HOW ALTRUISM CAN PREVAIL IN AN EVOLUTIONARY ENVIRONMENT

by Theodore C. Bergstrom* and Oded Stark**

Forschungsbericht/
Research Memorandum No. 320
April 1993

*Department of Economics
University of Michigan
Ann Arbor, MI 48109
USA

**Department of Economics
Harvard University
Cambridge, MA 02138
USA

We are grateful for generous advice and assistance from Carl Bergstrom, Gary Becker, and Andreu Mas-Collel.

Abstract

We study environments in which an individual gets a higher payoff from defecting than from cooperating and where "copies" of an individual are more likely to appear the higher is the individual's payoff. We demonstrate that even in single-shot prisoner's dilemma models (where cooperation benefits one's opponent at a cost to oneself), evolution can sustain cooperative behavior between relatives or neighbors.

In addition, we show that selfish individuals who conciously choose their actions may find it in their interest to be altruistic when there is some probability that their practices will be imitated.

Both genetic and cultural inheritance appear to be blunt instruments that do not operate on individuals in isolation. Those who inherit a genetic tendency to cooperate are more likely than others to enjoy the benefits of cooperative siblings. Similarly with culturel inheritance; altruism can prevail when individuals are likely to interact with others who share the same role model.

Why are economists convinced that *homo economicus* is selfish? No doubt we find considerable support for this hypothesis in the behavior of our colleagues. Beyond this, a plausible evolutionary argument for selfishness would assert that if natural selection favors those who receive high payoffs, and if altruists get lower payoffs than selfish individuals, then evolution will tend to eliminate altruists. In this paper, we will show that, paradoxically, evolution can sustain cooperative behavior between relatives or neighbors even in single-shot prisoner's dilemma models, where cooperation benefits one's opponent at a cost to oneself.[1]

In the first three sections of this paper, we consider two-player, two-strategy games in which a player who cooperates gets a payoff of R if his or her opponent cooperates, and S if the opponent defects. A player who defects gets T if his or her opponent cooperates, and P if the opponent defects. In a prisoner's dilemma game, $S < P < R < T$, so that *defect* is a dominant strategy for each player, and $S + T < 2R$, so that total payoffs are maximized when both cooperate. An individual's strategy is determined either by genetic inheritance or by imitating the behavior of parents or neighbors.

The discussion in the fourth section shows that selfish rational individuals who consciously choose their actions may find it in their interest to be altruistic when there is some probability that their practices will be imitated.

I. The Evolution of Genetically Transmitted Behavior toward Siblings

Not much is known about the environments that shaped our genes, and most economists do not believe that evolutionary hypotheses could explain human preferences. But since the fundamentals of mating, child-rearing, and sibling relations have changed little over the millennia, we believe that evolutionary theory can enrich the study of the economics of the family.[2]

---

[1] In this paper we identify *altruism* with playing *cooperate* in prisoner's dilemma. For explorations of more subtle connections between cooperative actions and preferences and the well-being of others, see Douglas Bernheim and Oded Stark (1988), Stark (1989), (1993), and Theodore Bergstrom (1989), (1992). Most human interaction occurs in environments that are more conducive to cooperation than prisoner's dilemma games. We have chosen the case of prisoner's dilemma in order to show that evolution can select for altruism even in a most hostile environment.

[2] We have good company in this heresy. Gary Becker (1976) and Jack Hirshleifer (1978) explore evolutionary theories of altruism in the family. Robert Frank (1988) and Arthur Robson (1992) propose evolutionary explanations for emotions and attitudes toward risk.

Population biologists (William D. Hamilton (1964); Richard Dawkins (1976)) predict altruistic behavior not only between parents and children but also among siblings and other close relatives. Dawkins' expression of this view in *The Selfish Gene* is that the replicating agent in evolution is the *gene* rather than the animal. Since a gene carried by one animal is likely to appear in its relatives, genes for helping one's relatives when the assistance is cheap enough will prosper relative to genes for total selfishness.

Altruistic Sororities without Sex

We introduce the logic of inheritance with a model of *asexual reproduction.* Individuals who survive to reproductive age will have two daughters. Each daughter inherits a genetically-programmed strategy (either *cooperate* or *defect*) from her mother and plays that strategy in a game of prisoner's dilemma with her sister. The larger her payoff in this game, the greater the probability that an individual survives and reproduces.

We claim that the only stable equilibrium[3] is one in which every individual cooperates with her sister. In a population of cooperators, each individual gets payoff R. A mutant who defects against her cooperating sister will get $T > R$. However, her good fortune will not be sustained by her descendants. Her daughters and their descendants will all defect, and hence will each get a payoff of $P < R$. In the long run, the mutant's descendants will reproduce less rapidly than the cooperators and will gradually disappear from the population.

A population of defectors, on the other hand, would be invaded by mutant cooperators. A mutant cooperator would face a defecting sister and get a payoff of S, while normal defectors get $P > S$. However, her daughters and their descendants will cooperate with their siblings and get payoffs of $R > P$. The mutant's descendants therefore reproduce more rapidly than the defectors and will eventually predominate.

The Occasional Altruism of Siblings in Sexually Reproducing Species

Human parents will not be surprised to find that siblings in sexually reproducing species are not always so cooperative. Depending on the payoff parameters in a prisoner's dilemma game, there can be a unique stable equilibrium with cooperators only, or a unique stable equilibrium with defectors only. For some parameter values, there are two stable equilibria - one with cooperators only and one with defectors only - and for some

---

[3] By "stable equilibrium", we mean an equilibrium that is dynamically stable. This should not be confused with the notion of Nash equilibrium in "evolutionary stable strategies" discussed in evolutionary game theory.

parameter values the only stable equilibrium is "polymorphic" with positive proportions of each type.

Consider a large sexually reproducing population in which mating is monogamous and random. For simplicity, assume that each individual either dies without mating or survives to mate and has exactly three offspring. Each offspring plays a game of prisoner's dilemma with each of its two siblings. The probability that an individual survives to reproduce is higher the greater its total payoff in these two games.

An individual's strategy depends on the contents of a single *genetic locus* that contains two genes, one selected at random from each of its parents' two genes. There are two kinds of genes: the *c* (cooperate) gene and the *d* (defect) gene, and three possible types of individuals, *cc* homozygotes who carry two *c* genes, *cd* heterozygotes who carry one *c* gene and one *d* gene, and *dd* homozygotes who carry two *d* genes. Type *cc* homozygotes always play *cooperate* and type *dd* homozygotes always play *defect*. If heterozygotes always defect, then the *d* gene is said to be dominant and the *c* gene to be recessive. If heterozygotes always cooperate, then the *c* gene is said to be dominant and the *d* gene is recessive. In determining the stability of equilibrium we do not assume that either *c* genes or *d* genes are intrinsically dominant, but allow the possibility that mutation could produce a dominant gene of either type *c* or type *d*. (Stability of equilibria against invasion both by recessive and by dominant mutants is not discussed here, but is studied by Bergstrom, 1992.)

Consider a population that consists entirely of cooperating *cc* homozygotes. Suppose that a *c* gene in one individual mutates to a *d* gene that is dominant. The mutant individual is a *cd* heterozygote who plays *defect*. The mutant individual gets a higher payoff than a normal member of the population, since it defects while its siblings cooperate. But whether the mutant genes will proliferate or disappear depends on whether the mutant individual's heterozygote offspring have higher or lower payoffs than normal individuals.

When the mutant *cd* types are rare, they almost certainly mate with *cc*'s. Each offspring of such a union will be a cooperating *cc* with probability 1/2 and a defecting *cd* with probability 1/2. A heterozygote offspring will defect. Each of its siblings with probability 1/2 will be a *cc* who cooperates and with probability 1/2 will be a *cd* who defects. The expected total payoff to the heterozygote defector in the games it plays with its two siblings is therefore $T + P$. Normal individuals with normal siblings cooperate with cooperating siblings and get payoffs of R from each game. Therefore, hetero-

zygote offspring of a mutant defector get higher payoffs and reproduce more rapidly than normal individuals if and only if $T + P > 2R$.

Now consider a population that consists entirely of type *dd*'s in which a *d* gene mutates to a *c* gene that is dominant, so that heterozygotes cooperate. A mutant type *cd* individual will almost certainly mate with a normal type *dd*. Each of its siblings with probability 1/2 will be a *cd* who cooperates and with probability 1/2 will be a *dd* who defects. The expected total payoff to the heterozygote cooperator in the games it plays with its two siblings is therefore $S + R$. Normal individuals with normal siblings defect against defecting siblings and get P from each game. Therefore, heterozygote offspring of a mutant cooperator get higher payoffs and reproduce more rapidly than normal individuals if $S + R > 2P$.

It follows that there is a stable equilibrium consisting entirely of type *cc* cooperators if $T + P < 2R$ and a stable equilibrium consisting entirely of type *dd* defectors if $S + R < 2P$. Prisoner's dilemma games can be found where one, both, or neither of these inequalities are satisfied. The possibilities are illustrated in Figure 1, where the game is normalized by setting $S = 0$ and $T = 1$. (With this normalization, the game is a prisoner's dilemma game if $R > P$ and $R > .5$. The region above the two dotted lines satisfies these conditions.) For parameter values in Region C, there is a *unique* stable equilibrium with a population consisting entirely of type *cc* cooperators. In Region D there is a *unique* stable equilibrium with a population consisting entirely of type *dd* defectors. Elsewhere it has been shown that in Region B there are two stable equilibria, one with a population of cooperators only and one with a population of defectors only, and that for parameter values in Region A, the only stable equilibrium is a polymorphic equilibrium. (See Carl Bergstrom and Bergstrom, 1992).

In an asexual population, cooperative siblings prevail because an individual's sibling is almost certainly "programmed" to treat her in the same way that she is programmed to treat her sibling. In a diploid sexual population, a mutant individual (with a dominant mutant gene) whose genes tell him to treat his siblings in a way different from normal will, with probability 1/2, have siblings who treat him just as he treats them. This probable similarity is sufficient to sustain cooperation in some but not all prisoner's dilemma games.

II. The Evolution of Behavior That is Acquired by Imitation

Similar results obtain when behavior toward one's siblings is acquired by imitation rather than through genetic hard-wiring. We assume that with probability v a child ran-

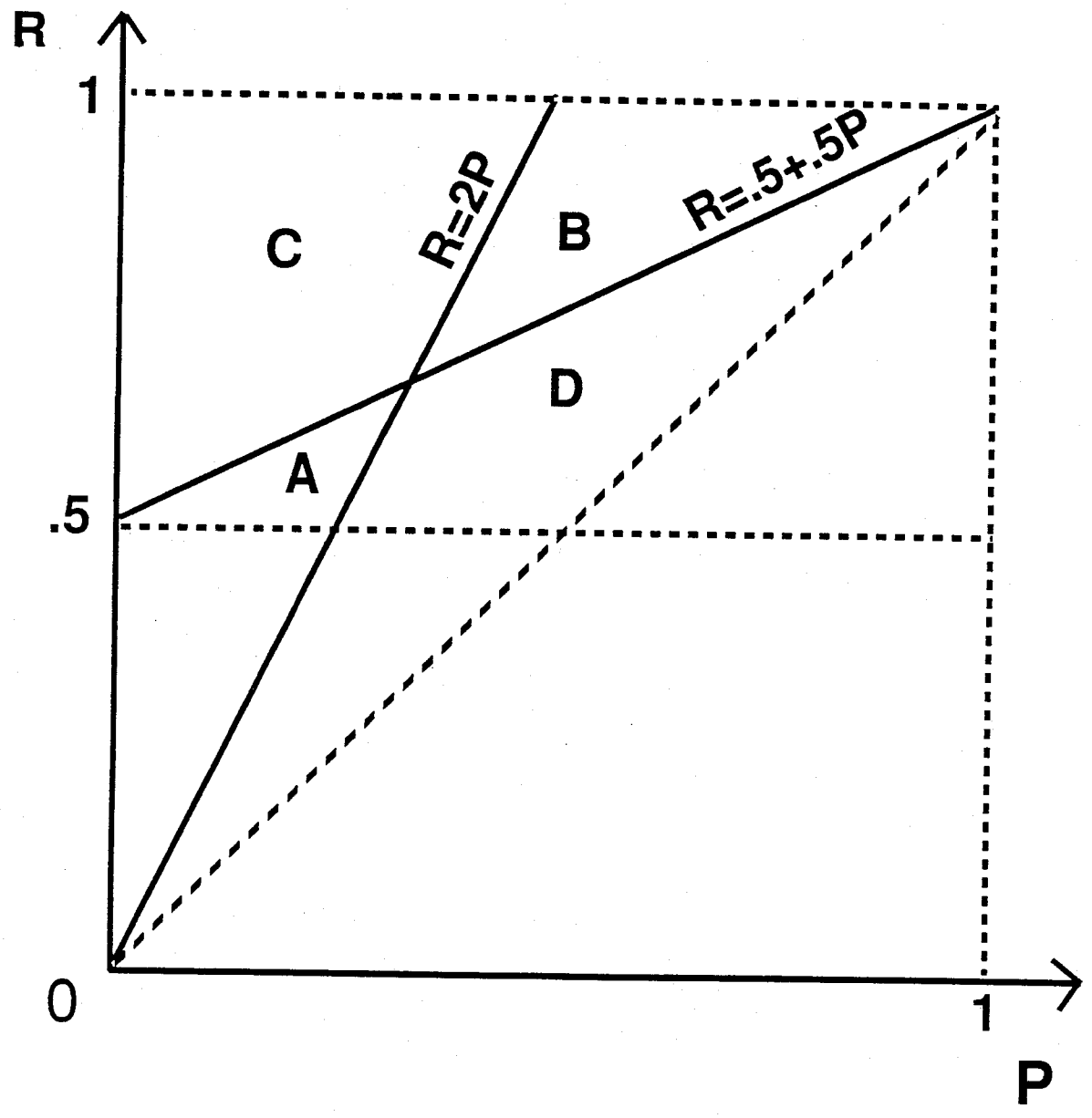

Figure 1. Equilibrium Regimes for Diploid Siblings

domly selects one parent as a role model and adopts that parent's strategy. With probability $1 - v$ the child chooses a random nonparent as a role model.[4] Each individual has two siblings and plays a game of prisoner's dilemma with each of them. The probability that an individual survives to reproduce is proportional to its average payoff in these two games.

Let mating be monogamous. Parent-couples can be one of three possible types: two-cooperator couples, "mixed couples" with one cooperator and one defector, and two-defector couples. Let x be the fraction of the adult population who are cooperators. If marriage is purely random, the fraction of marriages with mixed couples is $2x(1 - x)$. We define a parameter m where $0 \leq m \leq 1$ so as to allow mating patterns that are intermediate between the polar cases of purely random mating ($m = 0$) and purely assortative mating ($m = 1$). In the population at large, the proportion of two-cooperator couples is $x^2 + mx(1 - x)$, the proportion of two-defector couples is $(1 - x)^2 + mx(1 - x)$, and the proportion of mixed couples is $2(1 - m)x(1 - x)$.

Given the rules of imitation and our assumptions about mating, the proportions of cooperators and defectors who survive to reproduce will determine the expected proportions of each type of family, where family types are distinguished by the number of cooperating parents and the number of cooperating children. This in turn determines the expected proportions of cooperators and defectors who survive to reproduce in the next generation.[5]

Conveniently, the rate of change of the ratio of cooperators to defectors turns out to be linear in the proportion x of cooperators in the population. Specifically, this rate of change is proportional to $\alpha x + \beta(1 - x)$, where $\alpha = v^2(1 + m)(T - P) - 2(T - R)$ and $\beta = v^2(1 + m)(R - S) - 2(P - S)$. Depending on parameter values, the dynamics of this system falls into one of the following qualitatively distinct cases:

- $\alpha > 0$ and $\beta > 0$: The only stable equilibrium is a population consisting entirely of cooperators.
- $\alpha < 0$ and $\beta < 0$: The only stable equilibrium is a population consisting entirely of defectors.
- $\alpha < 0$ and $\beta > 0$: The only stable equilibrium is a polymorphic equilibrium in which the proportion of cooperators is $\beta/(\beta - \alpha)$.

---

[4] This is a variant of cultural transmission models developed by Luigi Cavalli-Sforza and Marcus Feldman (1980), and Robert Boyd and Peter Richerson (1985).

[5] Details are provided in an appendix that is available on request.

- $\alpha > 0$ and $\beta < 0$: There are two stable equilibria, one with a population of cooperators only and another with a population of defectors only.

There is a simple heuristic explanation of these results. The proportion of cooperators will increase or decrease depending on whether the average payoff to cooperators is higher or lower than that of defectors. If defectors were as likely as cooperators to have cooperative siblings, then defectors would get higher expected payoffs than cooperators. However, siblings are more likely to be similar than random pairs of individuals. As the proportion of one type in the population approaches zero, the probability that an individual of the rare type has a sibling of the rare type approaches $v^2(1 + m)/2$ [6] and the probability that an individual of the common type has a sibling of the rare type approaches zero.

When cooperators are rare, the expected payoffs of cooperators and defectors approach $Rv^2(1 + m)/2 + S(1 - v^2(1 + m)/2)$ and P, respectively, in each game they play. Since each individual plays with two siblings, the difference between the expected total payoff of rare cooperators and that of normal defectors is $v^2(1 + m)(R - S) - 2(P - S) = \beta$. A similar calculation shows that when defectors are rare the difference between the expected payoff of cooperators and the expected payoff of defectors is α. Therefore when α and β are both positive (negative), a population of cooperators (defectors) could not be invaded by defectors (cooperators), but a population of defectors (cooperators) would be invaded by cooperators (defectors). When $\beta < 0$ and $\alpha > 0$, normal defectors do better than rare cooperators and normal cooperators do better than rare defectors, so that there are two stable equilibria, one with cooperators only and one with defectors only. When $\beta > 0$ and $\alpha < 0$, there are no stable equilibria that have only one type of individual.

When there is perfectly assortative mating ($m = 1$) and children always imitate a parent ($v = 1$) then $v^2(1 + m)/2 = 1$ and $\alpha = \beta = 2(R - P) > 0$, so the only equilibrium has a population consisting entirely of cooperators. (Since in this case each child imitates its two identical parents, the outcome is the same as with asexual reproduction.) If mating is random ($m = 0$) and children always imitate a parent ($v = 1$), then $\alpha = 2R - T - P$ and $\beta = R + S - 2P$. The parameter values yielding the four possible types of equilibria are the same as for diploid inheritance and correspond to the regions

---

[6] An individual of the rare type will be of the same type as its sibling if both copy the same parent or if they copy different parents, but both parents are of the same type. The former event happens with probability $v^2/2$. The latter event happens with probability $mv^2/2$, since the probability that an individual of a rare type is mated to a similar individual approaches m as the proportion of the rare type approaches zero.

C, D, B, and A in Figure 1. If children never imitate a parent ($v = 0$), then $\alpha = 2(R - T) < 0$ and $\beta = 2(S - P) < 0$, and the only equilibrium is a population consisting entirely of defectors.

More generally, the greater is $v^2(1 + m)$, the greater the set of payoff parameters for which there is an equilibrium with all cooperators, and the smaller the set of parameters for which there is an equilibrium with all defectors. Thus the more likely it is that children imitate their parents, and the more likely it is that parents are of the same type, the more likely it is that cooperative behavior will prevail.

III. When does Provincialism Promote Cooperation?

Suppose several farmers live along a country road that loops around a lake. Each farmer plays prisoner's dilemma with his two nearest neighbors, and his income is the sum of the payoffs from these two games. When the farmers' sons take over, they decide whether to cooperate or defect after looking at the actions taken and payoffs received by their fathers and neighbors. Interesting patterns of cooperation emerge if the prisoner's dilemma games have parameters in region C of Figure 1, such that $S + R > 2P$ and $T + P < 2R$ (as for example when $S = 0$, $P = 1/4$, $R = 3/4$, and $T = 1$.)

Consider the case in which each farmer plays prisoner's dilemma with his two immediate neighbors, and each son imitates his father or an adjacent neighbor, depending on who has the highest total payoff. Any arrangement of farmers in which cooperators always appear in clusters of three or more, and defectors always appear in clusters of two or more will be stable.

Curious things happen if we assume that sons pay no attention to their fathers but instead imitate the more prosperous of their two neighbors. In this instance we find cultural patterns that "pick up their feet and walk down the road." One such pattern is a sequence of five adjacent farmers whose behavior forms the pattern CDCCC (where *C* denotes cooperation and *D* denotes defection) while all other farmers on the road defect. If there are at least eight farmers on the road, then each farmer's son will copy the behavior of his neighbor to the left. This moves the pattern CDCCC to the right by one farm in each generation. A long-lived chronicler of behavior at a single farm would see cycles in which spells of defections are interrupted by spells of cooperation.

Cycles are also generated when the farmers play prisoner's dilemma with their two nearest neighbors and the sons imitate the most prosperous of their father and his *four*

nearest neighbors. These cycles form "blinkers" which switch from a single defector surrounded by seven or more cooperators to a cluster of five defectors, then to a cluster of three defectors, and then back to a single defector. Thus the population of defectors alternately shrinks or expands, but never takes over entirely and never becomes extinct.[7]

IV. Maximizers and Imitators

Donald Cox and Oded Stark (1992) suggest that even selfish people have reason to be kind to their aged parents because this behavior may be "imprinted" on their children who, when they reach adulthood, will behave similarly without asking why. If this is so, it is in the interest of adults to treat their parents as they would like to be treated by their own children. But it is odd to assume that parents are *free to choose* according to their self-interest, while their childrens' behavior is predetermined by imprinting. One way out of this impasse is to suppose that parents do not know whether their children will be "imitators" or "maximizers".

For simplicity, let us consider single-parent, single-child families. Maximizers seek to maximize the expected value of $U(x,y)$ where x is what the maximizer does for her mother and y is what the maximizer's daughter does for the maximizer. Suppose that with probability π a daughter will simply imitate her mother's action, and with probability $1 - \pi$ the daughter will choose an action to maximize her expected payoff in the awareness that her own daughter may be an imitator. Therefore, a maximizing mother chooses x so as to maximize

$$\pi U(x,x) + (1 - \pi)U(x,y),$$

where y is the action that a maximizing daughter would take.

If the environment is stationary, the planning problem faced by each generation is the same as that faced by its predecessor, so that the maximizing action for a daughter will be the same as the maximizing action for her mother. Therefore, in equilibrium, everyone chooses $\bar{x}$, *where* $x = \bar{x}$ maximizes $\pi U(x,x) + (1 - \pi)U(x,\bar{x})$. The first-order necessary condition for $\bar{x}$ to be an equilibrium is $U_1(\bar{x},\bar{x}) + \pi U_2(\bar{x},\bar{x}) = 0$. This means that the marginal cost $-U_1(\bar{x},\bar{x})$ of kindness to one's parent equals π times the marginal benefit of kindness from one's child; the likelihood of not being imitated taxes kindness. If $U(x,y)$ is a concave function, the equilibrium choice of x by each generation is an

---

[7] Similar results were found in computer simulations on two-dimensional grids by Martin Nowak and Robert May (1992). Our one-dimensional examples have the advantage of showing these effects in a model simple enough to be studied with pad and pencil.

If $U(x,y)$ is a concave function, the equilibrium choice of x by each generation is an increasing function of π and the utility level enjoyed by each generation will be higher the closer π is to 1. Although maximizers always do at least as well as imitators, in families where imitation is more likely, *everyone* does better.

V. Conclusion

We have studied environments in which an individual gets a higher payoff from defecting than from cooperating and where "copies" of an individual are more likely to appear the higher is her payoff. Even in such unpromising soil, cooperation can persist and flourish. The reason is that both genetic and cultural inheritance are blunt instruments that typically do not operate on individuals in isolation. Those who inherit a genetic tendency to cooperate are more likely than others to enjoy the benefits of cooperative siblings. Similarly with cultural inheritance; altruism can prevail when individuals are likely to interact with others who share the same role model.

References

Becker, Gary S. "Altruism, Egoism, and Fitness: Economics and Sociobiology", *Journal of Economic Literature,* September 1976, **14**, 817-826.

Bergstrom, Carl and Bergstrom, Theodore C. "The Evolution of Altruism among Diploid Siblings who Play Games Like Prisoner's Dilemma", University of Michigan Working Paper, 1992.

Bergstrom, Theodore C. "Love and Spaghetti, The Opportunity Cost of Virtue", *Journal of Economic Perspectives,* Spring 1989, **3**, 165-173.

Bergstrom, Theodore C. "On the Evolution of Altruistic Ethical Rules for Siblings", University of Michigan Working Paper, 1992.

Bernheim, B. Douglas and Stark, Oded "Altruism within the Family Reconsidered: Do Nice Guys Finish Last?", *American Economic Review,* December 1988, **78**, 1034-1045.

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

Cavalli-Sforza, Luigi and Feldman, Marcus *Cultural Transmission and Evolution.* Princeton, N. J.: Princeton University Press, 1981.

Cox, Donald and Stark, Oded "Intergenerational Transfers and the Demonstration Effect", Harvard University Working Paper, 1992.

Dawkins, Richard *The Selfish Gene.* New York: Oxford University Press, 1976.

Frank, Robert *Passions within Reason.* New York: Norton, 1988.

Hamilton, William D. "The Genetical Evolution of Social Behavior. I and II", *Journal of Theoretical Biology*, 1964, **7**, 1-52.

Hirshleifer, Jack "Natural Economy Versus Political Economy", *Journal of Social and Biological Structures*, 1978, **1**, 319-337.

Nowak, Martin and May, Robert "Evolutionary Games and Spatial Chaos", *Nature,* 1992, **359**, 826-829.

Robson, Arthur "The Biological Basis of Expected Utility, Knightian Uncertainty, and the Ellsberg Paradox", University of Western Ontario Working Paper, 1992.

Stark, Oded "Altruism and the Quality of Life", *American Economic Review*, May 1989, *(Papers and Proceedings)*, **79**, 86-90.

Stark, Oded "Nonmarket Transfers and Altruism", *European Economic Review*, 1993 (forthcoming).