(1 - u)(1 + f) < 2$ and unstable when the opposite inequality holds. Either of two conditions renders the bD equilibrium unstable, in which case aC, in which all individuals internalize the fitness-enhancing norm, will be globally stable, which means the system moves to this equilibrium from any starting point. The first is that $(1-u)(1+f) > 2$. The second condition is that the cultural bias transmission coefficient β is sufficiently greater than 1/2. We consider the former condition implausible because it requires that $f > 1$, whereas positive fitness coefficients are rarely so large. However, the latter condition is quite plausible, because it may take only one parent to instill a norm in all offspring with high probability ("Mom taught me to be clean. Dad was always a slob"). Note that biased vertical transmission, $\beta > 1/2$, produces the same effect as oblique transmission, $\gamma > 0$, in the previous section.

10.4 The Internalized Norm as Hitchhiker

We now add a second phenotypic trait with two variants. Internalized norm A is promulgated by the group but imposes fitness cost s, with $0 < s < 1$, on those who adopt it. The normless state, S, is neutral, imposing no fitness cost on those who adopt it. An individual phenotype is then one of SD (internalizes neither norm), SC (internalizes only the fitness-enhancing norm), AD (internalizes only the fitness-reducing norm A), and AC (internalizes both the fitness-enhancing and fitness-reducing 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 because b individuals must be normless, i.e., SD. These three are bAC, bAD, and bSC. We represent the frequency of phenogenotype i by $P(i)$, for $i = a\mathrm{AC}, \ldots, \mathrm{SD}$.

As before, families are formed by random pairing and the offspring genotype obeys Mendelian segregation (an offspring is equally likely to inherit a gene from either parent) . As above, 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 famil-

Phenogenotype	*Frequency*
P(aaAACC)	P(aAC)$^2(1-u)^2(1+f)^2(1-s)^2\beta_o^2$,
P(aaAACD)	2P(aAC)P(aAD)$(1-u)^2(1-s)^2(1+f)\beta_o^2$,
P(abASCD)	2P(2aAC)P(bSD)$(1-u)(1+f)(1-s)\beta_o^2$,
P(bbSSDD)	P(bSD)$^2\beta_o^2$.

Table 10.4. Selected phenogenotypic frequencies. Note: 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 $(1-s)(1+f)(1-u)$ and the other with fitness 1.

ial phenogenotypes, which we can write as aaAACC,...,bbSSDD, only 14 of which can occur. For instance, aaAACC represents the case where both parents have the internalization allele a, and both parents internalize the fitness-reducing and the fitness-enhancing norm. Similarly, aaAACD represents the case where both parents have the internalization allele a, and both parents internalize the fitness-reducing norm A, but only one internalizes the fitness-enhancing norm C. Finally, abASCD represents the case where one parent carries the internalization norm and the other does not, the former internalizing both the fitness-reducing norm A and the fitness-enhancing norm C. 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 illustrative purposes, a few of the phenogenotypic frequencies are shown in Table 10.4.

The rules of cultural transmission are as before. If the familial phenogenotype is xyXYZW, where x and y are either a or b, X and Y are either A or S, and Z and W are either C or 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. Oblique cultural transmission occurs when an a individual with S phenotype, genetically capable of internalizing but culturally selfish, adopts the A phenotype in response not to parental socialization but to learning from other A-types in the population. This occurs more frequently the more A-types there are in the population (p) and the more effective are the society's institutions (deliberate or otherwise) for oblique transmission (γ), each aS individual switching at the rate γp, so that the gain in A phenotypes by this mechanism is $\gamma p(1-p)$, where $1-p$ is the frequency of aS-types in the population. Note that oblique transmission in this model is asymmetric: if there are A-types in the population, S-types may learn to become A-types, not the other way around, even if the population is predominantly of the S-type.

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 f from holding norm C, and a fitness loss s from holding the A norm. The individual's resulting fitness is then $(1-u)(1+f)(1-s)$. In the absence of positive assortment, $(1-u)(1+f)(1-s) > 1$ is a necessary condition for the fitness-reducing norm to evolve, so we assume this inequality holds; i.e., the direct individual fitness benefit due to having phenotype C must be sufficient to offset both the cost of having the internalization allele and the cost of expressing the fitness-reducing

norm. The fitness of the phenogenotypes that can occur with positive frequency are as shown in Table 10.5.

Phenogenotype	*Fitness*
aAC	$(1-u)(1-s)(1+f)$
aAD	$(1-u)(1-s)$
aSC	$(1-u)(1+f)$
aSD	$(1-u)$
bSD	1

Table 10.5. Fitnesses of five phenogenotypes.

The fitness of these phenotypes, along with the rules of genetic and cultural transmission given above, allow us to determine for any combination of frequencies of the phenogenotypes in Table 10.5 the change in frequencies that will occur as a result of the combined impact of genetic and cultural transmission. The population is in equilibrium when the frequency of each phenogenotype is constant from generation to generation. We can determine the possible population equilibria using 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 show that there are five equilibria, in which the whole population bears a single phenogenotype. These are aAC, in which all individuals internalize both the fitness-reducing and fitness-enhancing norms, aAD, in which only the fitness-reducing 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 the fitness-reducing nor the fitness-enhancing norm is transmitted from parents to offspring. But the aAD and the aSD equilibria are unstable, and hence will not survive an evolutionary process, so we can ignore them.

The analysis of the stability of the remaining equilibria, aAC, aSC, and bSD, is given in Gintis (2003a). The two a equilibria are stable when $s < \gamma$. This inequality expresses the key condition that the fitness-reducing norm cannot be evolutionarily stable unless the effectiveness of oblique transmission is sufficient to overcome the fitness cost of expressing the fitness-reducing norm. Groups with high levels of fitness-reducing norm expression solve the problem of rendering the fitness-reducing norm stable by increasing the effectiveness of oblique transmission so that the new converts to fitness-reducing norms compensate for the lower fitness of A-types.

It is no surprise, therefore, that the aSC equilibrium, in which internalization is possible but the fitness-reducing norm is not internalized, 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 $(1-u)(1+f) > 2$, which is highly unlikely, as we explained above. There are two reasons why the equilibria aSC, aAC, and bSD, all homogeneous populations with a single type, are stable. First, there are positive feedbacks in the oblique transmission process by which individuals are socialized, such that it is inoperative when the internalization gene is absent from the population, and may be powerful enough to offset

the fitness disadvantages of the A-types when the internalization gene is universally expressed. This explains why a stable equilibrium population is either all S or all A. Second the bSD ("no internalization, no norms") equilibrium is stable in that a and C are complements, meaning that in the absence of C, a cannot proliferate when rare, and conversely. We have not determined if stable mixed strategy equilibria exist, but for the above reasons we doubt that they could.

This analysis shows that if $s < \gamma$, the fitness-disadvantaged phenotype A coexists in a stable equilibrium with the fitness-enhancing phenotype C. We say that A hitchhikes on C because the fitness value of C renders the internalization allele a evolutionarily viable, and once this allele occurs in high frequency, the normed phenotype A is evolutionarily viable because its fitness cost s is less than the oblique transmission effect γ, which favors A.

10.5 The Gene-Culture Coevolution of a Fitness-Reducing Norm

To simplify the gene-culture interaction, the analysis of the previous section did not include an obvious challenge to the fitness-reducing norm A: when people update their behaviors they not only do so under the influence of schools, elders and the other bearers of oblique transmission, they also pay attention to the payoffs that they and people of different types are receiving, and this must disadvantage the A-types. We now add the payoff-based updating dynamic developed in §10.2 to our gene-culture 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 §10.2. In the current context, there are four phenotypes, and only a individuals will copy another phenotype, because they are the only type capable of internalizing a norm, and noninternalizers will not desire to mimic internalizers.

Let XY and WZ be two of the phenotypes AC, AD, SC, SD. We assume an a individual with 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. Thus, as in §10.2, the parameter η is the a measure of the strength of the tendency to shift from lower to higher payoff phenotypes.

Adding payoff-based cultural updating does not change the single phenogenotype equilibria, because when all equilibria consist of a single phenogenotype, in equilibrium an individual can never meet a distinct phenotype to which he might switch. We find that the aAD and aSD equilibria remain unstable, and payoff-based updating does not affect the conditions for stability of the normless equilibrium bSD. The condition $\gamma > s$ for stability of the fitness-reducing norm equilibrium aAC now becomes

$$\eta < \frac{\gamma - s}{1 - \gamma}\left(\frac{1}{s} - 1\right). \tag{10.7}$$

Note the similarity to the all-A equilibrium conditions (10.4–10.6) 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, even if the effect of socialization exceeds the fitness cost of the A-type. The condition $s > \gamma$ for stabil-

ity of the fitness-enhancing norm 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 $\gamma > s$.

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 fitness-enhancing internalized equilibrium aSC, rather than the equilibrium with both norms, aAC, stable. The intuition here is that the fitness-reducing norm A imposes a fitness cost s leading individuals to abandon it. The greater the rate at which this occurs, the larger must be the oblique socialization force γ that replenishes the stock of A-types in the group.

10.6 How Can Internalized Norms Be Altruistic?

As we have seen, internalized norms may reduce the fitness of group members. 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 to the individual (s), given the strength of the payoff-based updating process (η). 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 populations that persist over long periods for cultural institutions to promote social and eschew antisocial norms, and for A-types 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 and reproduction rates than societies that do not.

Weak group selection (§4.2) is sufficient for the proliferation of socially beneficial norms as long as the conditions for the stability of the equilibrium with the fitness-reducing norm (10.7) are met. A-types in groups at or near such an equilibrium if A is altruistic 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 itself likely to evolve such that the condition for the stability of the altruistic equilibrium (10.7) 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 (10.7) might not be satisfied in a long-term evolutionary dynamic. Recall that in Chapters 7 and 8 we asked a similar question. Having shown that culturally transmitted reproductive leveling and within-group segmentation practices

favor the evolution of a genetically transmitted altruistic predisposition (Chapter 7) and that intergroup hostilities are essential to this process (Chapter 8), we asked if these altruism-favoring conditions themselves could evolve. Here, instead, we explore the coevolution of three distinct aspects of a population: the distributions of its genotypes and phenotypes and the evolution of the process by which individuals update their socially learned traits. The third will require an exploration of the dynamics of γ, the effectiveness of its institutions of socialization, and η, the effect of payoff differences in inducing individuals to switch from altruist to selfish types. As we did in Chapters 7, 8, and 9, we will also determine if an initially rare altruistic trait can proliferate in a reasonable time frame, and if it is sustained in a stochastic environment.

Given the complexity of this task, selection on genes, learned behavior, and two aspects of a society's social learning system operating at both the individual and group level, we are not able to develop an illuminating analytical model, and so, as in previous chapters, 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 1000 groups, each initially comprising 12 members per generation, or a census size of 36, about the size of a Pleistocene hunter-gatherer group, arranged spatially on a torus (a 50×50 inner-tube-type grid with the opposite edges identified). Each group started with 2% *a*AC-types, 1% *a*AD-types, 1% *a*SC-types, 1% *a*SD-types, and 95% *b*SD-types. Table 10.6 summarizes the parameter choices of the simulation. We let $s = 0.03$, $f = 0.06$ and $u = 0.01$, common across all groups, because they represent individual-level costs and benefits unrelated to any group differences in 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, corresponding to β_G in §4.2, at 0.05 for all groups; i.e., a group of all A-types 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 A-type elders. Similarly the strength of payoff-based updating may vary across groups and over time. Each group initially was randomly assigned a value of γ and a value of η. Random variation in social learning arrangements ("institutional mutation") allowed η and γ to increase or decrease by 1% of their values. The migration rate was set to 25% per generation (very high for a genetic model but reasonable for a cultural model), and the mutation rate was set to 0.01% per generation, and migration was always to a neighboring group, individuals taking their phenotypic traits with them. As in Chapter 7, we assume that institutions are not free goods. In this case a more effective socialization system (greater γ) comes at the price of a larger fitness disadvantage for the A-types. The time they spend teaching altruistic behavior, for example, they cannot be seeking out mating opportunities and caring for their offspring.

We set the cost per A-type of γ to be $s\gamma$; i.e., setting $\gamma = 0.80$ in a group is equivalent to raising the fitness cost to A-types by 0.8s. We found in the simulations that s is inversely related to the long-run value of γ, as one might expect. The level of η, the lure of higher payoffs in motivating the regression from altruism to self-interest, is also socially determined. A-types, whose numbers are reduced by desertion to self-interest when η is substantial, can devote time and energy to reducing the lure of payoffs, teaching, for example, the value of non-material well-being. To reflect his

Simulation Parameter	*Value*
Initial frequency of aAC	2%
Initial frequency of aAD,aSC,aSD	1%
Initial frequency of bSD	95%
Fitness cost of altruism s	0.03
Gain from internalizing fitness-enhancing norms f	0.06
Fitness cost of internalization physiology u	0.01
Initial range of rate of oblique transmission γ	[0,0.9]
Initial range of imitation rate η	[0,0.9]
Initial group size	12
Conflict rate	10%
Cost of γ	$5s$
Cost of η	$5s$
Fitness contribution of A-type to group	0.05
Mutation rate	0.01%
Migration rate	25%
Number of groups	1000

Table 10.6. 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, f, u, as well as the fitness contribution of A-types and the mutation and migration rates are the same and unchanging for all groups and all generations.

we imposed a cost of $s(1 - \eta)$ on the A-types. Thus setting $\eta = 0.20$ in a group is equivalent to raising the fitness cost to A-types by $0.8s$.

In each generation, for each of the groups, we simulated the theoretical model as described in the previous sections and updated 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 a randomly selected 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, if 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 small probability for each generation, a group enters into conflict with another randomly chosen group. The group with higher fitness prevails, and members of the losing group copy the group-specific parameters of members of the winning group.

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 initial conditions were never critical. The parameter values always allow zero altruism to be a stable evolutionary equilibrium, but with as few as 2% initial A-types, altruism always stabilized at a high level. A run with the parameters described above is exhibited in Figure 10.1. There is always strong

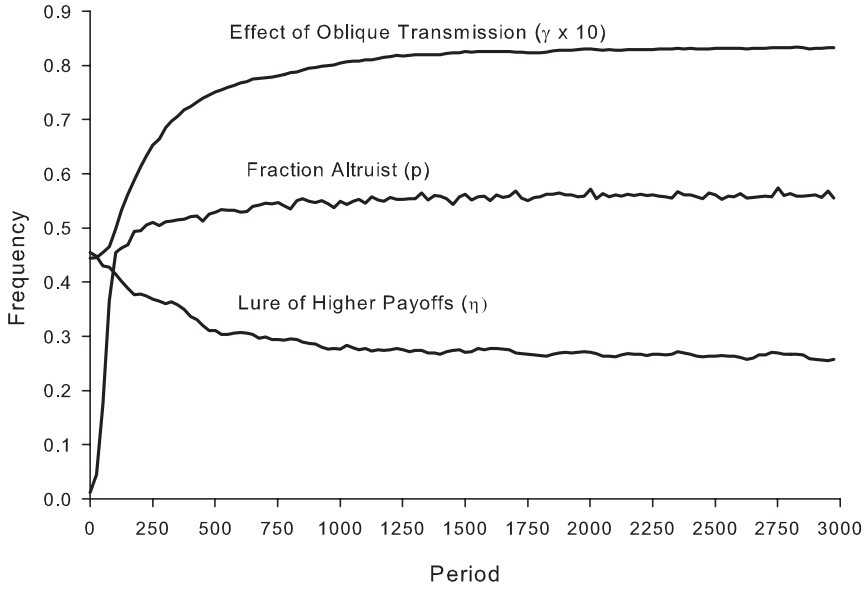

Figure 10.1. The evolution of endogenous parameters. In this simulation, the steady state fraction of Altruists is $p \approx 0.57$, the effect of oblique transmission stabilizes at $\gamma \approx 0.083$, and the rate of switching from A type to N type is $\eta \approx 0.26$.

selection favoring 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 high cost of reducing it is needed to prevent η from falling to very low levels in the long run.

Figure 10.1 shows the evolution of the endogenous parameters in this simulation. The fraction of A-types increases to about 57% by the end of the run. This value varies between 50% and 75%, depending on the costs, borne by A-types 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 (γ is multiplied by 10 in the figure).

Migration does not undermine the altruistic equilibrium, because most of the effects occur on the cultural rather than the genetic level, and migrants respond to the social learning environment of their new home.

The simulation thus identifies a wide range of parameter values under which a system of cultural transmission biased toward socialization of the young for altruism and minimizing the lure of material payoffs could itself evolve, and if it did that, these social learning arrangements would support a frequency of altruism in the population.

10.7 The Programmable Brain

Vertical, oblique, and payoff-based updating all affect the internalization of norms. Taking on a general rule of behavior as an objective rather than a constraint or an in-

strument toward some other end is likely to be costly for two reasons. First, a considerable fraction of the total available time of the members of most societies is spent teaching the young the proper way to behave, rather than providing for the nutritional and other needs of its members. But in addition to the cost of acquiring such a norm ($u > 0$), there is a further cost: 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 studied in Chapter 8 ("don't marry outside your religion") is an example of a personally costly general rule of behavior—costly because it reduces the size of the marriage pool.

Why, then, are humans so susceptible to internalizing general rules? If this susceptibility were subject to a purely payoff-based selection process, whether fitness- or payoff-sensitive, one might expect it to be eliminated from any population in which it appeared. What, then, accounts for the extraordinary success of general rules of behavior? An answer that we have found 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. A similar argument led John Stuart Mill to remark, "Being rational creatures [sailors] go to sea with it [the Nautical Almanac] already calculated; and all rational creatures go out upon the sea of life with their minds made up on the common questions of right and wrong, as well as on many of the far more difficult questions of wise and foolish" (1957[1861], p. 407).

Our 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?

Like those of other animals, our bodies produce 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 in Chapter 3 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, as well as social learning in some animals, and there is some reason to think that, for most animals most of the time, preferences induce behavior approximating that which 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 impossibility of altruism evolving by a fitness-monotone dynamic in a random mixing population is a pertinent example. Other examples were studied in Chapters 7 and 9. 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.

In our model of socialization, oblique transmission converts a fraction of self-regarding types into altruists. But we did not ask about the proximate motives for the altruists helping others. Does oblique transmission work by teaching children the golden rule or Kant's categorical imperative? By warning them that God may be watching?

These and other cognitive reasons for good behavior are no doubt involved, but the motivation to help others and to act ethically often short-circuits these reflective processes in favor of more visceral influences on behavior such as anger, shame, elation and guilt. To readers who share our horror of road rage and honor killings, the claim that visceral reactions are among the proximate motives for generous, fair-minded and civic actions may seem surprising. But it is true, and we think that a good case can be made that the social emotions evolved precisely because they motivated prosocial actions.

Symbol	Meaning
β	Bias of vertical transmission
β_o	Baseline fitness
η	Imitation rate
f	Fitness gain from C phenotype
γ	Rate of oblique transmission
γ_i	Strength of i's moral standard
λ_i	Strength of i's reciprocity motive
μ_{ij}	Punishment of j by i
ν_i	Strength of i's shame
p	Fraction of A's
π_i	Material payoff to i
s	Fitness cost of A phenotype
u	Fitness cost of a allele
τ_r	Degree of reproductive leveling (effective tax rate)
τ	Quorum threshold level
ζ	Segmentation rate

Table 10.7. Definition of symbols

11

Social Emotions

> This is the gist of human psychology. . . what the hero does all feel that they ought to have done as well. The sophisms of the brain cannot resist the mutual aid feeling, because this feeling has been nurtured by thousands of years of human social life and hundreds of thousands of years of prehuman life in societies.
>
> Pyotr Kropotkin, *Mutual Aid* Chapter VIII (1989[1903]) p. 277

> Let's not forget that the little emotions are the great captains of our lives and we obey them without realizing it.
>
> Vincent Van Gogh, *Letter to his brother Theo* Letter 603 (July 6, 1889)

> The heart has reasons that reason knows nothing about.
>
> Blaise Pascal, *Pensées* Number 277 (1995[1670])

Social emotions—love, guilt, shame, and others—are responsible for the host of civil and caring acts that enrich our daily lives and render living, working, shopping, traveling among strangers, sustaining social order, even conducting scientific research, feasible and pleasant. Adherence to social norms is underwritten not only by cognitively mediated decisions, but also by emotions (Frank, 1987, 1988; Ekman, 1992; Damasio, 1994; Elster, 1998; Boehm 2007). When Bosman et al. (2001) assayed the feelings of respondents in an ultimatum game, they found that low offers by the proposer provoked anger, contempt and sadness in the respondents, that the intensity of the self-reported emotions predicted the respondents' behavior, stronger emotions inducing rejections of low offers. Interestingly, the introduction of an hour-long "cooling-off" period between the offer and the respondent's choice of an action had no effect on either reported emotions or on the rejection behaviors of the respondents. Recall from Chapter 3 that Sanfey et al. (2003) found that those rejecting low offers in an ultimatum game experienced heightened levels of activation in the brain areas associated with disgust and anger.

One of the most important emotions sustaining cooperation is shame, the feeling of discomfort at having done something wrong not only by one's own norms but also in the eyes of those whose opinions matter to you. Shame differs from guilt in that, while both involve the violation of a norm, the former but not the latter is necessarily induced by others' knowing about the violation and making their displeasure known to the violator.

We will suggest that shame, guilt, and other social emotions may function like pain, in providing personally beneficial guides for action that bypass the explicit cognitive optimizing process that lies at the core of the standard behavioral model in economics and decision theory. Pain is one of the six so-called basic emotions, the others being pleasure, anger, fear, surprise, and disgust. Shame is one of the seven so-called social emotions, of which the others are love, guilt, embarrassment, pride, envy, and jealousy (Plutchik 1980, Ekman 1992). Basic and social emotions are expressed in all human societies, although their expression is affected by cultural conditions. For instance, in all societies one may be angered by an immoral act, or disgusted by an unusual foodstuff, but what counts as immoral or disgusting is, at least to some extent, culturally specific.

Antonio Damasio (1994) calls an emotion a "somatic marker," that is, a bodily response that "forces attention on the negative outcome to which a given action may lead and functions as an automated alarm signal which says: Beware of danger ahead if you choose the option that leads to this outcome…the automated signal protects you against future losses" (p. 173). Emotions thus may contribute to the decision-making process by working with, not against, reason. Damasio continues, analogizing emotions to physical pain: "suffering puts us on notice…it increases the probability that individuals will heed pain signals and act to avert their source or correct their consequences" (p. 264).

To explore the role of guilt and shame in inducing social behaviors we will consider a particular interaction having the structure of a public goods game (§3.2). In the public good setting, contributing too little to the public account may evoke shame if one feels that one has appropriated "too much" to oneself. Because shame is socially induced, being punished when one has contributed little triggers the feeling of having taken too much. In this case, the effect of punishment on behavior may not operate by changing the material incentives facing the individual, that is, by making clear that if he continues to free ride his payoffs will be reduced by the expected punishments in future rounds. Rather it evokes a different evaluation by the individual of the act of taking too much, namely, shame. This is the view expressed by Jon Elster (1998)" "material sanctions themselves are best understood as vehicles of the emotion of contempt, which is the direct trigger of shame" (p. 67). Thus, self-interested actions, per se, may induce guilt, but not shame. If one contributes little and is not punished, one comes to consider these actions as unshameful. If, by contrast, one is punished when one has contributed generously, the emotional reaction may be spite toward the members of one's group. This is one of the reasons why the "antisocial" punishment of high contributors in public goods experiments has such deleterious effects on the level of cooperation in a group.

We assume individuals maximize a utility function that includes five distinct motives: one's individual material payoffs, how much one values the payoffs to others, this depending on both one's unconditional altruism and one's degree of reciprocity, as well as one's sense of guilt or shame in response to one's own and others' actions. To this end, we will amend and extend a utility function derived from the work of Geanakoplos et al. (1989), Levine (1998), Sethi and Somanathan (2001), and Falk and Fischbacher (2006).

In Chapter 3, we presented experimental evidence consistent with the view that punishment not only reduces material payoffs of those who transgress norms, but also may recruit emotions of shame toward the modification of behavior. Indeed, we showed in §3.4 that in some societies many defectors react to being punished by increasing their contribution to the group, even when the punishment does not affect material payoffs, consistent with the shame response, while in other societies they react by counter-punishing contributors, consistent with an anger response. Social emotions in response to sanctions can thus either foster or undermine cooperation. Reacting to sanctions, then, is often not a dispassionate calculation of material costs and benefits, but rather involves the deployment of culturally specific social emotions. In Chapter 9 we showed that the altruistic punishment of shirkers by strong reciprocators can proliferate in a population and sustain high levels of cooperation, but we tacitly assumed that those punished would react prosocially rather than antisocially. Here, we focus on the manner in which social emotions and punishment of miscreants may be synergistic, each enhancing the effects of the other.

We first model the process by which an emotion such as shame may affect behavior in a simple public goods game. We then show that shame and guilt along with internalized ethical norms allow high levels of cooperation to be sustained with minimal levels of costly punishment, resulting in mutually beneficial interactions at limited cost. In §11.2, we ask how prosocial emotions such as shame might have evolved.

11.1 Reciprocity, Shame, and Punishment

Consider two individuals who play a one-shot public goods game in which each has a norm concerning the appropriate amount to contribute to the public project, and each (a) values his own material payoff, (b) may prefer to punish others who contribute insufficiently, (c) feels guilt if he contributes less than the norm; and finally (d) experiences shame if he is sanctioned for having contributed less than the norm. This psychological repertoire captures some of the motives that we think explain cooperation in behavioral experiments. The results that follow for a dyadic interaction generalize to an n-person interaction. A summary of the symbols used in this chapter appears in Table 11.1

In what follows, we represent the two players as i and j, where $j \neq i$. We assume each individual starts with a personal account equal to one unit. Each individual contributes to the public project an amount a_i, $0 \leq a_i \leq 1$, where $i = 1, 2$ refer to the two individuals, and each receives $\chi(a_1 + a_2)$ from the project, where $1/2 < \chi < 1$. Thus, the individuals do best when both cooperate ($a_i, a_j = 1$), but each has an incentive to defect ($a_i, a_j = 0$) no matter what the other does. In the absence of punishment, this two-person public goods game thus would be a prisoner's dilemma. But at the end of this cooperation period there is a punishment period, in which each individual is informed of the contribution of the other individual, and each individual may impose a penalty μ on the other individual at a cost

$$c(\mu) = c\frac{\mu^2}{2}. \tag{11.1}$$

This, and the other functional forms below, are chosen for expositional and mathematical convenience.

Letting μ_{ij} be the level of punishment of individual j by individual i, the material payoff to i is then given by

$$\pi_i = 1 - a_i + \chi(a_1 + a_2) - \mu_{ji} - c(\mu_{ij}). \tag{11.2}$$

In (11.2), the first two terms give the amount remaining in i's private account after contributing, the third term is i's reward from the public project, the fourth term is the punishment inflicted by j upon i, and the final term is the cost to i of punishing j.

We assume that the norm is that each should contribute the entire endowment to the public project. The results generalize to the case where the norm is less stringent. Individual i may wish to punish j by reducing j's payoffs, if i is a reciprocator (that is $\lambda_i > 0$) and j contributes less than the entire endowment. To represent the propensity of i to punish j for not contributing sufficiently, we assume that i's valuation of j's payoff is

$$\beta_{ij} = \lambda_i(a_j - 1), \tag{11.3}$$

where we assume $0 < \lambda_i < 1$, so that unless j contributed his entire endowment, i receives a subjective benefit from lowering j's material payoff that is proportional to j's shortfall. The parameter λ_i, $0 < \lambda_i < 1$, is the strength of i's reciprocity motive. The condition that $\lambda_i < 1$ ensures that individual i cannot value j's payoffs negatively more than he values his own positively. Thus should both payoffs increase proportionally, individual i cannot be worse off.

The shame experienced by i is a subjective cost proportional to the product of the degree to which he is punished by j, and the extent to which his contribution falls short of the norm, and is equal to $\nu_i(1 - a_i)\mu_{ji}$. Thus, punishment triggers shame, which is greater the more the individual has kept for himself rather than contributing to the public project, and the larger is ν_i, the susceptibility of individual i to feeling shame. Finally, i may feel guilt simply for having violated his internal standards of moral behavior. We represent this feeling by $-\gamma_i(1 - a_i)$, which is negative for $\gamma_i > 0$ unless i contributes the full amount to the project. The parameter γ_i is i's susceptibility to guilt.

The utility function of i is then given by

$$u_i = \pi_i + \beta_{ij}(1 - a_j + \chi(a_1 + a_2) - \mu_{ij}) - (\gamma_i + \nu_i\mu_{ji})(1 - a_i). \tag{11.4}$$

The first term is i's material payoffs, which are those from the public project net of his own contribution and minus the cost of being punished by j and the cost of punishing j, from equation 11.2. The second term is (using equation 11.3) i's evaluation of j's material payoffs, which are those from the project net of his own contribution and minus i's punishment of j.

We have not included the cost to j of punishing i, in the material payoffs of j that i takes account of when choosing his contribution level because we think it is unrealistic to imagine that i would seek to reduce j's payoffs by inducing j to bear costs so as to punish i. The third term is the guilt and punishment-induced shame that i experiences when i contributes less than the amount that would maximize the well-being of the two players, namely 1.

Given any level of j's contribution, we can represent individual i's behavior as the joint maximization of two objective functions. The first is, given j's contribution,

how much to punish j. The answer is to select μ_{ij} so as to equate the marginal cost of punishment ($dc/d\mu_{ij} = c\mu_{ij}$) with the marginal benefit of punishing j, which is β_{ij}. Given this level of punishment, i will then select the level of contribution that equates the marginal benefits of contributing, which are reduced punishment, guilt and shame, and the marginal costs of contributing, which involve forgoing some of one's endowment and contributing to the material payoffs of j, even though i values these negatively. Note that because the susceptibility to shame and the level of punishment received appear multiplicatively in this last term, punishment and shame are what economists call complements. This means that an increase in the susceptibility to shame increases the marginal effect of punishment on the individual's utility and therefore raises the marginal benefit that i will receive by contributing more. Similarly, an increase in the level of punishment raises the marginal effect of an enhanced susceptibility to shame on the actor's utility. Shame thus enhances what is termed the "punishment technology," the effectiveness of which is measured by the ratio of the utility loss inflicted on the target, including both the subjective costs and the reduction in payoffs from equation 11.2, to the marginal cost to the punisher of undertaking the punishment, which from equation 11.1 is $c\mu_{ij}$. This punishment effectiveness ratio for i's punishment of j is thus

$$\frac{1 + \nu_j(1 - a_j)}{c(\mu_{ij})}, \tag{11.5}$$

from which it is clear that the punishment of j is more effective the more susceptible to shame is j.

Because each individual's valuation of the payoffs of the other depends on the actions the other takes, it is clear that the actions taken by each will be mutually determined. For any given value of j's action, there will be an action—a best response—by i that maximizes his utility as expressed in equation 11.4. The best response function for individual i is shown in Figure 11.1, along with the analogous best response function for j. Their intersection is the mutual best response, and is therefore a Nash equilibrium. In Figure 11.1 we see that because of reciprocity, the best response a_i is an increasing function of a_j, and the a_i schedule shifts up when susceptibility to shame or guilt, or j's degree of reciprocity (ν_i, γ_i, λ_j), increases, corresponding to our intuitions concerning the model. There is also a minimal level of susceptibility to shame supporting positive contributions. The minimal level of shame that will induce a positive contribution is increasing in the cost of punishment and decreasing in i's susceptibility to guilt γ_i, j's level of reciprocity λ_j, and the productivity χ of the public project, again confirming our intuitions.

Suppose the level of shame of both individuals were to increase. This is shown in Figure 11.1 by the dashed lines. The result is a displacement of the mutual best response so that both individuals contribute more, and as a result the level of punishment is less. This is the sense in which we mean that because shame enhances the effectiveness of punishment, it economizes on the cost of punishment. When one individual's susceptibility to shame increases the other individual benefits and when this occurs for both, as in Figure 11.1, both benefit. Payoffs therefore are higher in a population that has inculcated a sense of shame in its members, as could be the case, for example,

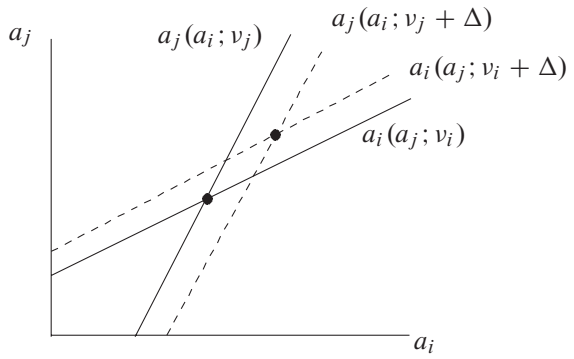

Figure 11.1. Mutual determination of contributions to a public project. The functions slope upwards because the individuals are reciprocators and shift as shown when susceptibility to shame, v, increases, because this enhances the effects of punishment. There is no reason to think that the function would take the linear form shown here.

through the kinds of population-wide internalization of norms studied in the previous chapter.

11.2 The Evolution of Social Emotions

Human behaviors systematically deviate from the model of the self-interested actor, and we think the evidence is strong that social emotions account for much of the discrepancy. But this description of behavior would be more compelling if we understood how social emotions might have evolved, culturally, genetically, or both. There are two puzzles here. First, social emotions are often altruistic, indicating actions benefiting others at a cost to oneself, so that in any dynamic in which the higher payoff trait tends to increase in frequency, social emotions would eventually disappear. We addressed this puzzle in the previous four chapters, showing that by the process of group competition, reproductive leveling, and norm internalization, vertically transmitted altruistic traits may evolve.

The second puzzle concerns social emotions per se. How could it ever be evolutionarily advantageous to bypass one's cognitive decision making capacities and let behavior be influenced by the visceral reactions associated with one's emotions? We addressed a similar question in the previous chapter: internalizing norms may be a way of economizing the costs of calculating benefits and costs in each situation, and of averting costly errors when the calculations go wrong. A related argument, we think, helps explain the evolutionary viability of social emotions.

Humans tend to be impatient, a condition we share with other animals (Stephens et al. 2002). We tend to discount future costs and benefits myopically, that is, more than either a fitness-based or a lifetime welfare-based accounting would require. The mismatch between our impatience and our fitness is in part due to the payoff to patient behaviors that resulted from the extended life histories and prolonged period of learning the skills associated with the distinctive skill-intensive human feeding niche

based on hunted and extracted foods. Prior to this period in human history, the importance of the future was more limited and largely concerned the survival of one's offspring. A genetically transmitted disposition to assist one's relatives may have produced a selective degree of patience as a by-product of kin-based selection, resisting stealing food from one's offspring, for example. But even if our genetic development in a cooperative social context has mitigated the extreme short-term benefits of lying, cheating, killing, stealing, and satisfying immediate bodily needs, such as wrath, lust, greed, gluttony, sloth, we nevertheless have a fitness-reducing bias toward behaviors that produce immediate satisfaction at the expense of our long-run well-being.

The internalization of norms and the expression of these norms in a social emotion such as guilt and shame addresses this problem by inducing the individual to place a contemporaneous value on the future consequences of present behavior, rather than relying upon an appropriately discounted accounting of its probable payoffs in the distant future. One may curb one's anger today not because there may be harmful effects next month, but because one would feel guilty now if one violated the norms of respect for others and the dispassionate adjudication of differences. One may punish others for behaving antisocially not because there are future benefits to be gained thereby, but because one is angered at the moment.

Do the social emotions thus function in a manner similar to pain? Complex organisms have the ability to learn to avoid damage. A measure of damage is pain, a highly aversive sensation the organism will attempt to avoid in the future. Yet an organism with complete information, an unlimited capacity to process information, and with a fitness-maximizing way of discounting future costs and benefits would have no use for pain. Such an individual would be able to assess the costs of any damage to itself, would calculate an optimal response to such damage, and would prepare optimally for future occurrences of this damage. The aversive stimulus, pain, could then be strongly distorting of optimal behavior. If you sprain your ankle while fleeing from a lethal predator, you might have a better chance of survival if you could override the pain temporarily. Because pain per se clearly does have adaptive value, it follows that modeling pain presupposes that the individual experiencing pain must have incomplete information and/or a limited capacity to process information, and/or an excessively high rate of discounting future benefits and costs. Are guilt and shame social analogues to pain?

If being socially devalued has fitness costs, and if the amount of guilt or shame that a given action induces is closely correlated with the level of these fitness costs that would otherwise not be taken account of, then the answer is affirmative. The same argument will hold not only for fitness costs, but for any effect, possibly operating through cultural transmission, that reduces the number of replicas an individual will generate.

11.3 The "Great Captains of Our Lives"

Shame and guilt, like pain, dispense with an involved optimization process by means of a simple message: whatever you did, undo it if possible, and do not do it again. Two types of selective advantage thus may account for the evolutionary success of shame and related social emotions. First, social emotions may increase the number of replicas,

by either genetic or cultural transmission, of an individual who has incomplete information (e.g., as to how damaging a particular antisocial action is), limited or imperfect information-processing capacity, and/or a tendency to undervalue costs and benefits that accrue in the future. Probably all three conditions conspire to induce people to respond insufficiently to social disapprobation in the absence of social emotions. The visceral reactions associated with these emotions motivate a more adequate response, one that will avert damage to the individual. Of course the role of social emotions in alerting us to negative consequences in the future presupposes that society is organized to impose those costs on norm violators. The social emotions thus may have coevolved with the reciprocity-based emotions motivating punishment of antisocial actions, modeled in the previous chapters.

The second selective advantage favoring the evolution of social emotions refers specifically to shame. The fact that the higher levels of shame among members of a group, the higher (in equilibrium) will be the sum of their payoffs also suggests that shame may evolve through the effects of group competition. As we have seen, where the emotion of shame is common, punishment of antisocial actions will be particularly effective and as a result seldom used. Thus groups in which shame is common can sustain high levels of group cooperation at limited cost and will be more likely to survive environmental, military and other challenges, and thus to populate new sites vacated by groups that failed.

As a result, selective pressures at the group level will also favor religious practices and systems of socialization that support susceptibility to shame for failure to contribute to projects of mutual benefit of the type modeled in the previous two sections.

It is quite likely, then, that the "moralistic aggression" that is involved in the altruistic punishment of miscreants and that motivated the punishment of shirkers in Chapter 9 also created a selective niche favorable to the emergence of shame and other social emotions, or what Christopher Boehm (2007) calls a conscience:

> The human conscience evolved in the Middle to Late Pleistocene as a result of subsistence turning to the hunting of large game. This required... cooperative band-level sharing of meat... bands had to gang up physically against their alphas to ensure efficient meat distribution. This sets the stage for morality to develop as a new, more socially-sensitive type of personal self-control became adaptive for individuals living in these punitive groups. Thus a conscience began to develop biologically. In turn... conscience transformed social control by making punitive sanctioning increasingly moral and also less lethal, as group ostracism and shaming evolved. (Boehm 2007, p. 1)

Combining the model of this chapter and that of Chapter 9, the emergence of shame would have reduced the costs of punishing transgressors incurred by the strong reciprocators. The reason for this is that gossip and ridicule could then suffice where physical, often violent, elimination from the group had been necessary in the absence of shame. The proliferation of strong reciprocators engaging in altruistic punishment that this cost reduction allowed would then have enhanced the advantages of shame.

Thus the moralistic aggression motivating the altruistic punishment of defectors may have coevolved with shame, each providing the conditions favoring the prolifera-

tion of the other. The groups in which this occurred initially, perhaps among our foraging ancestors in Africa, would have enjoyed survival advantages over other groups.

Symbol	Meaning
a_i	Individual i' contribution to the project
β_{ij}	Equals $\lambda_i(a_j - 1)$
χ	Each individual receives $\chi(a_1 + a_2)$ from project
γ_i	Individual i's guilt coefficient
i, j	Two players
λ_i	$\lambda_i(a_j - 1)$ is the value i placed on j's contribution
μ_{ij}	Level of punishment of j by i
ν_i	Individual i's shame coefficient
n	Group size
π_i	Material payoff to i

Table 11.1. Definition of symbols

12

Conclusion: Human Cooperation and Its Evolution

> It is true that certain living creatures, as bees and ants, live sociably one with another. . . and yet have no other direction than their particular judgments and appetites; nor speech, whereby one of them can signify to another what he thinks expedient for the common benefit: and therefore some man may perhaps desire to know why mankind cannot do the same.
>
> Thomas Hobbes, *Leviathan* Chapter 8 (1968[1651]).

> Any animal whatever, endowed with well-marked social instincts, the parental and filial affections being here included, would inevitably acquire a moral sense or conscience, as soon as its intellectual powers had become as well developed, or nearly as well developed, as in man.
>
> Charles Darwin, *The Descent of Man* Chapter IV (1998[1873]) pp. 71–72

About 55,000 years ago, a group of hunter-gatherers left Africa and began to move eastward along the shores of the Indian Ocean. They may have originated in the Upper Rift Valley in modern-day Kenya. They could have been the descendants of the cooperative early humans we described at the outset, living 30,000 years earlier at the mouth of the Klassies River far to the south. Wherever they came from, some eventually crossed hundreds of kilometers of open ocean before reaching Australia, just 15,000 years later. We do not know if they encountered or simply bypassed communities of *Homo floresienis*, who persisted in what is now Indonesia almost to the end of the Pleistocene. As they spread northward, they also encountered the Denisovan hominins, who inhabited parts of Asia as recently as 50,000 years ago. Another branch of the African exodus crossed the Levant and somewhat later occupied Europe, then home to the soon-to-be-extinct Neanderthals. Though the possibility of multiple human origins cannot be eliminated, it is now widely thought that the descendants of this small group eventually peopled the entire world and are the ancestors of all living humans (Foley 1996, Klein 1999).

This second great exodus from Africa is remarkable for its speed and eventual spread. One cannot resist speculating about the capacities that made these particular individuals such lethal competitors for the (also large-brained, ornament-wearing and tool-making) Neanderthals or that allowed the construction of oceangoing craft. Some attractive candidates can be ruled out. The physiological innovations allowing for more effective speech, rearrangement of respiratory tract and esophagus, for example, had occurred much earlier. Likewise, the dramatic expansion of hominid brain size had occurred before two million years ago. Richard Klein (2000) suggests a "selectively

advantageous mutation" that facilitated the cultural transmission of behaviors as a possible cause.

> Arguably this was the most significant mutation in the human evolutionary series for it produced an organism that could radically alter its behavior without any change in its anatomy and that could cumulate and transmit alterations at a speed that anatomical innovation could never match. (p. 18)

But, as Klein himself points out, the only evidence for such a super-mutation are the facts it is intended to explain (Klein 2000). Whether the source was a single revolutionary innovation or, as many now think (McBrearty and Brooks 2000), the result of a long process of incremental changes, the linguistic capacities and the cultural transmission of norms of social conduct that supported cooperation were a necessary part of the human repertoire that made the peopling of the world possible. These same capabilities must be part of any account of the remarkable success of humans as a species then and since.

12.1 The Origins of Human Cooperation

Humans became a cooperative species because our distinctive livelihoods made cooperation within a group highly beneficial to its members and, exceptionally among animals, we developed the cognitive, linguistic and other capacities to structure our social interactions in ways that allowed altruistic cooperators to proliferate.

Human reliance on the meat of large hunted animals and other high quality, large package-size, and hence high-variance foods meant that our livelihoods were risky, skill-intensive, and characterized by increasing returns to scale. Deploying skills that required years to acquire favored the evolution of large brains, patience, and long lives (Kaplan et al. 2000, Kaplan and Robson 2003). Organizing and sharing the returns to successful hunting additionally favored groups that developed practices of sharing information, food, and other valued resources (Boehm 2000). Moreover, the long period of dependency of human offspring on adults, in part the result of the prolonged learning curve associated with hunting and gathering, meant that there were substantial benefits to cooperative child-rearing practices extending beyond the immediate family. Prolonged juvenile dependency also generated a net food deficit for families with adolescent children, increasing the benefits of food-sharing among unrelated individuals and other forms of social insurance (Kaplan and Gurven 2005). Our experimental evidence, presented in Chapter 3, shows that among today's small-scale societies, those that are especially reliant on big game, like the Lamalera whale hunters that we studied in Indonesia, and those for whom livelihoods require either joint efforts in acquisition or sharing in distribution, are especially likely to exhibit the social preferences that underpin altruistic cooperation.

One of the reasons for the connection between the potential benefits of cooperation and the prevalence of cooperative behaviors that we discovered in our models and simulations is that where the benefits associated with cooperation relative to the costs are substantial, it is more likely that the evolutionary processes of gene-culture coevo-

lution will support populations with large numbers of cooperators, whether altruistic or mutualistic. A high ratio of benefits to costs makes cooperation an evolutionarily likely outcome because, as our models and simulations, for example, Figures 4.6, 9.1, and 9.4 confirmed, in virtually any plausible evolutionary dynamic in which stochastic shocks to payoffs and to behaviors play an important role, the likelihood that a population will develop and maintain cooperative practices is higher, the greater are the net benefits of cooperation.

But the fact that cooperation was group-beneficial in the environments of early humans does not explain why it evolved, for individuals bear the costs of their cooperative behaviors, while it is often others who enjoy the benefits. Thus, the distinctive human livelihood and associated cognitive capacities and longevity are necessary but not sufficient to explain the extent and nature of human cooperation. While benefits of cooperation accruing to the individual cooperator may sometimes offset the costs, this is not likely to have been the case in many situations in which cooperation was essential to our ancestors, including defense, predation and surmounting environmental crises. In these situations involving large numbers of individuals facing their possible demise, people with self-regarding preferences would not cooperate, regardless of their beliefs about what others would do. As a result, for cooperation to be sustained, social preferences would have to motivate at least some of those involved.

The distinctive human capacity for institution-building and cultural transmission of learned behavior allowed social preferences to proliferate. Our ancestors used their capacities to learn from one another and to transmit information to create distinctive social environments. The resulting institutional and cultural niches reduced the costs borne by altruistic cooperators and raised the costs of free-riding. Among these socially constructed environments, three were particularly important: group-structured populations with frequent and lethal intergroup competition, within-group leveling practices such as sharing food and information, and developmental institutions that internalized socially beneficial preferences.

These culturally transmitted institutional environments created a social and biological niche favorable to the evolution of the social preferences on which altruistic cooperation is based. We can only speculate, of course, about the initial appearance and proliferation of these preferences. But their emergence was highly likely for two reasons. The first is that the preferences that constitute strong reciprocity and some other social preferences could appear *de novo* as the result of only a small behavioral modification of either kin-based altruism or reciprocal altruism. In the case of kin-based altruism, those behaving altruistically toward kin may have simply ceased discriminating against the non-kin members of their groups. Likewise, a reciprocal altruist could become a strong reciprocator by simply deleting the proviso that one should condition one's behavior on expectations of future reciprocation.

The second reason why the emergence of social preferences among early humans would be highly likely is the vast number of foraging bands during the Late Pleistocene and earlier. Even if strong reciprocity initially emerged in a very small fraction of the human population, it is highly likely that over tens of thousands of generations and something like 150,000 foraging bands, it would have occurred that the strong reciprocators or other altruistic cooperators were prevalent in one or more such groups at some point. These bands would have done very well in competition with other bands.

We have sought to explain how humans came to develop these exceptional social preferences and the cooperative social practices that supported them, taking the distinctive nature of human ecology, diet, and life course as preexisting. This analytical simplification is almost surely historically inaccurate. The distinctive nature of human livelihoods, the importance of hunted and extracted as opposed to collected foods, apparently does not predate and is not the cause of the emergence of cooperation. Rather, it appears that the two developed in tandem.

Though we have not addressed this question, we think it likely that the models presented here, suitably amended, would illuminate the coevolution of human cooperation along with our distinctive diets, life histories, and livelihoods. The presence on the African savannah of large mammals vulnerable to attack by cognitively advanced predators must have given substantial advantages to the members of groups that developed means of coordinating the hunt and sharing its sporadically acquired prey. Correspondingly, groups that had learned how to cooperate in these ways would have benefited from preferentially targeting large animals, as opposed to food acquired in smaller packages, and thereby enlarging the place of hunted meat in their diet. Winterhalder and Smith (1992) write:

> only with the evolution of reciprocity or exchange-based food transfers did it become economical for individual hunters to target large game. The effective value of a large mammal to a lone forager... probably was not great enough to justify the cost of attempting to pursue and capture it... However, once effective systems of reciprocity or exchange augment the effective value of very large packages to the hunter, such prey items would be more likely to enter the optimal diet. (p. 60)

We think it likely that the distinctive aspects of the human livelihood thus coevolved with the distinctive aspects of our social behavior, most notably cooperation.

Two approaches inspired by standard biological models have constituted the workhorses of our explanation, multi-level selection and gene-culture coevolution. Could it be that altruistic cooperation became common among humans in the absence of these two processes? We think it empirically unlikely. The reason is that the kin-based and reciprocal altruism models, operating alone or in tandem, are peculiarly ill-suited to explain the distinctive aspects of human cooperation, for the reasons given in Chapter 4 and 6.

By contrast, explanations of the emergence and proliferation of cooperative behaviors based on gene-culture coevolution and multi-level selection are quite plausible. First, the models and simulations of our evolutionary past presented in the previous chapters provide strong evidence that in the relevant evolutionary environments, selective pressures based on the positive assortment of behaviors arising from the group-structured nature of human populations could have been a significant influence on human evolution. Second, we have also demonstrated the important contribution to the evolution of social preferences that could have been accomplished by the cultural transmission of empirically well-documented behaviors such as the internalization of norms, within-group leveling, and between-group hostility. Third, the nature of preferences revealed in behavioral experiments and in other observations of human behavior is consistent with the view that genuine altruism, a willingness to sacrifice one's own

interest to help others, including those who are not family members, and not simply in return for anticipated reciprocation in the future, provides the proximate explanation of much of human cooperation. These ethical and other-regarding group-beneficial social preferences are the most likely psychological consequence of the gene-culture coevolutionary and multi-level selection processes we have described.

12.2 The Future of Cooperation

Conclusive evidence about the origins of human cooperation will remain elusive given the paucity of the empirical record and the complexity of the dynamical processes involved. As in many problems of historical explanation, perhaps the best that one can hope for is a plausible explanation consistent with the known facts. This is what we have attempted to provide.

The challenge of explaining the origins of human cooperation has led us to the study of the social and environmental conditions of life of mobile foraging bands and other stateless small-scale societies that arguably made up most of human society for most of the history of anatomically modern humans. The same quest has made non-cooperative game theory (which assumes the absence of enforceable contracts) an essential tool. But as Ostrom (1990), Taylor (1996), and other authors have pointed out, most forms of contemporary cooperation are supported by incentives and sanctions based on a mixture of multilateral peer interactions and third-party enforcement, often accomplished by the modern nation-state.

It would thus be wise to resist drawing strong conclusions about cooperation in the 21st century solely on the basis of our thinking about the origins of cooperation in the Late Pleistocene. One may doubt, for example, that lethal intergroup conflict today contributes to the altruism, civic-mindedness or other social preferences that could underpin the more cosmopolitan forms of cooperation required to address global challenges such as climate change and epidemics.

But the fundamental challenges of social living and sustaining a livelihood that our distant ancestors faced are in many respects not fundamentally different from those we face today. Modern states and global markets have provided conditions for mutualistic cooperation among strangers on a massive scale. But altruistic cooperation remains an essential requirement of economic and social life.

The reason is that neither private contract or governmental fiat singly or in combination provides an adequate basis for the governance of modern societies. Social interactions in modern economies are typically at best quasi-contractual. Some aspects of what is being transacted are regulated by complete and readily enforceable contracts, while others are not. Transactions concerning credit, employment, information, and goods and services where quality is difficult to monitor provide examples of quasi-contractual exchanges.

Where contracting is absent or incomplete, the logic of Adam Smith's invisible hand no longer holds. Decentralized markets fail to implement efficient allocations. But governments typically lack the information, and often the motivation, necessary to provide adequate governance where markets fail or are absent.

We now know from laboratory experiments that subjects in marketlike situations with complete contracts tend to behave like the *Homo economicus* of the Adam Smith of *The Wealth of Nations*, but when their contracts are not complete their behavior fortunately resembles more the virtuous citizens of the Adam Smith of *The Theory of Moral Sentiments*. Thus, where the invisible hand fails, the handshake may succeed. Kenneth Arrow wrote (1971)

> In the absence of trust... opportunities for mutually beneficial cooperation would have to be foregone... norms of social behavior, including ethical and moral codes [may be]... reactions of society to compensate for market failures. (p. 22)

Thus, social preferences such as a concern for the well-being of others and for fair procedures remain essential to sustaining society and enhancing the quality of life.

In a world increasingly connected not just by trade in goods but also by the exchange of violence, information, viruses, and emissions, the importance of social preferences in underwriting human cooperation, even survival, may now be greater even than it was among that small group of foragers that began the exodus from Africa 55,000 years ago to spread this particular cooperative species to the far corners of the world.