J. theor. Biol. (2003) **220,** 407–418
doi:10.1006/jtbi.2003.3104, available online at http://www.idealibrary.com on IDEAL®

The Hitchhiker's Guide to Altruism: Gene-culture Coevolution, and the Internalization of Norms

HERBERT GINTIS*

Santa Fe Institute and Department of Economics, University of Massachusetts, 15 Forbes Avenue, MA 01060, Northampton, U.S.A.

(*Received on 5 December* 2001, *Accepted in revised form on 4 June* 2002)

An *internal norm* is a pattern of behavior enforced in part by internal sanctions, such as shame, guilt and loss of self-esteem, as opposed to purely external sanctions, such as material rewards and punishment. The ability to internalize norms is widespread among humans, although in some so-called "sociopaths", this capacity is diminished or lacking. Suppose there is one genetic locus that controls the capacity to internalize norms. This model shows that if an internal norm is fitness enhancing, then for plausible patterns of socialization, the allele for internalization of norms is evolutionarily stable. This framework can be used to model Herbert Simon's (1990) explanation of altruism, showing that altruistic norms can "hitchhike" on the general tendency of internal norms to be personally fitness-enhancing. A multi-level selection, gene-culture coevolution argument then explains why individually fitness-reducing internal norms are likely to be prosocial as opposed to socially harmful.

Introduction

An *internal norm* is a pattern of behavior enforced in part by internal sanctions, including shame, guilt and loss of self-esteem, as opposed to purely external sanctions, such as material rewards and punishments. Humans internalize norms through *socialization* by parents (*vertical transmission*) and extraparental conspecifics (*oblique* and *horizontal transmission*). The capacity to internalize norms is widespread among humans, although in some so-called "sociopaths", this capacity is diminished or lacking (Mealey, 1995). Human behavior is commonly modeled assuming agents have objective functions which they maximize subject to constraints. In these terms, the capacity to internalize norms means human agents have *socially programmable* objective functions. Human behavior thus depends not only on *beliefs*, which concern constraints on action (taking action X will lead to result Y), but *values*, which are the very *goals* of action.

Suppose there is one genetic locus that controls the capacity to internalize norms. I develop models of gene-cultural coevolution to show that if an internal norm is fitness enhancing, then the allele for internalization of norms is evolutionarily stable. Moreover, if the fitness payoff to the internalized norm is sufficiently large, or if there is a sufficiently high rate of phenotypic level assortative mating, the allele for internalization is globally stable.

Basic sociological theory holds that society's values are transmitted through the internalization of norms (Parsons, 1967; Grusec & Kuczynski, 1997). Successful societies tend to foster internal norms that enhance personal

*Tel.: +1-413-586-7756; fax: +1-413-586-6014.
E-mail address: hgintis@attbi.com (H. Gintis).
URL: http://www-unix.oit.umass.edu/~gintis.

0022-5193/03/$35.00

fitness, such as future-orientation, good personal hygiene, positive work habits, and control of emotions, as well as altruistic norms that subordinate the individual to group welfare, fostering such behaviors as bravery, honesty, fairness, willingness to cooperate, and empathy for others (Brown, 1991). People follow internal norms because they value this behavior for its own sake, in addition to, or despite, the effects the behavior has on personal fitness and/or perceived well-being. For instance, an individual who has internalized the value of "speaking truthfully" will do so even in cases where the net payoff to speaking truthfully would otherwise be negative. It follows that where people internalize a norm, the frequency of its occurrence in the population will be higher than if people follow the norm only instrumentally; i.e. when they perceive it to be in their interest to do so.

Why does the capacity to internalize norms have adaptive value? It might be argued that if a norm is fitness enhancing, a non-internalizing agent could simply *mimic* the behavior of an internalizer. But this assumes that agents maximize fitness. In general, however, in any species, individuals do not maximize *fitness* but rather a objective function that has evolved to reflect biological fitness more or less accurately for a given environment. If the *Homo sapiens* objective function were perfectly adapted, internalization would not be fitness-enhancing. But the rapid cultural evolution and highly variable environments (Richerson *et al.*, 2001) that characterized the period in which *Homo sapiens* developed doubtless led to a situation in which the unsocialized human objective function deviated strongly from fitness maximization. The internalization of norms thus permitted rapid cultural adaptation towards fitness-maximization, while a purely genetic adaptive process would have taken orders of magnitude longer in time. In short, non-internalizers fail to mimic internalizers not because they *cannot*, but because they *do not want to* (sociopaths, for instance, do often mimic internalizers—displaying empathy and helpfulness, for example—but only as long as it suits them to do so).

Altruism is the tendency of individuals to behave prosocially towards unrelated others (e.g. by helping those in distress and punishing anti-social behavior) at personal cost.† Adding an altruism norm allows us to model Herbert Simon's (1990) explanation of altruism. Simon suggested that altruistic norms could "hitch-hike" on the general tendency of internal norms to be personally fitness-enhancing. Of course, internal norms may persist even if they are fitness-*reducing* both for individuals and the group (Boyd & Richerson, 1992; Edgerton, 1992). I develop a multi-level gene-culture coevolutionary model to elucidate the process whereby altruistic internal norms will tend to drive out norms that are both socially harmful and individually fitness-reducing.

Socialization and Fitness-Enhancing Internal Norms

Suppose there is a norm **C** that can be internalized by a new member of society. Norm **C** confers fitness $1+t>1$, while the normless phenotype, denoted by **D**, has baseline fitness 1. There is a genetic locus with two alleles, {a} and {b}. Allele {a}, which is dominant, permits the internalization of norms, whereas {b} does not. We assume that possessing at least one copy of **a** imposes a fitness cost $u\in(0,1)$, on the grounds that there are costly physiological and cognitive prerequisites for the capacity to internalize norms.‡ We assume $(1+t)(1-u)>1$, so the cost of the internalization allele is more than offset by the benefit of the norm **C**. There are five phenogenotypes, whose fitnesses are listed in Fig. 1.§

Families are formed by random pairing, and offspring genotypes obey the laws of Mendelian segregation. Thus, there are six familial genotypes, **aaaa**, **aaab**, **aabb**, **abab**, **abbb**, and **bbbb**. We assume also that only the phenotypic

† For reviews of the evidence on the importance of altruism in human societies, see Sober & Wilson (1998), Gintis (2000a, b), and Fehr & Gächter (2002).

‡ Assuming $u>0$ is conservative, in that it biases the model against the global stability of the internalization allele. However, the contrasting assumption $u<0$ is also plausible. I will point out the implications of $u<0$ where appropriate.

§ Feldman *et al.* (1985) develop a model similar to ours. Their model, however, is haploid, assumes uniparental transmission, and the phenotypic trait is kin-altruistic. Ours, by contrast, is diploid, assumes biparental transmission, and abstracts from kin-altruism.

Individual Phenogenotype	Individual Fitness
aaC	(1-u)(1+t)
aaD	(1-u)
abC	(1-u)(1+t)
abD	(1-u)
bbD	1

FIG. 1. Fitnesses of the five phenogenotypes. Here u is the fitness cost of possessing the internalization allele, and t is the excess fitness value of possessing the norm **C**. Note that **bbC** cannot occur.

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 18 familial phenogenotypes, of which only 14 can occur. The frequency of familial phenogenotypes are as shown in Fig. 2, where $p(i)$ represents the frequency of phenogenotype $i =$ **aaC**, …, ***bbD***.

The rules of gene-culture transmission are as follows. If familial phenogenotype is **xyzwXY**, where **x,y,z,w** $\in$ {**a,b**}, X, $Y \in$ {**C,D**}, an offspring is equally likely to inherit **xz**, **xw**, **yz**, or **yw**. An offspring whose genotype includes a copy of the **a** allele is equally likely to inherit **X** or **Y**.‖ But an offspring of genotype **bb** always has the normless phenotype **D**. The transition table is shown in Fig. 3.¶

The above accounts only for parental transmission. In addition, extraparental transmission is ubiquitous in human society, in the form of social pressure (rumor, shunning, and ostracism), rituals (dancing, prayer, marriage, birth, and death), and in modern societies, formalized institutions (schools and churches).** To account for extraparental transmission, let p_C be the fraction of the population carrying the **C** phenotype, and let $\gamma \in [0, 1]$. We assume a fraction γp_C of **aa**-types and a fraction vp_C of **ab**-types who have not internalized **C** through parental transmission, are influenced by extraparental transmission to switch to **C**. **bb** types are not affected by extraparental transmission.

The resulting system consists of four equations in four unknowns (**bbC** cannot occur, and one offspring phenogenotype is dropped, since the sum of phenogenotypic frequencies equals unity). It is straightforward to check that there are three pure equilibria (i.e. equilibria in which the whole population bears a single phenogenotype). These are **aaC**, in which all agents internalize the fitness enhancing norm, **aaD**, in which the internalization allele is present but the phenotype **C** is absent, and **bbD**, in which neither the internalization allele nor the norm is present.

A check of the eigenvalues of the Jacobian matrix of the dynamical system shows that the **aaD** equilibrium is unstable. Eigenvalues of the system at the **aaC** equilibrium are given by

$$\left\{0, 1, \frac{1-\gamma}{2(1+t)}, \frac{1-\gamma}{1+t}\right\}.$$

The unit eigenvalue is semisimple,†† so the linearization of the equilibrium **aaC**, in which the fitness-enhancing norm is internalized, is stable. However, we cannot conclude that the nonlinear model itself is stable. Extensive simulations fail to find a case in which the **aaC** equilibrium is unstable.‡‡ Moreover, in the case where the **a** allele is incompletely dominant, the unit root disappears, so stability is assured (this remark applies as well to all cases in which unit roots appear, some of which are discussed below).

‖ Simulations show that the assumption that the **a** allele is dominant is not critical. The stability results described below continue to hold, and indeed more strongly, when **a** is less than fully dominant.

¶ Biased parental transmission, in which heterogeneous familial phenotypes are more likely to transmit one phenotype to offspring that the other (Cavalli-Sforza & Feldman, 1981) is discussed below.

** Extraparental transmission is generally individually costly and the benefits accrue to unrelated others. Hence it is a form of altruistic behavior, and ideally should not be introduced until our analysis of altruism is completed. We introduce it now purely for expositional purposes.

†† An eigenvalue is semisimple if its algebraic and geometric dimensions are equal. Semisimple unit roots of linear dynamical systems are stable.

‡‡ The process of coding this and the other models presented in this paper is tedious and error-prone. To ensure accuracy I wrote the simulations in two completely different languages, one Lisp-like (Mathematica) and the other procedural (C++), and verified that the results agreed to six decimal places over thousands of generations of simulation.

Familial Phenogenotype	Frequency in Reproductive Pool
aaaaCC	$p(\mathbf{aaC})^2(1-u)^2(1+t)^2/\bar{p}$
aaaaCD	$2p(\mathbf{aaC})p(\mathbf{aaD})(1-u)^2(1+t)/\bar{p}$
aaaaDD	$p(\mathbf{aaD})^2(1-u)^2/\bar{p}$
aaabCC	$2p(\mathbf{aaC})p(\mathbf{abC})(1-u)^2(1+t)^2/\bar{p}$
aaabCD	$2(p(\mathbf{aaC})p(\mathbf{abD})+p(\mathbf{aaD})p(\mathbf{abC}))(1-u)^2(1+t)/\bar{p}$
aaabDD	$2p(\mathbf{aaD})p(\mathbf{abD})(1-u)^2/\bar{p}$
ababCC	$p(\mathbf{abC})^2(1-u)^2(1+t)^2/\bar{p}$
ababCD	$2p(\mathbf{abC})p(\mathbf{abD})(1-u)^2(1+t)/\bar{p}$
ababDD	$p(\mathbf{abD})^2(1-u)^2/\bar{p}$
aabbCD	$2p(\mathbf{aaC})p(\mathbf{bbD})(1-u)(1+t)/\bar{p}$
aabbDD	$2p(\mathbf{aaD})p(\mathbf{bbD})(1-u)/\bar{p}$
abbbCD	$2p(\mathbf{abC})p(\mathbf{bbD})(1-u)(1+t)/\bar{p}$
abbbDD	$2p(\mathbf{aaC})p(\mathbf{aaD})(1-u)(1+t)/\bar{p}$
bbbbDD	$2p(\mathbf{bbD})^2/\bar{p}$

FIG. 2. Frequencies of phenogenotypes. Here, $\bar{p}$ is chosen so the sum of the frequencies is unity. Note that **aabbCC**, **abbbCC**, **bbbbCC**, and **bbbbCD** are not listed, since **bbC** cannot occur.

Familial Type	Offspring Phenogenotypic Frequency				
	aaC	**aaD**	**abC**	**abD**	**bbD**
aaaaCC	1				
aaaaCD	1/2	1/2			
aaaaDD		1			
aaabCC	1/2		1/2		
aaabCD	1/4	1/4	1/4	1/4	
aaabDD		1/2		1/2	
aabbCD			1/2	1/2	
aabbDD				1	
abbbCD			1/4	1/4	1/2
abbbDD				1/2	1/2
ababCC	1/4		1/2		1/4
ababCD	1/8	1/8	1/4	1/4	1/4
ababDD		1/4		1/2	1/4
bbbbDD					1

FIG. 3. Phenotypic inheritance is controlled by genotype. Note that **aabbCC**, **abbbCC**, **bbbbCC**, and **bbbbCD** are not listed, since **bbC** cannot occur.

The eigenvalues of the Jacobian matrix of the unnormed equilibrium **bbD** are given by

$$\left\{(0,0,1-u,\tfrac{1}{2}(1+t)(1-u)\right\}.$$

Therefore this equilibrium, in which no internalization occurs, is locally stable if $(1+t)(1-u)<2$, and unstable when the opposite inequality holds. There may exist equilibria involving more than one type of behavior, although the system is too complex to determine whether or not this is the case. Extensive simulations suggest that if such equilibria exist, they are not stable. I shall assume this is the case in this paper. It follows that for $t>2/(1-u)-1$, **aaC** is a globally stable equilibrium.§§

There are four plausible conditions that render the **bbD** equilibrium unstable, in which case **aaC** will be globally stable. The first is $u<0$, which means that the apparatus upon which internalization depends has net positive (pleiotropic) fitness effects independent from its contribution to the internalization of norms. The second is that t is sufficiently large that $(1+t)(1-u)>2$. Third, if parental transmission is sufficiently biased in favor of **C**, the internalization equilibrium is globally stable.

The fourth condition leading to the global stability of the **aaC** equilibrium is that there is some assortative mating that overcomes the tendency of the internalization allele to become "diluted". Suppose each type mates with another of its type with probability ζ, and with an random member of the population with probability $1-\zeta$. Then the eigenvalues of the

§§ The above result depends on our assumption of unbiased parental transmission. Suppose, however, that a fraction of offspring who would acquire norm **C** under unbiased transmission in fact acquire **D**. In this case, inspection of the eigenvalues of the Jacobian tells us that the **aaC** equilibrium is locally stable provided $\delta<t$ and the **bbD** equilibrium is stable provided $(1-\delta)(1+t)(1-u)<2$. Thus, parental transmission biased against the internalizable norm **C** is hostile to internalization.

bbD equilibrium become

$$\{\tfrac{1}{2}\zeta(1-u), \tfrac{1}{2}\zeta(1+t)(1-u), 1-u,$$
$$\tfrac{1}{2}\zeta(1+t)(1-u)(1+\zeta)\}. \quad (1)$$

Therefore, there is always a degree of assortative mating that renders the **bbD** equilibrium unstable. Thus, it is plausible that some combination of assortative mating, parental transmission biased towards the internal norm, and high returns to the internal norm, assures the global stability of the **aaC** equilibrium.‖‖

Altruism

We now add a second dichotomous phenotypic trait with two variants. Internal norm **A** is altruistic in the sense that its expression benefits the group, but imposes fitness loss $s \in (0,1)$ on those who adopt it. The normless state, **B**, is neutral, imposing no fitness loss on those who adopt it, but also no gain or loss to other members of the social group.

We assume **A** has the same cultural transmission rules as **C**: individuals who have a copy of allele **a** inherit their phenotypes from their parents, while **bb** individuals always adopt the normless phenotype **BD**. In addition, there is extraparental transmission, as before. There are now three genotypes and four phenotypes, giving rise to nine phenogenotypes that can occur, which we denote by **aaAC**, **aaAD**, **aaBC**, **aaBD**, **abAC**, **abAD**, **abBC**, **abBD**, and **bbBD**, and three that cannot occur, **bbAC**, **bbAD**, and **bbBC**. We represent the frequency of phenogenotype i by $p(i)$, for $i =$ **aaAC**, …,**bbBC**.

We maintain the assumption that families are formed by random pairing and the offspring genotype obeys Mendelian segregation. We assume also that only the phenotypic traits of parents, and not which particular parent expresses them, are relevant to the transmission process. Therefore there are nine family phenotypes, which can be written as **AACC**, **AACD**, **AADD**, **ABCC**, **ABCD**, **ABDD**, **BBCC**, **BBCD**, and **BBDD**. It follows that there are 54 familial phenogenotypes, which we can write as **aaaaAACC**,…,**bbbbBBDD**, only 36 of which can occur. We write the frequency of familial phenogenotype j as $p(j)$, and we assume the population is sufficiently large that we can ignore random drift. For illustrative purposes, here are a few of the phenogenotypic frequencies:

$$p(\textbf{aaaaAACC}) = p(\textbf{aaAC})^2(1-s)^2(1+t)^2(1-u)^2/\bar{p},$$

$$p(\textbf{aaaaAACD}) = p(\textbf{aaAC})p(\textbf{aaAD})(1-s)^2(1+t)(1-u)^2/\bar{p},$$

$$p(\textbf{aaaaAACD}) = 2p(\textbf{abAC})p(\textbf{abBD})$$
$$p(\textbf{abAD})p(\textbf{abBC}))(1-s)(1+t)(1-u)^2/\bar{p},$$

$$p(\textbf{bbbbBBDD}) = p(\textbf{bbBD})^2/\bar{p},$$

and so on, where $\bar{p}$ is chosen so the sum of the frequencies is unity:

$$\bar{p} = p(\textbf{aaaaAACC}) + \cdots + p(\textbf{bbbbBBDD}).$$

The rules of cultural transmission are as before. If familial phenogenotype is **xyzwXYZW**, where **x,y,z,w** $\in$ {**a**,**b**}, **X,Y** $\in$ {**A**,**B**}, and **Z,W** $\in$ {**C**,**D**}, an offspring is equally likely to inherit **xz**, **xw**, **yz**, or **yw**. An offspring whose genotype includes a copy of the **a** allele is equally likely to inherit **X** or **Y**, and equally likely to inherit **Z** or **W**. Offspring of **bb** geneotype always have the normless phenotype **BD**, unless they are socialized extraparentally.¶¶ The transition table is shown in Fig. 4.

We assume both genotypic and phenotypic fitness, as well as their interactions, are multiplicative. Thus, the fitness of the nine phenogenotypes that can appear with positive frequency are as shown in Fig. 5. The resulting system consists of eight equations in eight of the nine offspring phenogenotypes. One offspring phenogenotype is dropped, since the sum of phenogenotype frequencies must be unity.

It is straightforward to check that there are five pure equilibria. These are **aaAC**, in which all agents internalize both the altruistic and fitness enhancing norms, **aaAD**, in which only the altruistic norm is internalized, **aaBC**, in which

‖‖ The same results hold for a haploid version of the model, except that there is no unit root in the **aC** equilibrium. Moreover, if **a** is not completely dominant in the diploid model, the unit root disappears and the equilibrium is unambiguously stable.

¶¶ Extensive simulations show that if **a** is incompletely dominant, the results described below continue to hold.

Familial type	Offspring Phenogenotypic Frequency								
	aaAC	**aaAD**	**aaBC**	**aaBD**	**abAC**	**abAD**	**abBC**	**abBD**	**bbBD**
aaaaAACC	1								
aaaaABCC	1/2		1/2						
aaaaBBCC			1						
aaaaAACD	1/2	1/2							
aaaaABCD	1/4	1/4	1/4	1/4					
aaaaBBCD			1/2	1/2					
aaaaAADD		1							
aaaaABDD		1/2		1/2					
aaaaBBDD				1					
aaabAACC	1/2				1/2				
aaabABCC	1/4		1/4		1/4		1/4		
aaabBBCC			1/2				1/2		
aaabAACD	1/4	1/4			1/4	1/4			
aaabABCD	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8	
aaabBBCD			1/4	1/4			1/4	1/4	
aaabAADD		1/2				1/2			
aaabABDD		1/4		1/4		1/4		1/4	
aaabBBDD				1/2				1/2	
aabbABCD					1/4	1/4	1/4	1/4	
aabbBBCD							1/2	1/2	
aabbABDD						1/2		1/2	
aabbBBDD								1	
abbbABCD					1/8	1/8	1/8	1/8	1/2
abbbBBCD							1/4	1/4	1/2
abbbABDD						1/4		1/4	1/2
abbbBBDD								1/2	1/2
ababAACC	1/4				1/2				1/4
ababABCC	1/8		1/8		1/4		1/4		1/4
ababBBCC			1/4				1/4		1/2
ababAACD	1/8	1/8			1/4	1/4			1/4
ababABCD	1/16	1/16	1/16	1/16	1/8	1/8	1/8	1/8	1/4
ababBBCD			1/8	1/8			1/4	1/4	1/4
ababAADD		1/4				1/2			1/4
ababABDD		1/8		1/8		1/4		1/4	1/4
ababBBDD				1/4				1/2	1/4
bbbbBBDD									1

FIG. 4. Cultural and biological transition parameters.

only the fitness-enhancing norm is internalized, **aaBD**, in which agents carry the gene for internalization of norms, but no norms are in fact internalized, and **bbBD**, in which internalization is absent, and neither altruistic nor fitness-enhancing norms are transmitted from parents to offspring. A check of the eigenvalues of the Jacobian matrix shows that the **aaAD** and the **aaBD** equilibria are unstable.

The Jacobian of the altruistic internalization equilibrium **aaAC** has eigenvalues

$$\left\{0, 1, \frac{1}{2(1+t)}, \frac{1}{(1+t)}, \frac{1-\gamma}{1-s}, \frac{1-\gamma}{2(1-s)(1+t)}, \frac{1-v}{4(1-s)(1+t)}, \frac{1-v}{4(1-s)(1+t)}\right\}.$$

It is easy to check that the linearization of this equilibrium is stable if $s<\gamma$, since the unit root is semisimple. We cannot conclude that the equilibrium itself is necessarily stable for all parameters s, t, u, γ, and v satisfying the above inequalities. However, many simulations under varying parameter sets have failed to turn up an instance of instability.***

Individual Phenogenotype	Individual Fitness	Individual Phenogenotype	Individual Fitness
aaAC	(1-u)(1-s)(1+t)	aaAD	(1-u)(1-s)
aaBC	(1-u)(1+t)	aaBD	(1-u)
abAC	(1-u)(1-s)(1+t)	abAD	(1-u)(1-s)
abBC	(1-u)(1+t)	abBD	(1-u)
bbBD	1		

FIG. 5. Payoffs to nine phenogenotypes.

The eigenvalues of the Jacobian for the **aaBC** equilibrium are

$$\left\{0, 1, \frac{1}{2(1+t)}, \frac{1}{(1+t)}, \frac{1}{2(1+t)^2(1-u)^2}, \frac{1}{2}(1-s), \frac{1-s}{4(1+t)}, 1+\gamma-s\right\}.$$

*** When allele **a** is incompletely dominant, the unit root disappears, so the **aaAC** equilibrium is unambiguously stable.

Thus, **aaBC** is stable when $\gamma < s$, and unstable when the opposite inequality holds.

Finally, the eigenvalues of the Jacobian for the **bbBD** equilibrium are

$$\{0, 0, 0, 0, 1-u, \tfrac{1}{2}(1+t)(1-u),$$

$$\tfrac{1}{4}(1-s)(1+t)(1-u), \tfrac{1}{2}(1-s)(1-u)\}.$$

As in the single phenotype case, this is unstable if $u<0$ or $(1+t)(1-u)>2$, and is stable if either of the opposite inequalities hold. Moreover, it can be shown that adding assortative mating leads to the instability of the **bbBD** equilibrium under the same conditions as in the single phenotype case, shown in eqn (1).

In sum, under plausible conditions, one internalization equilibrium is stable—the altruism equilibrium when $\gamma > s$ and the nonaltruism equilibrium when $s > \gamma$. Since we expect s to be small, whereas the ubiquity of extraparental transmission favors a high γ, the altruism equilibrium appears the more plausible of the two. Under not implausible conditions, either a high return to the internal norm, assortative mating of agents that internalize, or pleiotropism in the form $u<0$, the only stable equilibrium of the system involves internalization. The dynamics, which we present below, support this conclusion.

Copying High-Fitness Phenotypes: The Replicator Dynamic

The above models of cultural transmission have been strongly criticized in the literature for suggesting that agents adopt norms *independent of their perceived payoffs*. In fact, people do not always blindly follow the norms that have been inculcated in them, but at least at times treat compliance as a strategic choice (Wrong, 1961; Gintis, 1975). The "oversocialized" model of the individual developed above may be improved by adding a phenotypic copying process reflecting the fact that agents shift from lower to higher payoff strategies. We represent this process as a *replicator dynamic* (Taylor and Jonker, 1978; Samuelson, 1997; Nowak and Sigmund, 1998; Gintis, 2000b). In the current context, there are four phenotypes whose relative fitness ranks them as **BC** > **AC** > **BD** > **AD**, and only agents with a copy of the **a** allele will copy another phenotype, since only such types are capable of internalizing a norm, and non-internalizers will not desire to mimic internalizers.

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

It is easy to see that adding a replicator dynamic does not change the single phenogenotype equilibria. By checking the eigenvalues of the Jacobian matrix, we find that the **aaAD** and **aaBD** equilibria remain unstable, and the replicator dynamic does not affect the conditions for stability of the unnormed equilibrium **bbBD**. The condition $\gamma > s$ for stability of the altruism equilibrium **aaAC** now becomes

$$\alpha < \frac{\gamma - s}{1 - \gamma}, \tag{2}$$

so a sufficiently strong replicator dynamic can undermine the stability of the **aaAC** equilibrium.††† The condition $s > \gamma$ for stability of the non-altruism internalization equilibrium **aaBC** when the replicator dynamic is included now becomes

$$\alpha > \frac{\gamma - s}{1 + \gamma - s},$$

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

In sum, adding a replicator dynamic changes the stability properties of the model in only one important way: a sufficiently strong replicator process can render the non-altruistic yet internalized equilibrium **aaBC**, rather than the

††† This model also has a semisimple unit root, so stability was checked by extensive simulations. A similar result holds for the haploid model.

‡‡‡ This model also has a semisimple unit root, so stability was checked by extensive simulations. A similar result holds for the haploid model, except the relevant inequality for stability of **aaBC** becomes the much stronger, and hence implausible, inequality $\alpha > (\gamma - s)/s(1 + \gamma - s)$.

altruistic equilibrium, **aaAC**, stable. Realistically, while the replicator process is key to understanding social change, in general we expect this to be a relatively weak force, and certainly too weak to undermine altruistic norms, unless they incur substantial fitness costs. Norms do not come labeled "altruistic norm", "instrumental practice". Rather, they are inextricably intermingled. It is quite common to believe that immoral acts lead to disease, for instance, just as does poor hygiene. Human psychology has to be uncritical absorbers of hosts of beliefs and values, only a small fraction of which can be seriously questioned by an individual member of society conditions us.

Why is Altruism Predominantly Prosocial?

Internal norms may be either pro- or anti-social. Indeed, there are many accounts of social norms that are severely socially costly, such as those involving invidious displays of physical prowess (Edgerton, 1992). The reason for the feasibility of anti-social norms is that once the internalization gene has evolved to fixation, there is nothing to prevent group-harmful phenotypic norms, such as our **A**, from also emerging, provided they are not excessively costly in comparison with the strength of the replicator process. The evolution of these phenotypes directly reduces the overall fitness of the population.

Yet as Brown (1991) and others have shown, there is a tendency in virtually all successful societies for cultural institutions to promote prosocial and eschew anti-social norms. The most reasonable explanation for the predominance of prosocial norms is *gene-culture coevolutionary multi-level selection*: societies that promote prosocial norms have higher survival rates than societies that do not (Parsons, 1964; Cavalli-Sforza & Feldman, 1981; Boyd & Richerson, 1985; Boyd & Richerson, 1990; Soltis *et al.*, 1995). Note that the usual arguments against the plausibility of genetic group selection do not apply to our model. This is because altruism is (a) phenotypic, and (b) "hitchhikes" on the fitness-enhancing phenotypic norm. Because altruism is phenotypic, and because a high degree of cultural uniformity can be maintained within groups, a high ratio of between-group to within-group variance on the phenotypic trait is easily maintained, and hence high-payoff groups quickly outpace low-payoff groups. Because altruism hitchhikes, the mechanism that generally undermines group selection, a high rate of inter-group migration (Maynard Smith, 1976; Boorman & Levitt, 1980) does not undermine internalization, as long as altruistic individuals adopt the **A** norms of the groups to which they migrate.

To test this argument, I 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 256 groups, each initially comprising 100 members, arranged spatially on a torus (a 16×16 grid with the opposite edges identified). Each group was seeded with ten **aaAC** types, ten **aaBC** types, 74 **bbBD** types, and one each of the other possible types. In all groups, $t = 0.3$ and $u = 0.05$. Each group was then randomly assigned a value of γ between 0 and 0.60, a value of α between 0 and 0.5, a value of s between 0.01 and 0.10, and a value of ζ (the degree of assortative mating) between 0 and 0.70. Each group was also assigned an **A** phenotype with a fitness effect between -1 and 1, such that if a group's **A**-fitness effect is q, and if a fraction f of the group exhibit the **A** phenotype, then each member of the group has its fitness augmented by fq.

In each round, for each of the 256 groups, I simulated the model as described in the previous sections, and update the frequencies of the various types in each group, according to the fitness effect of their **A** phenotype and the fraction of the group that exhibits this phenotype. A fraction of each group (typically 5%) then migrated to a neighboring group. If altruistic migrants adopt the **A** norm of their new groups, migration never undermines the stability of the altruism equilibria. To be conservative, we assume here that agents take their genes with them to a new group, and they take their **C** phenotype to this new group (since **C** is the same for all groups), but they abandon their **A** phenotype when they migrate (e.g. an **aaAC** type becomes a **aaBC** type in the new group). This assumption is maximally geared to

undermine the altruism phenotype since immigrants never exhibit altruism. We then allow for some random drift in the individual groups parameters s, α, γ, and ζ, as well as the payoff of altruism to the group, q.

Our final modeling assumption is that when group size drops below a minimum (generally, I set this to zero or ten agents), it is replaced by a copy of a randomly chosen other group.

I ran this model many times with varying numbers of rounds, and varying the parameters described above. The system always stabilized by 100 periods, and the specific assumptions concerning the parameters were never critical. The following conclusions hold for these simulations:

(a) Groups exhibiting the non-internalization equilibrium **bbBD** were quickly driven from the population, except under the joint assumptions that altruistic agents become non-altruistic when they migrate, and the migration rate is over 20%.

(b) The equilibrium fraction of groups for which the non-internalization equilibrium was stable was highly variable and dependent upon the specifics of the initial distribution of groups. Thus in equilibrium, though all groups were near the internalization equilibria (**aaAC** and **aaBC**), in some this was globally stable and in others, only locally.

(c) The equilibrium fraction of the population exhibiting the altruistic phenotype was greater than 85% (the mean at the start of each simulation was 10%).

(d) All but the highest prosocial **A** phenotypes were eliminated from the population, so that the mean fitness effect of the altruistic phenotypes was greater than 0.9 (the maximum possible was 1.0, and the mean at the start of the simulation was approximately zero). In particular, no groups with anti-social norms ever survived in equilibrium.

(e) The mean level of extraparental socialization, γ was also very high, being at least 45% (the mean at the start of the simulation was 30%).

(f) The mean strength of the replicator dynamic, α, was 9% (the mean at the start of the simulation was 25%). This shows that while a higher α helps individuals, because they are then more likely to move to high-fitness phenotypes, it hurts the groups they are in, and on balance lowers group fitness.

(g) The altruistic equilibrium was attained as long as the initial average assortative mating probability ζ was at least 12.5%. An increase in the rate of assortative mating led to more globally stable altruism equilibria, but had no measurable effect on the equilibrium values of other variables.

(h) The emergence of the **aaAC** agents and the elimination of anti-social internal norms were both due to population growth alone. The extinction and replacement of groups by more successful groups accounted only for the change in the frequency of γ and α (extinctions began to occur after 20 rounds, and more than 1500 extinctions typically occurred in a 100 round simulation).

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

Altruistic Punishment Can Sustain Cooperation When Altruistic Participation Cannot

Consider the following "social dilemma".‖‖‖ Each member of a group can either cooperate or shirk. Shirking costs nothing, but adds nothing to the payoffs of the group members. Cooperating costs $s^* > 0$, but contributes an amount $f^* > s^*$ shared equally by the other members. Selfish individuals will always shirk in this situation, so the potential gains from cooperating will be forgone. If the situation is repeated sufficiently frequently with the same players, and if the group

‖‖‖ This section is an elaboration on Boyd *et al.* (2001).

is sufficiently small, cooperation can be sustained even with selfish players (Trivers, 1971; Axelrod & Hamilton, 1981). However, with large groups and/or infrequent repetition, universal shirking is virtually inevitable (Boyd & Richerson, 1988), as has been confirmed repeatedly in experiments with humans (Ledyard, 1995).

Given the potential gains to society of people internalizing the altruistic norm **A**="always cooperate" in the above situation, the absence of this norm in society suggests that the cost s^* is simply too high to sustain as an equilibrium of the **aaAC** form. However, experimental results (Fehr & Gächter, 2002) and ethnographic data (Boehm 1993, 2000), not to mention everyday observation, suggest that the threat of being punished by other group members for shirking may serve to sustain cooperation where the internalized value of cooperating does not.

Punishment can succeed where the norm of cooperation cannot because the expected cost per period of punishing shirkers is typically much smaller that the cost of cooperating. This is because (a) punishment such as shunning and ostracizing are inherently low cost, yet effective when directed by large numbers against a few transgressors; and (b) the punishment need be carried out only when shirking occurs, which is likely to be infrequent in comparison with the number of times cooperation must be carried out.

Nevertheless, altruistic punishment likely has strictly positive cost, so a selfish individual still will refrain from engaging in this activity. While a genetic group selection model can explain the evolutionary stability of altruistic punishment (Gintis, 2000b), such models are sensitive to group size and migration rates (Eshel, 1972; Rogers, 1990). The gene-culture coevolutionary model presented in this paper, by contrast, suffers less from these problems.

To see this, suppose a fraction p of a group with n members consists of altruistic punishers. To prevent intentional shirking by selfish agents, each must be prepared to inflict a punishment s^*/pn on a shirker. Suppose a fraction q of the group nevertheless shirks (or perhaps is simply perceived to shirk under conditions of imperfect information). Then the total amount of punishment per altruistic punishment is $s = qs^*/p$. If p is large (as in our simulations) and q is small (as is likely to be the case except under extreme conditions, since no one has an incentive to shirk), then this value of s will be close to zero for each altruistic punisher. But then the altruism equilibrium will be stable according to eqn (2), even when it would be violated for $s = s^*$.

Since the fitness costs of altruistic punishment are low, a replicator dynamic is unlikely to render the altruism equilibrium unstable in this case. Moreover, there is evidence that altruistic acts serve as costly signals of agent fitness (Gintis *et al.*, 2003), in which case the altruistic phenotype is cannot be undermined by the tendency to shift from lower to higher payoff phenotypes.

Conclusion

We have developed a plausible model of altruistic cooperation and punishment that does not depend on repeated interaction, reputation effects, or multi-level selection. The latter obtains because there is no net within-group penalty to either the altruistic gene or the altruistic norm, even though there is a penalty to individuals carrying the gene and behaving according to the norm.

One shortcoming of our model is that payoffs are assumed constant, whereas in many cases, we would expected payoffs to be frequency dependent, as when group members are engaged in a non-cooperative game. For instance, the payoff to being self-interested may increase when agents are predominantly altruistic. In a related paper (Gintis, 2003), I show that such a situation gives rise to a heterogeneous equilibrium, in which both altruists and self-interested types participate. Since the payoffs to the two types are equal in equilibrium, once again we can dispense with multi-level selection in specifying an equilibrium with a positive level of altruism.

There are two objections that biologists naturally raise to this model of altruism. First, if the **C** norm is individually fitness enhancing while the **A** is not, why is there not a genetic mutation (for instance at another genetic locus) that allows the individual to distinguish between altruistic and fitness-enhancing behaviors, and hence to eschew the former? The answer is that

A- and **C**- type behaviors are exhibited only on the phenotypic level, and hence have no clear inherent characteristics according to which such a gene could discriminate. Moreover, if such an inherent characteristic does exist for a particular **A**-type norm, that type would be driven to extinction. But there are so many varieties of cultural norms that others, unaffected by this mutation, would arise to replace the one to which people have become "immune". Finally, we should note that generally the *degradation* of the genetic capacity to discriminate is much more likely than the *emergence* of such a capacity, since the latter, being complex in the case of **A**-type norms, requires the existence of a sequence of one-point mutations that are each fitness-enhancing, finally leading to the capacity to discriminate. This is implausible in the current context.

A second objection to our model of altruism is that we have assumed rather than provided an explanation of why the internalization of norms—having a programmable objective function—is individually fitness enhancing. Why would an agent gain from an altered objective function when he always has the option of obeying a norm when it is his interest to do so, and violating the norm when it is not? However, agents do not maximize fitness, but rather an objective function that is itself subject to selection. In a constant environment, this objective function will track fitness closely. In a changing environment, natural selection will be too slow, and the objective function will not track fitness closely. Cultural transmission and the ensuing increase in social complexity produced such a rapidly changing environment in human groups. Imitation (the replicator dynamic) will not correct this failure, because agents copy objective-function-successful, not fitness-successful, strategies. In this situation, there are large fitness payoffs to the development of a *non-genetic mechanism for altering the agent's objective function*, together with a *genetic mechanism for rendering the individual susceptible to such alteration*. Internalization of norms, which may be an elaboration upon imprinting and imitation mechanisms in non-human animals, doubtless emerged by virtue of its ability to alter the human objective function in a direction conducive to higher fitness. There is not to my knowledge a confirmed instance of internalization in nonhuman animals. This may in part be due to the fact that the relevant research has not been carried out. Yet there are obvious reasons to doubt that internalization might be important, because cultural transmission in nonhuman animals is relatively rudimentary.

There is not to my knowledge a confirmed instance of internalization in non-human animals. This may in part be due to the fact that the relevant research has not been carried out. Yet there are obvious reasons to doubt that internalization might be important except perhaps in the case of animals with a long history of domestication by humans (dogs come to mind). This is because cultural transmission in animals is relatively rudimentary. The proximate cause of complex behaviors in animals lies in genetically expressed mechanisms, the ultimate cause being the contribution of these mechanisms to the fitness of the individuals who express them. Since culture is very important in the fitness of humans, internal norms become the proximate cause of complex behaviors in humans, but the ultimate cause of the capacity to internalize, and the content of internal norms themselves, in the first instance remains the same: the capacity to enhance the fitness of the individuals who express them. However, we have seen that there is a *second instance* in this case: altruistic norms can hitchhike on personally fitness-enhancing norms. Were this not the case, human society as we know it would not exist.

It would be a serious mistake to conclude that the socialization process in humans is sufficiently powerful to permit *any* pattern of norms to be promulgated by internalization. For instance, many have suggested that it would be better if people acted on the principle of contributing to society according to one's ability, and taking from society according to one's needs. Whatever the moral standing of such a principle, no society has lasted long when its incentives have been based on it. Our model suggests one reason why such a principle might fail: the operation of the replicator dynamic. In this case, the payoff to defectors from the norm is simply too high to prevent its erosion. There may be other criteria determining what types of altruistic

norms are likely to emerge from the gene-culture coevolutionary process described in this paper. For instance, behaviors that are altruistic, but very similar to ones that are personally fitness enhancing may be relatively easy to internalize; e.g. since it is generally fitness enhancing to speak truthfully, it may be relatively easy to move the decision to speak truthfully from the realm of instrumental calculation to that of the moral realm of right and wrong. Similarly, altruistic punishment may be widespread because it is generally prudent to develop a reputation for punishing those who hurt us, and it is a short step to turning this prudence into a moral principle.

I would like to thank Robert Boyd, Marcus Feldman, Samuel Bowles, Marci Gintis, Eric Alden Smith, John E. Stewart, and Claus Wedekind for helpful comments, and the John D. and Catherine T. MacArthur Foundation for financial support.

REFERENCES

Axelrod, R. & Hamilton, W. D. (1981). The evolution of cooperation. *Science* **211,** 1390–1396.

Boehm, C. (1993). Egalitarian behavior and reverse dominance hierarchy. *Curr. Anthropo.* **34,** 227–254.

Boehm, C. (2000). *Hierarchy in the Forest: The Evolution of Egalitarian Behavior*. Cambridge, MA: Harvard University Press.

Boorman, S. A. & Levitt, P. (1980). *The Genetics of Altruism*. New York: Academic Press.

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

Boyd, R. & Richerson, P. J. (1988). The evolution of cooperation. *J. theor. Biol.* **132,** 337–356.

Boyd, R. & Richerson, P. J. (1990). Group selection among alternative evolutionarily stable strategies. *J. theor. Biol.* **145,** 331–342.

Boyd, R. & Richerson, P. J. (1992). Punishment allows the evolution of cooperation (or anything else) in sizeable groups. *Ethol. Sociobiol.* **113,** 171–195.

Boyd, R., Gintis, H., Bowles, S. & Richerson, P. J. (2001). Altruistic punishment in large groups evolves by interdemic group selection. Working Paper.

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

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

Edgerton, R. B. (1992). *Sick Societies: Challenging the Myth of Primitive Harmony*. New York: The Free Press.

Eshel, I. (1972). On the neighbor effect and the evolution of altruistic traits. *Theor. Popul. Biol.* **3,** 258–277.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature* **415,** 137–140.

Feldman, M. W., Cavalli-Sforza, L. L. & Peck, J. R. (1985). Gene-culture coevolution: models for the evolution of altruism with cultural transmission. *Proc. Natl Acad. Sci.* **82,** 5814–5818.

Gintis, H. (1975). Welfare economics and individual development: a reply to talcott parsons. *Quart J. Econ.* **89,** 291–302.

Gintis, H. (2000a). *Game Theory Evolving*. Princeton, NJ: Princeton University Press.

Gintis, H. (2000b). Strong reciprocity and human sociality. *J. theor. Biol.* **206,** 169–179.

Gintis, H. (2003). Solving the puzzle of human prosociality. *Rationality and Society* **15,** forthcoming.

Gintis, H., Smith, E. A. & Bowles, S. (2001). Costly signaling and cooperation. *J. Theor. Biol.* **213,** 103–119.

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

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

Maynard, S. J. (1976). Group selection. *Quar. Rev. Biol.* **51,** 277–283.

Mealey, L. (1995). The sociobiology of sociopathy. *Behav. Brain Sci.* **18,** 523–541.

Nowak, M. A. & Sigmund, K., (1998). Evolution of indirect reciprocity by image scoring. *Nature* **393,** 573–577.

Parsons, T. (1964). Evolutionary universals in society. *Am. Sociol. Rev.* **29,** 339–357.

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

Richerson, P. J., Boyd, R. & Bettinger, R. L. (2001). Was agriculture impossible during the pleistocene but mandatory during the holocene? A climate change hypothesis. *American Antiguity* **66,** 387–411.

Rogers, A. R. (1990). Group selection by selective emigration: the effects of migration and kin structure. *Am. Nat.* **135,** 398–413.

Samuelson, L. (1997). *Evolutionary Games and Equilibrium Selection*, Cambridge, MA: MIT Press.

Simon, H. (1990). A mechanism for social selection and successful altruism. *Science* **250,** 1665–1668.

Sober, E. & Wilson, D. S. (1998). *Unto Others: The Evolution and Psychology of Unselfish Behavior*, Cambridge, MA: Harvard University Press.

Soltis, J., Boyd, R. & Richerson, P. (1995). Can group-functional behaviors evolve by cultural group selection: an empirical test. *Curr. Anthropo.* **36,** 473–483.

Taylor, P. & Jonker, L. (1978). Evolutionarily stable strategies and game dynamics. *Math. Biosci.* **40,** 145–156.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *Quar. Rev. Biol.* **46,** 35–57.

Wrong, D. H. (1961). The oversocialized conception of man in modern sociology. *Am. Sociol. Rev.* **26,** 183–193.