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

Herbert Gintis

SFI WORKING PAPER: 2001-10-058

SANTA FE INSTITUTE

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

Herbert Gintis

January 9, 2002

Abstract

The *internalization of norms* refers to the tendency of human beings to adopt social norms from parents (*vertical transmission*) or socializing institutions (*oblique transmission*). *Authority* rather than *contribution to fitness* accounts for the adoption of internalized norms. Suppose there is one genetic locus that controls whether or not an individual is capable of internalizing norms. We extend classical models (Cavalli-Sforza and Feldman 1981, Boyd and Richerson 1985) to show that *if adopting a norm is fitness enhancing, fixation of the allele for internalization is locally stable, and with a small amount of oblique transmission, fixation is globally stable*. We use this framework to model Herbert Simon's (1990) explanation of altruism. Simon suggested that altruistic norms could 'hitchhike' on the general tendency of the internalization of norms to be fitness-enhancing. We show that *the altruistic phenotype evolves if and only if there is a sufficient level of oblique transmission of internalizable norms*. This result holds even when there is a strong horizontal transmission process biased against the altruistic norm. We then use a gene-culture coevolutionary group selection argument to explain why internalized traits are likely to be pro- as opposed to anti-social.

*Department of Economics, University of Massachusetts and Santa Fe Institute, hgintis@mediaone.net, http://www-unix.oit.umass.edu/~gintis. I would like to thank Robert Boyd, Marcus Feldman, Eric Alden Smith, John E. Stewart, and Claus Wedekind for helpful comments, and the John D. and Catherine T. MacArthur Foundation for financial support.

1 Introduction

A *norm* is a phenotypically expressed rule governing social behavior. The *internalization of norms* refers to the tendency of human beings to adopt norms from parents (*vertical transmission*) or socializing institutions (*oblique transmission*). *Authority* rather than *contribution to fitness* explains the internalization of norms (Cavalli-Sforza and Feldman 1973, Feldman and Cavalli-Sforza 1976). This paper provides an analytical model and computer simulations of the internalization of norms, based on the coevolution of genes and culture (Lumsden and Wilson 1981, Feldman, Cavalli-Sforza and Peck 1985, Boyd and Richerson 1985, Feldman and Cavalli-Sforza 1987, Aoki and Feldman 1987, Durham 1991, Feldman and Zhivotovsky 1992).[1]

Suppose there is one genetic locus that controls whether or not an individual is capable of internalizing norms from parents and elders. We extend classical models (Cavalli-Sforza and Feldman 1981, Boyd and Richerson 1985) to show that *if internalizing a norm is fitness enhancing, the allele for internalization can evolve to fixation through vertical transmission*. This is a Baldwin effect (Baldwin 1896, Simpson 1953, Ancel 1999)—a genetic mechanism that predisposes the individual to adopt a fitness-enhancing phenotypic trait. We then add a fitness-sensitive horizontal transmission mechanism—the replicator dynamic familiar from evolutionary game theory—showing that the tendency of agents to switch from lower to higher-payoff norms enlarges the basin of attraction of the internalization allele.

By *altruism* we mean 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.[2] We use our framework to model Herbert Simon's (1990) explanation of altruism. Simon argued that altruistic norms, which are by definition fitness-reducing to their carriers, could 'hitchhike' on the general tendency of the internalization of norms to be fitness-enhancing. We show that *the altruistic phenotype can evolve if there is a sufficient level of oblique transmission*, even when there is a strongly fitness-biased horizontal transmission process. We then use a gene-culture coevolutionary group selection argument to explain why internalized traits are likely to be pro- as opposed to anti-social.

Society's values are transmitted through the internalization of norms (Durkheim

[1] Our use of the term 'oblique transmission' differs from that of Cavalli-Sforza and Feldman (1981) and Boyd and Richerson (1985), who use a *biological* conception of bilateral transmission from *influential elders*. We use oblique transmission in the *sociological* sense of internalization *via* social institutions, such as churches, schools, or tribal rituals. The difference is significant only one situation, which is discussed below.

[2] For reviews of the evidence on the importance of altruism in human societies, see Sober and Wilson (1998), Bowles and Gintis (1998), Fehr and Gächter (forthcoming), Gintis (2000a), and Gintis (2000b).

1951, Boas 1938, Benedict 1934, Mead 1963, Geertz 1963, Parsons 1967, Grusec and Kuczynski 1997). All known cultures foster norms that enhance personal fitness, such as prudence, personal hygiene, and control of emotions. Cultures also universally promote norms that subordinate the individual to group welfare, fostering such behaviors as bravery, honesty, fairness, willingness to cooperate, refraining from overexploiting a common pool resource, voting and otherwise participating in the political life of the community, acting on behalf of one's ethnic or religious group, and identifying with the goals of an organization of which one is a member, such as a business firm or a residential community (Brown 1991).

Of course, even when socially approved norms are fitness-reducing both for individuals and groups, they may persist, and people may widely and voluntarily conform to such norms and punish those who do not (Boyd and Richerson 1992, Kollock 1997, Glaeser, Laibson, Scheinkman and Soutter 2000, Tonkiss and Passey 2000). The models developed herein extend naturally to this phenomenon as well. As we shall argue below, the prevalence of pro- over anti-social norms is due to the ability of groups with prosocial norms to outcompete groups with antisocial norms.

In our first model, there is a dichotomous phenotypic trait, one version of which, (**C**), is a norm that enhances personal fitness and another, (**D**), is the absence of the norm, and is fitness-neutral. The norm **C**, can be internalized from parents. The capacity for internalization is encoded in a single diploid genetic locus, with two alleles **a** and **b**. We treat **a** as dominant, in the sense that an agent with at least one copy of **a** can acquire norm **C** by vertical transmission from parents, while **bb**-types cannot, and hence have phenotype **D** whatever the familial phenotypes.[3]

In our second model, we add a second dichotomous phenotypic trait, one version of which (**A**) is an altruistic norm—group-beneficial but personally fitness-reducing—and the other version, (**B**), the absence of the norm, is fitness-neutral. This trait is controlled by the same genetic locus, so **A** can be inherited from parents by agents who have at least one copy of allele **a**. Offspring of genotype **bb** thus have phenotype **B** whatever the phenotype of their parents. We find that this altruistic trait is doomed to extinction if only vertical transmission and fitness-sensitive horizontal transmission are available, but if we add oblique transmission through socialization institutions, the altruistic phenotype can evolve. *Altruistic traits can thus hitchhike on the fitness-enhancing capacity of internalization, but only if extra-familial, prosocial, cultural institutions are sufficiently powerful.*

[3]The model also works under incomplete dominance, but we here restrict discussion to the complete dominance case.

2 Norm Internalization and Vertical Cultural Transmission

Suppose there is a norm **C** that can be internalized by a new member of society through vertical transmission. Norm **C** confers fitness $(1 + t > 1)$, while the normless phenotype, which we denote by **D**, has baseline fitness 1. We assume that possessing a copy of the internalization allele **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. There are thus six phenogenotypes, whose fitnesses are listed in Figure 1.[4]

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

Figure 1: Fitnesses of the Six 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 under vertical or oblique transmission, **bbC** cannot occur after the first generation.

We assume families are formed by random pairing, and offspring genotype obeys 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 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 appear after the first generation. The frequency of familial phenogenotypes are as shown in Figure 2, where $p(i)$ represents the frequency of phenogenotype i, so p(**aaC**) is the frequency of **aaC** individuals, and so on.

The rules of cultural 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 Figure 3.[5]

[4] 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.

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

Figure 2: Frequencies of Phenogenotypes. Here, $\overline{p}$ is chosen so the sum of the frequencies is unity. Note that **aabbCC**, **abbbCC**, **bbbbCC**, and **bbbbCD** are not listed, since they cannot occur after the first generation. The table also assume $p(\mathbf{bbC}) = 0$.

The resulting system consists of four equations in four unknowns—four of the six offspring phenogenotypes (the offspring phenogenotype **bbC** disappears after the first generation for any initial distribution of familial phenogenotypes, and one offspring phenogenotype is dropped, since the sum of phenogenotype frequencies must be unity). It is straightforward to check that there are three 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 in fact no parent has the norm **C**, 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

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

[5]Biased vertical transmission, in which heterogeneous familial phenotypes are more likely to transmit one phenotype to offspring that the other (Cavalli-Sforza and Feldman 1981) is discussed below.

Familial Type	Offspring Phenogenotypic Frequency					
	aaC	**aaD**	**abC**	**abD**	**bbC**	**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		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

Figure 3: Phenotypic Inheritance is Controlled by Genotype. Note that **aabbCC**, **abbbCC**, **bbbbCC**, and **bbbbCD** are not listed, since they cannot occur after the first generation.

The unit eigenvalue is semisimple,[6] 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 show that the **aaC** equilibrium is stable for all permissible parameter values.[7] One surprising implication is that the stability of the **aaC** equilibrium does not depend on the fitness cost u of the internalization allele, although the size of the basin of attraction does.[8] For example, if we set $t = 0.2$ and $u = 0.3$, agents with the internalization allele are quite severely disadvantaged whether or not they have the norm **C**. Nevertheless, a simulation shows that the **aaC** equilibrium is stable (although with a small basin of attraction).[9] Indeed, in this case Fisher's Fundamental Theorem (Fisher 1930)

[6] An eigenvalue is semisimple if its algebraic and geometric dimensions are equal. Semisimple unit roots are stable.

[7] 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.

[8] This occurs because we have treated **b** as recessive. If **b** is additive this result disappears.

[9] Throughout this paper, the "basin of attraction" of a phenogenotype i is the set of initial conditions, consisting of a fraction of type i and the rest consisting of allele **b** and the fitness-neutral phenotypes alone, that converge to an equilibrium in which all agents are of genotype **aa**.

fails, since if we seed the population with all **aaC** except for a small number of **bbD**, average fitness is higher than at fixation with all **aaC**. Simulations show that fitness declines in the first period and increases monotonically thereafter, although I have not sought to prove this analytically.

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

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

Therefore this equilibrium, in which no internalization occurs, is locally stable if $(1+t)(1-u) < 2$, and locally 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. If this is the case, it follows that for $t > 2/(1-u) - 1$, **aaC** is a globally stable equilibrium.

The inequality $t > 2/(1-u) - 1$ implies an extremely high return to the internalizable norm, so the latter result is not strong. Suppose, however, we add oblique transmission, assuming a fraction $\gamma \geq 0$ of **aa**-types and **ab**-types who have inherited the selfish behavior **D** from their parents, are induced by socialization institutions to switch to the norm **A**. A check of the appropriate eigenvalues indicates that, as long at $u < 2t/(1+t)$, which is extremely plausible, then **aaC** is globally stable if and only if

$$\gamma > \frac{u}{1-u}\frac{1-t+u(1+t)}{2t-u(1+t)}.$$

For instance, if $u = t/4$, then $\gamma > 0.14$ guarantees global stability. We should note that if we use the biological definition of oblique transmission, the latter assertion fails, and the **aaC** equilibrium is never globally stable.

The above result depends on our assumption of unbiased vertical transmission. Suppose, however, that a fraction δ of offspring who would acquire norm **B** 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 vertical transmission biased against the internalizable norm **A** is hostile to internalization.[10]

This model of vertical transmission has been widely criticized for suggesting that agents adopt norms independent of their fitness effects. In fact, people do

[10]Anticipating our treatment of horizontal transmission, we note that if the rate of horizontal transmission is α and the rate of oblique transmission is γ, then the condition for stability of the **aaC** equilibrium is

$$\delta < \frac{1+t}{(1-\alpha)(1-\gamma)} - 1,$$

so both horizontal transmission in favor of higher fitness phenotypes, and oblique transmission of the internalizable norm are favorable to internalization.

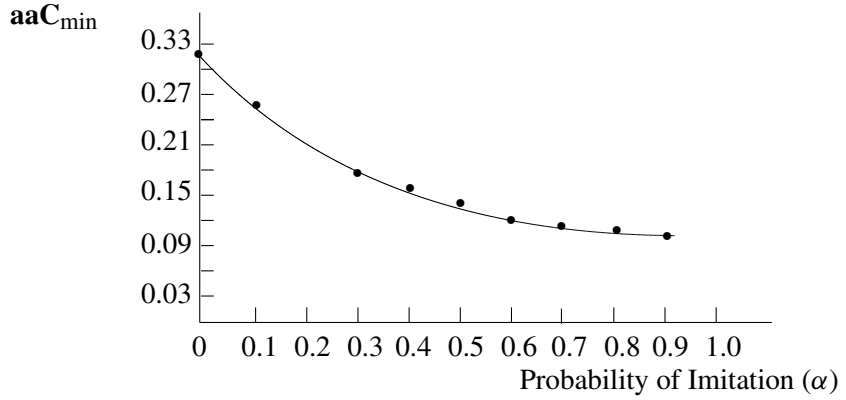

Figure 4: The Effect of the Replicator Dynamic on the Accessibility of the Internalization Equilibrium **aaC**. Here, $\mathbf{aaC}_{\min}$ is the endpoint of the basin of attraction of the internalization (**aaC**) equilibrium.

not always blindly follow social rules, but at least at times treat compliance as a strategic choice (Wrong 1961, Gintis 1975). We remedy our "oversocialized" model of choice behavior, adding a phenotypic imitation process, whereby agents shift from lower payoff to higher payoff strategies. We represent this process as a *replicator dynamic* (Taylor and Jonker 1978, Samuelson 1997, Nowak and Sigmund 1998, Gintis 2000b). We assume with some probability α an agent observes the phenotype and fitness of another agent, and switches to that phenotype if the other agent's payoff is higher than his own. Defining $p_A = \alpha(p(\mathbf{aaC}) + p(\mathbf{abC}) + p(\mathbf{bbC}))$, this give rise to the following set of equations, where primed values are post-horizontal transmission:

$$
\begin{aligned}
p(\mathbf{aaC})' &= p(\mathbf{aaC}) + p_A p(\mathbf{aaD}) \\
p(\mathbf{abC})' &= p(\mathbf{abC}) + p_A p(\mathbf{abD}) \\
p(\mathbf{bbC})' &= p(\mathbf{bbC}) + p_A p(\mathbf{bbD}) \\
p(\mathbf{aaD})' &= p(\mathbf{aaD})(1 - p_A) \\
p(\mathbf{abD})' &= p(\mathbf{abD})(1 - p_A) \\
p(\mathbf{bbD})' &= p(\mathbf{bbD})(1 - p_A)
\end{aligned}
$$

We find that adding the replicator dynamic enlarges the basin of attraction of the internalization equilibrium **aaC**. This is illustrated in Figure 4, where $\mathbf{aaC}_{\min}$ is the lower bound of the basin of attraction of the **aaC** equilibrium.

We conclude that

a. There are three equilibria in which the whole population bears a single phenogenotype: **aaC**, **aaD**, and **bbD**. The **aaD** equilibrium is unstable. The Jacobian of the

aaC equilibrium has a semisimple unit root, so we cannot ensure by analytical means that the equilibrium is stable. However, extensive simulations show that this equilibrium is locally stable for all permissible parameter values.

b. The **bbD** is locally stable if $(1 + t)(1 - u) < 2$, and locally unstable when the opposite inequality holds. It follows that for $t > 2/(1 - u) - 1$, **aaC** is a globally stable equilibrium.

c. for plausible values of u, a sufficiently high level (usually small) of oblique transmission renders the **aaC** equilibrium globally stable.

d. Adding a replicator dynamic (phenotypic imitation of high-fitness strategies) enlarges the basin of attraction of the internalization equilibrium **aaC**.

3 Altruism and Vertical Transmission

Suppose we add a second dichotomous phenotypic trait with two variants. 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, which we denote by **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 norm **A** has the same cultural transmission rules as norm **C**. In effect, individuals who have the internalization allele simply inherit their phenotypes from their parents, while individuals without the internalization allele always adopt normless phenotype **BD**. Such individuals may later change phenotype through a process of imitating more successful phenotypes, but we leave this issue for later. Since there are now three genotypes and four phenotypes, there are twelve phenogenotypes, which we denote by **aaAC**, **aaAD**, **aaBC**, **aaBD**, **abAC**, **abAD**, **abBC**, **abBD**, **bbAC**, **bbAD**, **bbBC**, and **bbBD**. We represent the frequency of phenogenotype i by $p(i)$, so $p($**aaAC**$)$ is the frequency of **aaAC** individuals, and so on. Note that **bbAC**, **bbAD**, and **bbBC** cannot arise through vertical or oblique transmission.

We maintain the assumptions that families are formed by random pairing and the offspring genotype obeys the laws of 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 fifty-four familial phenogenotypes, which we can write as **aaaaAACC**...**bbbbBBDD**, not all of which can arise through vertical and oblique transmission. 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, we list a few of the fifty four familial phenogenotypic frequencies:

$$
\begin{aligned}
p(\mathbf{aaaaAACC}) &= p(\mathbf{aaAC})^2(1-s)^2(1+t)^2(1-u)^2/\overline{p},\\
p(\mathbf{aaaaAACD}) &= p(\mathbf{aaAC})^2(1-s)^2(1+t)(1-u)^2/\overline{p},\\
p(\mathbf{ababABCD}r &= 2p(\mathbf{abAC})\,p(\mathbf{abBD})\\
&\qquad + p(\mathbf{abAD})(\mathbf{abBC}))(1-s)(1+t)(1-u)^2/\overline{p},\\
p(\mathbf{bbbbBBDD}) &= p(\mathbf{bbBD})^2/\overline{p},
\end{aligned}
$$

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

$$\overline{p} = p(\mathbf{aaaaAACC}) + \ldots + p(\mathbf{bbbbBBDD}).$$

The rules of cultural transmission are as follows. If familial phenogenotype is $xyzwXYZW$, where $x, y, z, w \in \{\mathbf{a}, \mathbf{b}\}$, $X, Y \in \{\mathbf{A}, \mathbf{B}\}$, and $Z, W \in \{\mathbf{C}, \mathbf{D}\}$, 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. But and offspring of genotype **bb** always has the normless phenotype **BD**. The transition table is shown in Figure 6 which, to save space, does not include familial and offspring phenogenotypes that cannot occur after the first generation through vertical and oblique transmission.

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 Figure 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.

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		

Figure 5: Payoffs to Nine Phenogenotypes. Note that types **bbAC**, **bbAD**, and **bbBC** cannot arise from vertical or oblique transmission.

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				
aaabBBCC			1/4				1/4		
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				1/4				1/4	
aabbABCD					1/4	1/4	1/4	1/4	
aabbBBCD							1/2	1/2	
aabbABDD						1/2		1/2	
aabbBBDD									
abbbABCD								1	
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

Figure 6: Cultural and Biological Transition Parameters. Familial types that cannot arise after one generation of vertical and oblique transmission are not depicted.

It is straightforward to check that there are five equilibria in which the whole population bears a single phenogenotype. 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 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 matrices of the model shows that **aaAC**, **aaAD** and **aaBD** are all unstable. The linearized version of the equilibrium **aaBC**, in which the fitness-enhancing norm is internalized but the altruistic norm is not, is stable. However, the Jacobian matrix of the system has a single semi-simple unit root, so we cannot conclude from the stability of the linearized system that the nonlinear model is stable. Extensive simulations show that it is stable. The equilibrium **bbBD** is stable for $t < 2/(1-u) - 1$, and unstable when the opposite inequality holds.

We conclude that *unbiased vertical transmission alone cannot lead to the expression of the altruistic phenotype.*

4 Oblique Transmission and the Stability of Altruism

Suppose we add *oblique transmission* to the model. Specifically, we assume a fraction $\gamma \geq 0$ of **aa**-types and a fraction $\nu \geq 0$ of **ab**-types who have inherited the selfish behavior **B** from their parents, are influenced by oblique transmission to switch to the altruistic behavior **A**.[11]

The Jacobian of the full internalization equilibrium **aaAC** has nonzero eigenvalues

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

It is easy to check that this equilibrium is stable if $s \leq \gamma$, $s \leq (1+\nu)/2$, $1-\gamma \leq 2(1-s)(1+t)$, and $1-\nu \leq 4(1-s)(1+t)$. Except for the requirement $s \leq \gamma$, these are all very weak conditions, satisfied for instance if $s < 0.50$. The condition $s < \gamma$ adds specificity to our conclusion from the last section: assuming $s < 0.50$,

[11] We do not investigate oblique transmission in favor of the normless phenotype, because we know by our previous analysis that the resulting equilibrium will be the global stability of the normless phenotype. Note that the assumption that γ does not depend on the fraction of the population with the **C** phenotype reflects our assumption that oblique transmission is *via* socialization institutions rather than influential elders that carry the **C** trait.

the internalization of norms can sustain altruism in equilibrium only if there is a strictly positive rate of oblique transmission of norms.

Since the Jacobian of the social equilibrium has a unit root, we cannot conclude that this equilibrium is necessarily stable for all parameters s, t, u, γ, and ν satisfying the above inequalities. However, many simulations under varying parameter sets have failed to turn up an instance of instability. Indeed, the basin of attraction of the social equilibrium is normally quite large. For instance, setting $t = 0.2$, $s = 0.05$, $u = 0.01$, and $\gamma = \nu = 0.11$, we find $\mathbf{aaAC}_{\min} \approx 0.37$.

The contrast between the model without the altruistic norm **A** and the current model is dramatic. In the earlier model, oblique transmission was unnecessary to achieve stability of the internalization equilibrium. In the current case, while oblique transmission is necessary for the full internalization equilibrium **aaAC** to be locally stable, even high levels of oblique transmission do not render this equilibrium globally stable. Indeed, simulations indicate that $\mathbf{aaAC}_{\min}$ does not fall below about 0.35 for plausible values of the parameters of the model.

A check of the eigenvalues of the Jacobian for the **bbBD** equilibrium shows why this is the case. It can be shown that this equilibrium is unstable when

$$s < -\frac{1 + u + \nu(1 - u)}{(u(1 - \nu) + 2\nu)(1 - u)},$$

and is stable when the opposite inequality holds. Since this expression is negative, *the norm* **A** *must be fitness-enhancing for the* **aaAC** *equilibrium to be globally stable*.

We should note that if we use the biological notion of oblique transmission (from influential elders) rather than the sociological (*via* socialization institutions), the normless equilibrium is always locally stable. Thus when the current level of expression of the altruistic phenotype is very low, socialization institution can ensure altruism will increase, whereas influential elder transmission cannot.

We would expect that adding a replicator dynamic to this model would both increase the basin of attraction of the internalization allele, and lower the equilibrium frequency of the altruistic norm. Simulations show that this is in fact the case. For instance, setting $t = 0.2$, $s = 0.05$, $u = 0.01$, and $\gamma = \nu = 0.11$, we find that the minimum initial fraction of the population having the **a** allele that leads to the fixation of the **a** allele falls from 37% with no social imitation, to 14% when agents switch to higher payoff types with probability $\alpha = 0.5$, and then to only 9% when $\alpha = 0.9$. The equilibrium frequency of altruism is also strongly inversely related to the strength of the replicator process, as shown in Figure 7.

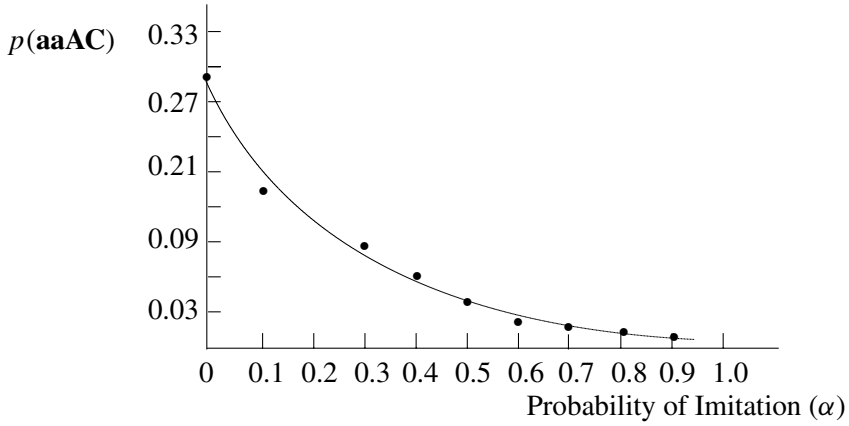

Figure 7: The Effect of the Replicator Dynamic on the Level of Altruism in the Internalization Equilibrium involving **aaAC** and **aaBC**. The fitness-enhancing norms **C** is fully expressed in all cases. This simulation assumed $t = 0.2$, $s = 0.05$, $u = 0.01$, and $\gamma = \nu = 0.11$,

5 Why is Altruism Predominantly Prosocial?

Internalization is universal in human societies, but internalized altruistic norms may be either pro- or anti-social. Indeed, there are many accounts of social norms that are severely socially costly (invidious displays of wealth or physical prowess, glorification of aggression, beliefs in supernatural causes of and remedies for sickness and crop failure, to mention a few). The reason for this is that once the internalization gene has evolved to fixation, there is nothing to prevent non-fitness-enhancing phenotypic norms, such as our **A**, from also emerging, provided they are not excessively costly. 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 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 group selection*: societies that promote prosocial norms have higher survival rates than societies that do not (Parsons 1964, Cavalli-Sforza and Feldman 1981, Boyd and Richerson 1985, Boyd and Richerson 1990, Soltis, Boyd and Richerson 1995). Note that the usual arguments against the plausibility of genetic group selection do not apply to our model at all. First, the internalization of norms is a purely individual selection argument. Second, the key condition facilitating group selection, a high ratio of between-group to within-group variance, is easily maintained, since the traits that evolve occur on the phenotypic level, in

the form of norms that can be relatively uniform with social groups. Third, the mechanism that generally undermines group selection, a high rate of migration across groups (Maynard Smith 1976, Boorman and Levitt 1980), can be offset by strengthening socialization institutions (increasing γ).

6 Conclusion

If phenotypically expressed norms are fitness-enhancing, a gene favoring the internalization of these norms through vertical transmission from parents is locally stable, and for plausible parameter values, is globally stable with sufficient oblique transmission. Adding a replicator dynamic, which responds to the fitness of phenotypes, enhances the process of internalization.

When we add a fitness-reducing "altruistic" phenotypic norm, we find that vertical transmission alone is incapable of rendering the internalization equilibrium even locally stable. Adding a small amount of oblique transmission does render internalization locally stable, and simulations show that with plausible parameter values, the basin of attraction of the internalization equilibrium is quite large. However, the basin of attraction of the non-internalization equilibrium is also large, and is not significantly reduced by increasing the amount of oblique transmission. Adding a replicator dynamic does considerably reduce the size of the non-internalization equilibrium, but it also reduces the amount of altruism in equilibrium rather strongly.

These observations give us a plausible story for the emergence and incidence of altruistic behavior of the sort described by Simon (1990). Because many phenotypic norms are fitness-enhancing for humans living in complex social environments, genes promoting the internalization of norms from parents and socializing institutions evolve in a manner involving a Baldwin effect: the fitness value of the phenotype leads the genome to incorporate alleles that strongly promote the acquisition of the phenotype.

References

Ancel, L. W., "A Quantitative Model of the Simpson-Baldwin Effect," *Journal of Theoretical Biology* 196 (1999):197–209.

Aoki, Kenichi and Marcus W. Feldman, "Toward a Theory for the Evolution of Cultural Communication: Culture Coevolution of Signal Transmission and Reception," *Proceedings of the National Academy of Sciences* 84 (October 1987):7164–7168.

Baldwin, J. M., "A New Factor in Evolution," *American Naturalist* 30 (1896):441–451.

Benedict, Ruth, *Patterns of Culture* (Boston: Houghton Mifflin, 1934).

Boas, Franz, *General Anthropology* (Boston: Heath, 1938).

Boorman, Scott A. and Paul Levitt, *The Genetics of Altruism* (New York: Academic Press, 1980).

Bowles, Samuel and Herbert Gintis, "The Moral Economy of Community: Structured Populations and the Evolution of Prosocial Norms," *Evolution & Human Behavior* 19,1 (January 1998):3–25.

Boyd, Robert and Peter J. Richerson, *Culture and the Evolutionary Process* (Chicago: University of Chicago Press, 1985).

— and — , "Group Selection among Alternative Evolutionarily Stable Strategies," *Journal of Theoretical Biology* 145 (1990):331–342.

— and — , "Punishment Allows the Evolution of Cooperation (or Anything Else) in Sizeable Groups," *Ethology and Sociobiology* 113 (1992):171–195.

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

Cavalli-Sforza, L. and M. W. Feldman, "Models for Cultural Inheritance: Within Group Variation," *Theoretical Population Biology* 42,4 (1973):42–55.

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

Durham, William H., *Coevolution: Genes, Culture, and Human Diversity* (Stanford: Stanford University Press, 1991).

Durkheim, Emile, *Suicide, a Study in Sociology* (New York: Free Press, 1951). Translated by John A. Spaulding and George Simpson. Edited, with an Introduction by George Simpson.

Fehr, Ernst and Simon Gächter, "Altruistic Punishment in Humans," *Nature (*forthcoming).

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

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

— and — , "Towards a Theory for the Evolution of Cultural Communication: Coevolution of Signal Transmission and Reception," *Proceedings of the National Academy of Sciences* 84 (1987):7164–7168.

— , Luca L. Cavalli-Sforza, and Joel R. Peck, "Gene-Culture Coevolution: Models for the Evolution of Altruism with Cultural Transmission," *Proceedings of the National Academy of Sciences* 82 (1985):5814–5818.

Fisher, Ronald A., *The Genetical Theory of Natural Selection* (Oxford: Clarendon Press, 1930).

Geertz, Clifford, *Peddlers and Princes: Social Change and Economic Modernization in Two Indonesian Towns* (Chicago: University of Chicago Press, 1963).

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

— , *Game Theory Evolving* (Princeton, NJ: Princeton University Press, 2000).

— , "Strong Reciprocity and Human Sociality," *Journal of Theoretical Biology* 206 (2000):169–179.

Glaeser, Edward, David Laibson, Jose A. Scheinkman, and Christine L. Soutter, "Measuring Trust," *Quarterly Journal of Economics* 65 (2000):622–846.

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

Kollock, Peter, "Transforming Social Dilemmas: Group Identity and Cooperation," in Peter Danielson (ed.) *Modeling Rational and Moral Agents* (Oxford: Oxford University Press, 1997).

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

Maynard Smith, John, "Group Selection," *Quarterly Review of Biology* 51 (1976):277–283.

Mead, Margaret, *Sex and Temperament in Three Primitive Societies* (New York: Morrow, 1963).

Nowak, Martin A. and Karl Sigmund, "Evolution of Indirect Reciprocity by Image Scoring," *Nature* 393 (1998):573–577.

Parsons, Talcott, "Evolutionary Universals in Society," *American Sociological Review* 29,3 (June 1964):339–357.

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

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

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

Simpson, John Gaylord, "The Baldwin Effect," *Evolution* 7 (1953):110–117.

Sober, Elliot and David Sloan Wilson, *Unto Others: The Evolution and Psychology of Unselfish Behavior* (Cambridge, MA: Harvard University Press, 1998).

Soltis, Joseph, Robert Boyd, and Peter Richerson, "Can Group-functional Behaviors Evolve by Cultural Group Selection: An Empirical Test," *Current Anthropology* 36,3 (June 1995):473–483.

Taylor, P. and L. Jonker, "Evolutionarily Stable Strategies and Game Dynamics," *Mathematical Biosciences* 40 (1978):145–156.

Tonkiss, Fran and Andrew Passey, *Trust and Civil Society* (New York: St. Martin's Press, 2000).

Wrong, Dennis H., "The Oversocialized Conception of Man in Modern Sociology," *American Sociological Review* 26 (April 1961):183–193.