A Mechanism for Social Selection and Successful Altruism

HERBERT A. SIMON

Within the framework of neo-Darwinism, with its focus on fitness, it has been hard to account for altruism, behavior that reduces the fitness of the altruist but increases average fitness in society. Many population biologists argue that, except for altruism to close relatives, human behavior that appears to be altruistic amounts to reciprocal altruism, behavior undertaken with an expectation of reciprocation, hence incurring no net cost to fitness. Herein is proposed a simple and robust mechanism, based on human docility and bounded rationality, that can account for the evolutionary success of genuinely altruistic behavior. Because docility—receptivity to social influence—contributes greatly to fitness in the human species, it will be positively selected. As a consequence, society can impose a "tax" on the gross benefits gained by individuals from docility by inducing docile individuals to engage in altruistic behaviors. Limits on rationality in the face of environmental complexity prevent the individual from avoiding this "tax." An upper bound is imposed on altruism by the condition that there must remain a net fitness advantage for docile behavior after the cost to the individual of altruism has been deducted.

IT IS OF NO LITTLE MOMENT FOR THE HUMAN FUTURE WHETHER people are necessarily and consistently selfish, as is sometimes argued in population genetics and economics, or whether there is a significant place for altruism in the scheme of human behavior. Do centrally important institutions like business and government depend entirely on motivating participants through their selfish interests in order to operate successfully? Is reciprocal altruism (actually a form of self interest) the only kind that can survive?

In recent years there have been many attempts to derive theoretical answers to these questions from the first principles of natural selection (*1*). Most of the answers give a central, almost exclusive, role to self-interest, and, apart from altruism to close kin, leave little room for genuine, as distinct from reciprocal, altruism.

The proposal in this paper can be read as an "even if" argument. Even if we accept the genes of individual persons as the controlling sites for natural selection—the assumption most antagonistic to altruism—a mechanism can be described that selects for altruistic behavior well beyond altruism to close kin and beyond support from expected reciprocity or social enforcement. The mechanism will select for behavior that reduces the fitness of the altruist while increasing average fitness in the society.

The argument does not deny the existence of social mechanisms for transmitting behavior traits; in fact, socially learned behavior is central to the theory. Nor is it concerned with the many forms of behavior usually called "altruistic" that are unrelated to biological fitness. The argument shows that even though altruistic behavior, strictly defined, is penalized, altruism can still be positively selected.

Essentially, the theory accounts for altruism on the basis of the human tendency (here called docility) to learn from others (more accurately, the tendency to accept social influence)—which is itself a product of natural selection. Because of the limits of human rationality, fitness can be enhanced by docility that induces individuals often to adopt culturally transmitted behaviors without independent evaluation of their contribution to personal fitness.

The author is professor of computer science and psychology, Department of Psychology, Carnegie-Mellon University, Pittsburgh, PA 15213.

Altruism

By altruism I mean behavior that increases, on average, the reproductive fitness of others at the expense of the fitness of the altruist. Fitness simply means expected number of progeny. An exchange in which both parties are compensated for what they initially cede does not count as altruism but as enlightened self-interest (sometimes called soft or reciprocal altruism). Still, the boundaries are tricky, as we shall see.

Notice that "altruism" and "selfishness" in genetics bear no close resemblance to these terms in everyday language. Presumably, Don Juan was fitter than Croesus or Caesar. From a genetic standpoint, the amassing of wealth or power does not count at all toward fitness, only the amassing of progeny. By the same token, liberality with wealth or willingness to cede power do not constitute genetic altruism. Altruism means forgoing progeny.

We could debate at some length whether, either at the present time or earlier in the history of our species, wealth and power have or had any strong connection with genetic fitness. If the connection is weak, then the evolutionary argument that people are essentially selfish in the everyday sense of that word—that is, striving only for economic gain, power, or both—is correspondingly weakened. Under those circumstances, there could be any amount of altruism, in the usual sense of that term, without any behavior that would qualify as altruistic in a genetic sense.

In this article, I am concerned with fitness, altruism, and selfishness only in the genetic meanings of those terms. In the concluding section I will return briefly to desire for wealth and power as human motives. In any event, our goal is not to establish how much or how little altruism, in either sense, there is in human behavior, but rather to show that altruism on a substantial scale is not inconsistent with the strictest neo-Darwinian assumptions.

The Neo-Darwinian Analysis

The acceptance by many modern geneticists of the axiom that the basic unit of selection is the "selfish gene" quickly led to the production of population models that left little room for the survivability of altruistic behavior (*2*). If altruism incurred any cost in fitness, that is, in reduced potential or reproduction, then it could not compete against selfishness.

To be sure, it was recognized that altruism was viable under several specific (and rather narrow) conditions. First, altruism toward close relatives could increase fitness through the genes shared with those relatives. But the closest relatives (except identical twins) have only half their genes in common, and this fraction drops by a factor of two with each step of distance in the relationship. Consanguinity can account for altruism only toward close kin (*3*).

The second qualification is that, if several mixed societies (trait groups) contain varying fractions of altruists and non-altruists, then (i) the groups with the larger fractions of altruists may outbreed the groups with smaller fractions, (ii) as a result, the fraction of altruists in the entire population may increase for some time, (iii) even though the fraction of altruists in each separate group will necessarily decrease (*4*).

Of course, if the groups inbreed, then, in the long run, as the least altruistic (and least successful) groups became extinct or nearly so, the number of altruists in the entire population would begin to decrease, and altruism would ultimately become extinct. If, however, the population members periodically mixed thoroughly for purposes of reproduction, then the fraction of altruists in the total could continue to increase indefinitely.

All of these results can be formalized with relatively simple mathematical models. I will borrow heavily from these mathematical formulations, but my assumptions will be different from those in the model just described (*5*).

In addition to the models mentioned above, several explicit theories analyze the co-evolution of culturally transmitted and genetically transmitted traits. Among the most prominent of these are the theories of Cavalli-Sforza and Feldman, Lumsden and Wilson, and Boyd and Richerson (*6–8*). I will discuss them after I have presented my own model.

A Simple Model of Altruism

Consider a population consisting of n individuals, of two types, A and S, in proportions p and $1 - p$, respectively. The individuals of type A are altruistic, while those of type S are selfish. Each A expresses a behavior that contributes b offspring to members of the population (including himself), the recipients being chosen at random. The cost of this altruistic behavior is that each A has c fewer children than he or she otherwise would have. The average number of offspring, F_A, and F_S, of each A and S will be: $F_A = X - c + bp$, and $F_S = X + bp$, where X is the number of offspring in the absence of altruistic behaviors, the same for both types of individuals. All individuals, including altruists, can serve as recipients to the np altruists, and selfish S individuals incur no cost of altruism. Since c is positive, selfish individuals always have more offspring than altruistic ones. To the degree that the behaviors are heritable, selfish individuals will therefore be found with greater relative frequency in each succeeding generation.

Notice that the total contribution of each altruist to the population is b, assumed independent of the size of the population. Under an alternative assumption, which does not affect our main conclusions, each altruist contributes b to the fitness of each member of the population, thereby making the total contribution of the altruist bn, where n is the size of the population. In this latter case, the contribution is a "public good"—its consumption by one member does not decrease the amount available to others. (An attractive garden visible to passersby is an example.)

As was mentioned earlier, if there are a number of groups instead of one, and if the groups are segregated during most of their life cycle but intermingle thoroughly while reproducing, then altruists may have greater net fitness than non-altruists and may grow in numbers at the expense of the latter. Systems with this property are called "structured demes," and mathematical models of them are examined in considerable detail by Wilson (*4*).

Social Learning and Altruism

With only a single change of assumption, which I will now motivate, my simple model can be converted into one in which altruists are fitter than selfish individuals even within a single, self-contained population that is not a structured deme. In this system, altruism will not only survive, but will gradually permeate the entire population (*9*).

The human species is notable, although not unique among animals, in requiring for survival many years of nurture by adults. In most human societies, the survival and fitness even of adults depends on the assistance, or at least forbearance, of other adults. Leaving aside active hostility from others, even access to food and shelter cannot be ensured in most societies without the consent of others.

The human species also has a notable ability to learn, and especially to learn from other people, particularly with the help of language. We will use the term "social learning" to refer to learning from others in the society.

Social learning makes two major contributions to an individual's fitness. First, it provides knowledge and skills that are useful in all of life's activities, in particular, in transactions with the environment. Second, goals, values, and attitudes transmitted through social learning, and exhibited in the speech or behavior of the learner, often secure supportive responses from others. For brevity, we will call the knowledge and skills of the first kind "skills," and those of the second kind "proper behaviors."

Learning of both kinds obviously contributes to fitness. We will use the term "docile" (in its dictionary meaning of "disposed to be taught") to describe persons who are adept at social learning, who accept well the instruction society provides them. Individuals differ in degree of docility, and these differences may derive partly from genetic differences. There are differences in intelligence (cognitive ability to absorb what is taught) and in motivation (propensity to accept or reject instruction, advice, persuasion, or commands).

Docile persons tend to learn and believe what they perceive others in the society want them to learn and believe. Thus the content of what is learned will not be fully screened for its contribution to personal fitness. This tendency derives from the difficulty—often an impossibility—for individuals to evaluate beliefs for their potential positive or negative contribution to fitness. For example, many of us believe that less cholesterol would be beneficial to our health without reviewing (or even being competent to review) the medical evidence. Hundreds of millions of people believe that behaving in a socially acceptable way will enhance the probability of enjoying blissful immortality.

Belief in large numbers of facts and propositions that we have not had the opportunity or ability to evaluate independently is basic to the human condition, a simple corollary of the boundedness of human rationality in the face of a complex world. We avoid most hot stoves without ever having touched them. Most of our skills and knowledge, we learned from others (or from books); we did not

discover or invent them. The contribution of docility to fitness is enormous.

Guilt and shame, although perhaps genetically independent of docility, also serve most people as strong motivators for accepting social norms. Guilt is particularly important because it can operate independently of the detection of nonconformity.

In analogy with earlier simple models, I assume a population made up of two kinds of people: those who are docile, D, and those who are not, S. We assume that both kinds of people are identical in fitness, except that docile people, because of the skills and proper behaviors they have acquired, produce an average d more offspring than the others. Thus, $F_D = X + d$, while $F_S = X$. Clearly, docile people will increase in relative number in the society.

Now if the society coexists in its environment with other societies, we may also compare the relative rates of growth of these societies. As in the models of qualified altruism that we have already examined, there may be certain altruistic behaviors that, although costly to the fitness of the individual who exhibits them, have more than a compensating advantage for other individuals in the society.

A society that instilled such behaviors in its docile members would grow more rapidly than one that did not; hence such behaviors would become, by evolution at the social level, a part of the repertory of proper behaviors of successful societies. Societies that did not develop such a repertory would be less fit than those that did, and would ultimately disappear. But could the altruism ultimately survive within the more successful societies?

To answer this question, I add altruism acquired by social learning to the model and see how docile-altruistic individuals fare relative to selfish ones. I will now simply call docile-altruistic individuals "altruistic," F_A, as in the previous models: $F_A = X + d - c + b(c)p$ and $F_S = X + b(c)p$. Again, p is the percentage of altruists in the population; X is the number of offspring in the absence of altruistic behaviors; d is the gross increase in A's offspring due to A's docility; c is the net cost to A, in offspring, of altruistic behavior acquired through the docility mechanism; $b(c)$, which replaces the b of the previous model, is the number of offspring contributed to the population of A's altruistic behavior. I express this number as a function of c, because the amount of altruism exacted from A, and its corresponding contribution of fitness to others, depend on the society's definition of proper behavior, itself subject to cultural evolution.

Under these assumptions, an individual who is docile, enjoying the advantage (d) of that docility, will consequently also accept the society's instructions to be altruistic as part of proper behavior. Because of bounded rationality, the docile individual will often be unable to distinguish socially prescribed behavior that contributes to fitness from altruistic behavior. In fact, docility will reduce the inclination to evaluate independently the contributions of behavior to fitness. Moreover, guilt and shame will tend to enforce even behavior that is perceived as altruistic. Hence the docile individual will necessarily also incur the cost, c, of altruism.

Now unlike the previous model, in this case, because $F_A - F_S = d - c$, the fitness of altruists will actually exceed the fitness of selfish individuals as long as d exceeds c, that is, as long as the demands for altruism that society imposes on docile individuals are not excessive compared with the advantageous knowledge and skills acquired through docility. If this condition is satisfied, the proportion of altruists will increase.

Suppose there are decreasing marginal returns from altruism, so that $d^2b/dc^2 < 0$. In the short run (that is, for fixed p), it will be optimal for the society to fix c at the level where $db/dc = 1$, but the long-run optimal strategy will be to demand less altruism initially so as to increase the absolute number of docile individuals as rapidly as possible, that is, to set $p(db/dc) = 1$. For small p, this implies that db/dc will be large, hence that c will be small or even zero [if $(dc/dc)_0 < (1/p)$]. As p grows, social demands on the altruists can be increased correspondingly—the greater the fraction of altruists in the society, the more altruistic it can be.

In this scheme of things, altruism is a relative matter, for only a subset of the altruist's behaviors reduce fitness. Moreover, the altruist is rewarded, in advance, by the "gift" of docility; altruism is simply a by-product of docility. Docile persons are more than compensated for their altruism by the knowledge and skills they acquire, and moreover not all proper behaviors are sacrificial. (Learning to drive in the right lane is a proper behavior, but not sacrificial.) The term "altruism" applies only to the sacrificial subset of the behaviors engendered by docility.

If docility were something the individual deliberately chose, one might even rename the accompanying altruism "enlightened selfishness." But docility (at least its genetic component) is bestowed, not chosen, and with the bestowal goes the propensity to adopt proper behaviors, including altruistic ones. By virtue of bounded rationality, the docile person cannot acquire the personally advantageous learning that provides the increment, d, of fitness without acquiring also the altruistic behaviors that cost the decrement, c.

Three final observations: first, altruism includes the effort individuals spend to induce and enforce learning and proper behavior in others. The docility mechanism will work only if there are providers of skills and knowledge as well as recipients. But nurturing and enforcing behaviors will be learned as an essential component of the proper behaviors of altruism. In enforcement are included carrots as well as sticks—praising and nurturing others who exhibit proper behavior, as well as frowning on, shunning, or otherwise punishing those who do not.

Second, the fitness advantage of altruists would be decreased if individuals could feign proper behavior without detection. (They would be motivated to do so only when they knew the behavior was altruistic.) There are probably severe limits, however, as to how far deception will be successful (*10*).

Third, the effectiveness of the docility mechanism would be impaired if individuals could discriminate perfectly proper behaviors that were "for their own good" from those that were altruistic. But people can discriminate only very imperfectly between beneficial and altruistic proper behaviors.

Moreover, much of the value of docility to the individual is lost if great effort is expended evaluating each bit of social influence before accepting it. Acceptance without full evaluation is an integral part of the docility mechanism, and of the mechanisms of guilt and shame.

Comparison with Alternative Models

I return now to the models of Cavalli-Sforza and Feldman, Lumsden and Wilson, and Boyd and Richerson and compare their mechanisms for altruism with the docility mechanism.

Cavalli-Sforza and Feldman (*6*, footnote 6), examining the interaction between cultural and genetic transmission of traits (*6*, pp. 102–107 and pp. 133–143), show that a selectively disadvantageous trait can spread to a whole population, where by a disadvantageous trait they mean "a maladaptive social custom (for example, one creating some degree of danger to life that is not compensated for by other advantages in Darwinian fitness) or a custom decreasing fertility . . . , or an infectious disease" (*6*, p. 106).

They do not consider, however, traits that, while maladaptive to individuals, confer net benefits on the population (altruistic behaviors); nor do they explain why negative selection of maladaptive social customs does not remove them, either by positive selection of those individuals who reject them, or by selection or social norms, or

both. Many sociobiologists would therefore regard their model as incomplete, holding constant things that evolutionary forces would change in the long run. The mechanism I have proposed avoids both of these difficulties.

Lumsden and Wilson provide no mechanism for altruism other than altruism toward close kin and reciprocal or "soft" altruism (*11*).

Boyd and Richerson (*8*, chap. 7, footnote 6) introduce a mechanism that produces altruism by "conformist transmission," which is, essentially, preferential selection of the behaviors individuals encounter most frequently. Conformist transmission has something in common with the docility mechanism, but differs from it in several crucial respects.

Degree of conformism, in the initial version of the Boyd and Richerson model (*8*, pp. 206–213) depends solely on frequency of exposure, without individual differences between conformers and defecters. If such differences are introduced for traits that are individually disadvantageous, there will be negative selection of conformers and positive selection of rejecters until the traits disappear.

The authors recognize this difficulty (*8*, p. 213) and introduce the possibility of individuals rejecting individual culturally transmitted traits. They then show that for rather special circumstances (involving migration among groups living in varying environments) conformist transmission (hence altruism) could be stably maintained.

But the docility mechanism I have proposed accounts for altruism even in a homogeneous environment, and does not depend on the frequency with which a trait is encountered. Finally, it is considerably simpler and more robust than conformist transmission, depending only on a couple of system parameters.

This review of these alternative theories of altruism shows that altruism based on docility provides a simpler mechanism, valid under a wider range of conditions, than the others.

Implications for Economics and Politics

The existence of heritable docility, and the consequent possibility for a society to cultivate and exploit altruism, has very strong implications for social theory, including economics, and the theories of political institutions and other organizations. I will mention just a few such implications as examples.

First, goals like gaining wealth and power might become very strong motivations even if they made no direct contribution to genetic fitness. If it were advantageous to the success of a society for people to seek wealth or power, then these could be taught and rewarded as proper behaviors. The dangers of early assassination (and consequent deficit of offspring) to those who exercise power could be absorbed in the term c, among the costs of altruism. In particular, the desire for glory becomes, in this framework, an understandable human motive.

Motives like wealth, power, and glory would be difficult to sustain if associated with major costs to fitness. They are readily sustained if they are both useful to the society and nearly neutral for individual fitness. Power motives might have net value to the society by providing leaders who enhance the society's ability to organize to exploit resources or defend against enemies. Wealth-amassing motives might be useful if they created more wealth than was drawn off by those who strove for gain.

Consider next an example from politics. It has been difficult to explain what self interest leads many people to go to the polls on election day. Any single vote is unlikely to change an election outcome, so it should seem pointless to a rational person to exert effort to vote. Even a small opportunity cost of casting a ballot is too much. But a society that includes voting among the proper behaviors can, at a minute cost to the fitness of altruists, secure their participation in elections.

Many other troublesome issues of public goods can be explained in the same way—contributions to charity and volunteer work being important examples. Of course other motives may also help to cause these behaviors. People may volunteer in order to make useful acquaintances. There are many possibilities, but no reason to rule out altruism as an important motivation.

Finally, many people exhibit loyalties to organizations and organization goals that seem wholly disproportionate to the material rewards they receive from the organization or its success (*12*). In particular, few people (including top executives) receive rewards from business firms that are proportional to the profits. Yet executives and other employees seem often to make decisions in terms of their expected effects on the firm's profitability. And empirical evidence suggests little difference in the relative efficiencies of profit-making and non-profit firms in the same industry (for example, health care, water supplies, education) (*13*). With profits or without, people often identify with organization goals and organizational survival.

All these topics deserve a more thorough treatment than they are given here. Mentioning them suggests what a wealth of possible behaviors opens up when we admit docility as a major mechanism of social transmission.

As a final caution, I repeat that what I have called altruism, a partial sacrifice of genetic fitness, may be very different from the forgoing of wealth and power that is called altruism in common discourse. Nothing in the model predicts that we will not see people attending to their economic interests in most of their everyday behavior; or for that matter, that we will not see them giving away a large part of the wealth they have taken great pains to amass.

In our century, we have watched two great nations, the Peoples' Republic of China and the Soviet Union, strive to create a "new man," only to end up by acknowledging that the "old man" —perhaps we should say the "old person"—self-interested and concerned with his or her economic welfare or the welfare of family, clan, ethnic group, or province, was still alive and well. It will be important to reexamine this striking historical experience, not in terms of oversimple models of the "selfish gene," but in a framework that acknowledges that altruism, either as defined socially or as defined genetically, is wholly compatible with natural selection and is an important determinant of human behavior.

REFERENCES AND NOTES

1. G. C. Williams, *Adaptation and Natural Selection: a Critique of Some Current Evolutionary Thought* (Princeton Univ. Press, Princeton, 1966); R. D. Alexander, *Annu. Rev. Ecol. Syst.* **5**, 325 (1974).
2. R. Dawkins, *The Selfish Gene* (Oxford Univ. Press, Oxford, 1976).
3. W. D. Hamilton, *J. Theor. Biol.* 7, 1 (1964).
4. D. S. Wilson, *The Natural Selection of Populations and Communities* (The Benjamin-Cummings Press, Menlo Park, CA, 1980).
5. The reader who wants to pursue the mathematics further will find Wilson's book (*4*, footnote 4, pp. 23–32) indispensable.
6. L. L. Cavalli-Sforza, M. W. Feldman, *Cultural Transmission and Evolution* (Princeton Univ. Press, Princeton, NJ, 1981).
7. C. Lumsden and E. O. Wilson, *Genes, Mind, and Culture* (Harvard Univ. Press, Cambridge, MA, 1981).
8. R. Boyd and P. J. Richerson, *Culture and the Evolutionary Process* (Univ. of Chicago Press, Chicago, IL, 1985).
9. The docility mechanism described in this section was introduced less formally by H. A. Simon, *Reason in Human Affairs* (Stanford Univ. Press, Stanford, CA, 1983).
10. This issue has been examined by R. H. Frank, *Passions Within Reason* (Norton, New York, NY, 1988).
11. C. J. Lumsden and E. O. Wilson, *Promethean Fire* (Harvard Univ. Press, Cambridge, MA, 1983), pp. 30–32.
12. H. A. Simon, *Administrative Behavior* (Macmillan, New York, ed. 3, 1976).
13. B. A. Weisbrod, *Science* **244**, 541 (1989).

14. I am very grateful to a number of geneticists and others who have contributed to my education on this subject by reading and commenting upon earlier drafts of this paper, including D. T. Campbell, J. F. Crow, R. C. Lewontin, D. S. Wilson, and E. O. Wilson. Supported by the Personnel and Training Programs, Psychological Sciences Division, Office of Naval Research, under contract N00014-86-K-0768.