Ethics, Evolution, and Games Among Family and Neighbors

Theodore Bergstrom
University of California Santa Barbara[1]

April, 2009

[1]The author is grateful to Carl Bergstrom and Ken Binmore for encouragement and useful advice. Section 3 of this paper is quite incomplete. A better version is planned for the near future.

Abstract

Several similar maxims, known as "Golden Rules" are found in the writings of moral philosophers and religious teachers. Though similar, these maxims appeal to different principles; and do not always recommend the same actions nor lead to the same equilibrium outcome in interactive games. This paper examines some of these rules and explores the way that they may emerge as a result of biological or social evolution.

1 Ethics, Altruism, and Utility

Golden Rules

Several similar maxims, known as "Golden Rules" are found in the writings of moral philosophers and religious teachers. Though similar, these maxims appeal to different principles; and do not always recommend the same actions. There are at least four distinct variants of the golden rule.

The first version is the "Love-thy-neighbor" rule. This rule appears in the Hebrew Old Testament as well as in the Taoist writings of Lao Tze:

- "Thou shalt love thy neighbor as thyself." *Leviticus 19:18*
- "Regard your neighbor's gain as your gain, and your neighbor's loss as your own loss." Lao Tze [14]

A second version, is the "Do-unto-others" rule. This rule is found in the Christian New Testament, in teachings of the Jainist religion of ancient India, and in the writings of Aristotle:

- "Do unto others as you would have them do unto you." *Luke 6:31*
- "We should behave toward friends as we would wish friends to behave toward us." Aristotle [9]
- "A man should wander about treating all creatures as he himself would be treated. "Jainist *Sutrakritanga* 1.11.33

A third version is the "Negative do-unto-others" rule found in the writings of Confucius, in the Hebrew Talmud and in ancient religious texts of Hinduism, Buddhism, and Zoroastroism.

- "Never impose on others what you would not choose for yourself." Confucius *Analects*
- "That which thou likest not being done unto thyself do not unto thy neighbor. That is the whole of Torah and the remainder is but commentary." Hillel the elder, *Babylonian Talmud, Shabbat 31a*
- "This is the sum of duty: do not do to others what would cause pain if done to you." Hindu *Mahabharata 5:1517*
- "Hurt not others in ways that you yourself would find hurtful." Buddhist *Udana-Varga 5:18*

- "Whatever is disagreeable to yourself do not do unto others." Zoroastrian *Shayast-na-Shayast 13:29*

The Love-thy-neighbor form of the golden rule calls for empathy with one's neighbor that would produce altruistic behavior. The Do-unto-others and Negative do-unto-others also require individuals to account for the the effects of their own actions on the payoffs of others when choosing how to act.

The previous three rules do not explicitly recognize the equilibrium effects in game interactions where they are adopted by more than one player. Immanuel Kant's *Categorical Imperative* is a golden rule that accounts for the reciprocal effects that would arise if a maxim were universally adopted. This principle can arguably be found as well in the oldest known instance of a golden rule–a rule that appears in an Egyptian tale, *The Eloquent Peasant* which was written between 2000 and 1750 BC.

- "Act only according to that maxim whereby you can at the same time will that it should become a universal law." Kant, *Foundations of the Metaphysics of Morals* [8]

- "Do for one who may do for you, that you may cause him thus to do." *The Eloquent Peasant*, translated from the original papyri by R.B. Parkinson [13]

Utility, Symmetric Games, and Golden Rules

To understand the differences among these golden rules, it is useful to interpret them as instructions for how individuals should form the utility functions that guide their choices in simple games. We will assume that individuals have well-defined private payoffs that are determined independently of ethical rules. Ethical rules recommend behavior that is guided not only by one's own private payoff, but also by the private payoffs of one's neighbors.

We begin with a class of games that seems particularly well-suited for application of golden rules; the class of symmetric two-player games. Each player chooses an action from a set S of possibilities. A player who takes action x while his neighbor takes action y has private payoff that is determined by the function $u(x, y)$. Since the game is assumed to be symmetric, it must be that if when one takes action x and one's neighbor takes action y, the neighbor receives private payoff $u(y, x)$.

The rule "Love thy neighbor as thyself" commands each neighbor to value the private payoff of his neighbor as highly as his own.[1] If an individual chooses action x and his neighbor chooses y, then the sum of his own private payoff and that of his neighbor is $u(x,y) + u(y,x)$. Therefore this rule requires that when one's neighbor is playing y, one should choose x to maximize

$$L(x,y) = u(x,y) + u(y,x). \tag{1}$$

The Do-unto-others rule asks an individual to act toward his neighbor in the way that he would if the payoffs were reversed. When two neighbors are playing a symmetric game, if a player takes action x and his neighbor takes action y then the neighbor's private payoff will be $u(y,x)$. Therefore the Do-unto-others rule requires that if one's neighbor plays y, one should choose x to maximize

$$D(x,y) = u(y,x). \tag{2}$$

To characterize the Negative do-unto-others rule, we need to decide how to distinguish an instruction not to harm others from an instruction to help others. Of course it is logically possible to make this rule equivalent to the Do-unto-others rule by supposing that failure to help a neighbor is equivalent to harming the neighbor. Some followers of the negative golden rule may interpret this rule in just this way. But here we take what seems a natural interpretation of the language in which "not doing things to our neighbors which we wouldn't want being done to us" differs from "doing for our neighbor whatever we would wish him to do for us." To make this distinction, we define an action x_0 such that one who takes this action neither harms nor benefits his neighbor. The Negative do-unto-others rule requires that if my neighbor is taking action y, my choice of actions should be constrained to those actions such that my neighbor is no worse off than he would be if I took action x_0. This constraint requires that $u(y,x) \geq u(y,x_0)$. Thus the Don't-do-unto-others rule requires that if one's neighbor plays y, one should choose x to maximize

$$\begin{aligned} N(x,y) &= u(x,y) \text{ if } u(y,x) \geq u(y,x_0) \\ &= -\infty \text{ if } u(y,x) < u(y,x_0). \end{aligned} \tag{3}$$

For symmetric games, the Kantian rule asks that one take the action that would maximize one's own well being if this action were to be copied

[1] Following this rule implies that one impose a meaningful cardinality on utility up to affine transformations. In evolutionary applications, the appropriate cardinal measure would be one's expected number of descendants in the long run. In many ethically-motivated discussions, $u(x,y)$ would be one's von Neumann Morgenstern utility.

by one's neighbor. This requires that regardless of the other's action y, individuals should choose the action x that maximizes the utility function:

$$K(x, y) = u(x, x). \tag{4}$$

Some Benchmark Games

The Christmas Gift Game

At Christmas time, people try to buy objects for their friends that the friends have not purchased for themselves; typically because they did not think these objects were worth their cost. Consider two friends, both of whom drink cheap scotch. Each considers spending \$80 to buy a bottle of expensive scotch for the other. Neither buys this brand for himself, since they value it only at \$60. In this game, there are two possible strategies "Give the scotch" G and "Don't give the scotch" D. The private payoffs in this game are $u(D, D) = 0$, $u(D, G) = 60$, $u(G, D) = -80$ and $u(G, G) = -20$.

A follower of the Love-thy-neighbor rule will choose action x to maximize the payoff function $L(x, y) = u(x, y) + u(y, x)$. Since the sum of the payoffs to the two players is greater if one does not give the scotch, D is a dominant strategy for both players. Likewise, followers of the Kantian rule would choose D since $u(D, D) > u(G, G)$. A follower of the Negative do-unto-others rule would choose not to give, under the reasonable assumption that the strategy D does not harm the other. Thus the outcome with players of any of these three types would be that neither gives.

But, for followers of the Do-unto-others rule, G is a dominant strategy. This is true since each person would rather receive a bottle of expensive scotch than not. The only Nash equilibrium in the game played between two Do-unto-others players is that both give the expensive scotch, although this leaves both worse off than they would have been if they had given nothing.

The Redistribution Game

Two neighbors play the following game. A coin is tossed. If it comes up heads, Neighbor A receives \$100 and Neighbor B receives \$20. If it comes up tails, the prizes are reversed. When he sees the outcome, each neighbor can choose to give as much of his prize as he wishes to the other. The private payoffs of the players are expected utilities with concave von Neumann Morgenstern utilities of one's own post-transfer wealth. Each neighbor observes

the result of the coin flip and is allowed to give any non-negative amount of his prize to the other.

It is straightforward to show that those who follow the Love-thy-neighbor rule and those who follow the Kantian rule will choose to give \$40 to the other if their own prize is \$100 and nothing to the other if their own prize is \$20. The equilibrium outcome after gifts is therefore an equalitarian outcome, where each player receives \$50 with certainty.

Since not giving does not harm the other player, a reasonable interpretation of the Negative do-unto-others rule is that neither player would choose to give anything to the other. If both players play by the Do-unto-others rule, each would chooses the action that is most privately beneficial to the other. Therefore each player would give his entire prize to the other. Thus both do-unto-others rules result in a final distribution in which one player gets \$100 and the other gets \$20. With the negative do-unto-others rule, there are not transfers and the winner of the initial lottery has the higher income, while with the positive do-unto-others rule, the loser of the initial lottery emerges with the higher income after transfers.

Love Can Be Better Than Reciprocity

Two neighbors play a game in which their private payoffs constitute a prisoners' dilemma. Each player has access to two possible strategies, C (cooperate) and D (defect). Private payoffs are given by Table 1, where $T > R > P > S$. If the neighbors act selfishly, each will choose the strategy D, although both would be better off if both chose C.

Let us assume that $T + S > 2R$, so that the sum of the neighbors' payoffs is maximized when one of them cooperates and the other defects. In this case, the Love-thy-neighbor rule mandates different behavior from that required by the Do-unto-others and Kantian rules.

Table 1: Prisoners' Dilemma with Selfish Payoffs

		Neighbor 2	
		C	D
Neighbor 1	C	R, R	S, T
	D	T, S	P, P

Suppose that the neighbors behave according to the Do-unto-others rule. Followers of this rule will choose the action that they would wish the other to take. Since $R > S$ and $T > P$, each player would always prefer that the other to play C. Therefore, for Do-unto-others players, C is a dominant strategy. Individuals who follow the Kantian rule would also always choose strategy C, since the payoff to each when both choose C exceeds the payoff when both choose D. It follows that the outcome where both choose C is a dominant strategy equilibrium in any population abiding either by the Kantian rule or the Do-unto-others rule.

Those who follow the Love-thy-neighbor rule act according to a utility function L, where $L(C, C) = R + R = 2R$, $L(D, C) = L(C, D) = T + S$, and $L(D, D) = P + P = 2P$. Since, by assumption, $T + S > 2R$, it follows that for either player, the best response to the other's action is to do the opposite. Therefore there are two Nash equilibria. These are the two outcomes in which the neighbors choose opposite strategies. For this game, we see that the total payoff in equilibrium for two Love-thy-neighbor types exceeds that for the two players who abide by Do-unto-others or Kantian rules.

Reciprocity Can Be Better Than Love

But it can also happen that the Love-thy-neighbor rule results in a lower combined payoff for the two players. Suppose that $T > R > P > S$ and that $T + S < 2P$. As before, Do-unto-others players and Kantian players will choose strategy C, and the only Nash equilibrium for two such players is the outcome where both choose C and receive payoffs R.

For Love-thy-neighbor players, there are two Nash equilibria, one of which is Pareto inferior to the outcome where both cooperate. The outcome in which both choose C will be a Nash equilibrium, since $L(C, C) = R + R > 2P > T + S = L(D, C) = L(C, D)$. But the outcome where both players choose D is a Nash equilibrium, since $L(D, D) = P + P > T + S = L(D, C) = L(C, D)$.

In this example, when $T + S > 2R$, a society that adopts the Love-thy-neighbor ethic will have greater total wealth than one that adopts Do-unto-others or the Kantian ethic. On the other hand, if $T + S < 2P$, a society with the Do-unto-others or the Kantian ethic is likely to outperform one with a Love-thy-neighbor ethic, since the only equilibrium in the former case yields payments of R to everyone, while in the latter case some pairs of neighbors may reach the equilibrium in which they receive $P < R$.

A partial resolution of the two Golden Rules

We have seen that the Love-thy-neighbor and Kantian golden rules do not always recommend the same behavior. Since moralists rarely distinguish between these rules, it is not surprising to find that, at least sometimes, communities governed by the two different maxims arrive at the same social outcome.

In particular, we will show that sometimes, but not always, the symmetric Nash equilibria of the games played by Love-thy-neighbor types and of the games played by Kantian types are the same. We will follow John Maynard Smith [11] in calling a symmetric Nash equilibrium an evolutionarily stable strategy (ESS) for a symmetric game.

Let us assume that the individual payoff function $u(x, y)$ is twice continuously differentiable and defined on a closed, convex subset of the Euclidean plane. Let $u_i(x, y)$ denote the partial derivative of $u(x, y)$ with respect to its ith argument and let $u_{ij}(x_1, x_2)$ denote the ijth element of the Hessian matrix of $u(x, y)$.

For a player who follows the Love-thy-neighbor rule, if the other player takes action y, the partial derivative of $L(x, y)$ with respect to action x is

$$L_1(x, y) = u_1(x, y) + u_2(y, x). \tag{5}$$

For a player who follows the Kantian rule, if the other player takes action y, the partial derivative of $K(x, y) = u(x, x)$ with respect to x is

$$K_1(x, y) = u_1(x, x) + u_2(x, x). \tag{6}$$

At an ESS (interior symmetric Nash equilibrium) both players must choose the same action $\bar{x}$. Therefore an equilibrium $(\bar{x}, \bar{x})$ for Love-thy-neighbor players must satisfy:

$$0 = L_1(\bar{x}, \bar{x}) = u_1(\bar{x}, \bar{x}) + u_2(\bar{x}, \bar{x}), \tag{7}$$

and an equilibrium $(\bar{x}, \bar{x})$ for Kantian players must satisfy:

$$0 = K_1(\bar{x}, \bar{x}) = u_1(\bar{x}, \bar{x}) + u_2(\bar{x}, \bar{x}). \tag{8}$$

Thus the first-order necessary conditions for an ESS are the same for a society of Love-thy-neighbor players as for a society of Kantian players. But this does not necessarily imply that the set of equilibria are the same. The second derivatives $L_{11}(x, y)$ and $K_{11}(x, y)$ are not identical. An action that is a local maximum for a player acting according to the payoff function L

may be a local minimum for a player acting according to K, or vice versa. The second-order condition for $\bar{x}$ to be a best response to the action $\bar{x}$ by one's neighbor is

$$0 \geq L_{11}(\bar{x}, \bar{x}) = u_{11}(\bar{x}, \bar{x}) + u_{22}(\bar{x}, \bar{x}) \tag{9}$$

for Love-thy-neighbor players and

$$0 \geq K_{11}(\bar{x}, \bar{x}) = u_{11}(\bar{x}, \bar{x}) + 2u_{12}(\bar{x}, \bar{x}) + u_{22}(\bar{x}, \bar{x}) \tag{10}$$

for Do-unto-others players.

From equations 9 and 10, we see that

$$K_{11}(\bar{x}, \bar{x}) = L_{11}(\bar{x}, \bar{x}) + 2u_{12}(\bar{x}, \bar{x}) \tag{11}$$

From these facts, we can draw several implications that relate the set of equilibria for Love-thy-neighbor players to those for Kantian players.

Proposition 1 *The evolutionary stable states for games played by Love-thy-neighbor players and those for games played by Kantian players are related as follows:*

1. *If $u_{12}(x, x) < 0$ for all x, then every ESS for a population of Love-thy-neighbor players is an ESS for a population of Kantian players.*
2. *If $u_{12}(x, x) > 0$ for all x, then every ESS for a population of Kantian players is an ESS for a population of Love-thy-neighbor players.*
3. *If $u_{12}(x, x) = 0$ for all x, then the set of ESS outcomes is the same for Kantian players as for Love-thy-neighbor players.*
4. *If the function $u(x_1, x_2)$ is a concave function, then the set of ESS outcomes is the same for Kantian players as for Love-thy-neighbor players.*[2]

2 Natural Selection and Ethics

An interesting evolutionary argument can be made to explain the fact that many successful religions pay at least lip service to the various golden rules.

[2]To prove this result, note that if u is a concave function, its Hessian evaluated at $(\bar{x}, \bar{x})$ must be negative semi-definite. From this it follows that the second-order conditions, 9 and 10 are both satisfied.

Typically these maxims are directed toward one's dealings with "neighbors or friends," people with whom one maintains a continuing relationship. As the Folk Theorem of game theory reminds us, in repeated games between players who are able to observe each others' actions and reward or punish accordingly, almost any feasible outcome can be a Nash equilibrium. The commandments issued by religious leaders, particularly when augmented by threats of divine retaliation,[3] may sometimes serve to coordinate players on equilibrium strategies that lead to group prosperity. Prosperous groups are more likely to spread their influence than those who coordinate on less efficient equilibria. Therefore the doctrines that sustain group prosperity are likely to spread more widely.

In the world that we observe, the unselfishness demanded by the Love-thy-neighbor rule, the Do-unto-others rule, or the Kantian rule seems to be rarely achieved, even by those who profess to believe in them. Our profession's friendless old workhorse, *homo economicus* may exceed real humans in his complete disregard for the well-being of others. But anyone who drives on the freeway or reads the newspaper is likely to concede that much of what we observe is better predicted by the behavior of selfish agents than by that of adherents to Golden Rules.

It is interesting to seek intermediate models that predict behavior that is neither completely selfish nor entirely unselfish. The Negative do-unto-others rule is one such less demanding middle ground. This rule, for example, does not insist that the wealthy share their wealth equally with the poor, but it does demand that they not abuse and further impoverish them.

Evolutionary biology provides a rich source of models of human behavior that lie between the extreme egoism of *homo economicus* and the unselfishness of the golden rules.

Hamilton's rule

The great evolutionary biologist, William Hamilton [6], proposed a rule of altruism for living organisms that is not an entreaty to behave as Hamilton or the Deity would like them to. "Hamilton's rule" is a prediction of the degree of altruism to be found in a population that has been shaped by natural selection. Hamilton's rule generalizes the Love-thy-neighbor rule to allow for individuals who care about others, but less intensely as they care about themselves. Traditional statements of the Golden Rule are not clear on who is to be regarded as a neighbor and hence deserving of symmetric

[3]The full text of Leviticus 19:18 is "Thou shalt not avenge, nor bear any grudge against the children of thy people, but thou shalt love thy neighbour as thyself: I am the LORD."

treatment. Hamilton quite explicitly defines a degree-of-relationship that determines the extent of altruism towards another.

Hamilton stated this rule as follows"

> "The social behavior of a species evolves in such a way that in each distinct behavior-evoking situation the individual will seem to value his neighbors' fitness against his own according to the coefficients of relationship appropriate to that situation." [6], p 19.

Hamilton's rule has a straightforward interpretation in a model where behavior is genetically inherited. In this interpretation, one's "fitness" is defined as the expected number of one's long run biological descendants. The coefficient of relationship between two individuals is determined by their genetic relatedness. Specifically, it is the probability that if one of these individuals has a rare mutant allele, then he shares this allele, by inheritance, with the other. In sexual diploid organisms, with no inbreeding of close relatives, the coefficient of relationship between two full siblings is 1/2, that between half siblings it is 1/4, between (full) cousins it is 1/8, between parent and offspring it is 1/2 and between grandparent and grandchild it is 1/4.

Hamilton stated his rule as a prediction that individuals will act so as to maximize their "inclusive fitness" where inclusive fitness is a weighted sum of one's own fitness and that of one's relatives, with weights being proportional to their degree of relatedness. For a symmetric two-player game, suppose that the fitness of an individual who takes action x when his relative takes action y is $F(x, y)$. Then Hamilton's measure of inclusive fitness for players 1 and 2 is

$$H(x, y) = F(x, y) + rF(y, x) \tag{12}$$

Hamilton's discussion focusses on a special class of games in which he implicitly assumed an additive structure. For a symmetric two-person interaction, this structure implies that the fitness of an organism that takes action x when its relative takes action y is

$$F(x, y) = b(y) - c(x). \tag{13}$$

In this case, the inclusive fitness takes the special form

$$H(x, y) = rb(x) - c(x) + b(y) - rc(y). \tag{14}$$

In a symmetric Nash equilibrium, each player chooses $\bar{x}$ to maximize $rb(x) - c(x)$. The first-order condition for this maximization is $rb'(x) = c'(x)$ which

is to say that a player would help his neighbor so long as the marginal cost of assisting is no greater than the fraction r of the marginal gain to the neighbor from this assistance.

Many interesting social and economic interactions (including the prisoners' dilemma example that we presented above) are not of this additive form. The costs of helping a neighbor may depend on the neighbor's actions, and the benefits from a partner's cooperation may depend on one's own actions.[4] Maynard Smith [10] proposed that the concept of inclusive fitness could be extended to games with general payoff functions. He conjectured that in symmetric games, whether or not the game is additive, equilibrium populations would consist of players who use "evolutionary stable strategies" (ESS), which are symmetric Nash equilbria in a game where payoffs are given by players' inclusive fitness functions.

Several authors, including Alan Grafen [5], Marcus Feldman and Luca Cavalli-Sforza [3], [4], Scott Boorman and Paul Levitt [2], and Theodore Bergstrom [1])have examined genetically based models in which there is natural selection have argued that Maynard Smith's conjecture was not correct. For games that lack Hamilton's additive structure, the set of symmetric Nash equilibria for players with inclusive fitness utility functions does not coincide with the set of equilibria of the most natural evolutionary genetic models of natural selection.

Maynard Smith was persuaded by the critiques of Grafen and others, and agreed that for symmetric non-additive games between relatives, Hamilton's inclusive fitness should be replaced by the following function, which had been introduced by Grafen, and which he and W. G. S. Hines [7] called "personal fitness:"

$$V(x, y) = rF(x, x) + (1 - r)F(x, y). \tag{15}$$

In the Appendix of this paper, we sketch the argument for why the function $V(x, y)$ is appropriate. More detailed discussions can be found in Boorman and Levitt [2] and in Bergstrom [1].

The inclusive fitness function H generalizes the Love-thy-neighbor function L to allow preferences that value a neighbor's fitness as a fraction $r < 1$ of one's own fitness. The personal fitness function V generalizes the Kantian function to allow preferences over actions based in which the probability that one's own act will be mimicked by the neighbor is $r < 1$.

If the effect of actions on fitness has the additive structure implicitly

[4]Hamilton defined inclusive fitness almost 10 years before G. R. Price and John Maynard Smith [12] introduced game theory to evolutionary biologists. It is therefore not surprising that he did not model familial interactions as a game.

assumed by Hamilton, there is no difference between the behavior of an inclusive fitness maximizer and that of a personal fitness maximizer. In this case, Equation 13 personal fitness reduces to:

$$V(x, y) = rb(x) - c(x) + (1 - r)b(y). \tag{16}$$

Thus a personal fitness maximizer chooses x to maximize $rb(x) - c(x)$, which results in the same choice made by the inclusive fitness maximizer.

Proposition 2 *The evolutionary stable states for games played by inclusive fitness players and those for games played by personal fitness players are related as follows:*

1. *If $F_{12}(x, x) < 0$ for all x, then every ESS for inclusive fitness players is an ESS for a population of personal fitness players.*
2. *If $F_{12}(x, x) > 0$ for all x, then every ESS for a population of personal fitness players is an ESS for a population of inclusive fitness players.*
3. *If $F_{12}(x, x) = 0$ for all x, then the set of ESS outcomes is the same for personal fitness players as for inclusive fitness players.*
4. *If the function F is a concave function, then the set of ESS outcomes is the same for personal fitness players as for inclusive fitness players.*

3 Asymmetric Games

Games between relatives of different ages or different sexes often have a strongly asymmetric payoff function. For example, older siblings may be able to bully their younger siblings and deprive them of resources, or they may help their parents with the upbringing of their juniors. In species where siblings are born in different years and never interact directly, the amount of resources that an older child takes from its mother may affect her health and the survival probability of later-born children, while the actions taken by later-born siblings have no effects on their older siblings.

An individual's strategy in an asymmetric game will typically be a function that maps each possible familial role into the action that an individual will take if cast in this role. For example, an individual may be genetically instructed to take one action if finds itself to be the older sibling and a different action if it finds itself to be the younger sibling. This leads to an interesting choice about the appropriate way to model the genetic transmission of strategies.

One possible model assumes that the function that determines ones action, given one's familial role, is controlled by the genes in a single genetic locus. At the opposite extreme is a model in which it is assumed that the action one takes if one is a younger sibling and the action one takes if one is an older sibling are controlled by genes in two distinct genetic loci and that these loci are "unlinked" in the sense that the assortment of genes at these two loci are statistically independent. Intermediate between these two polar models are genetic models of *linkage disequilibrium*, such that "behavior if younger" and "behavior if older" are controlled by two distinct genetic loci, but the contents of these loci are correlated, rather than statistically independent.

Quite remarkably, we find that if behavior in different familial roles is determined by separate, unlinked genetic loci, then the Nash equilibrium for games with inclusive fitness payoffs coincide with stable monomorphic equilibria. However, if the function that maps familial roles into actions is determined by a single genetic locus, then stable monomorphic equilibrium, in general, coincides with Nash equilibrium for a generalization of personal fitness payoffs rather than inclusive fitness payoffs.

Payoff functions and reciprocity

We will consider two-person games where each player is equally likely to be cast in one of two roles. Let $F^i(z_1, z_2)$ be the payoff to the player cast in role i if the action of the role 1 player is z_1 and that of the role 2 player is z_2. A strategy for either player is a vector, specifying an action to be taken if assigned to each role. If player A chooses strategy $x = (x_1, x_2)$ and B chooses strategy $y = (y_1, y_2)$, and if A is assigned role 1 and B role 2, A's fitness will be $F^1(x_1, y_2)$ and B's will be$F^2(x_1, y_2)$. If the roles are reversed, A's fitness will be $F^2(y_1, x_2)$ and B's will be $F^1(y_1, x_2)$. Since the roles are equally likely to be cast in either way, if A's strategy is x and B's is strategy y, then A's expected fitness will be

$$F(x, y) = \frac{1}{2}\left(F^1(x_1, y_2) + F^2(y_1, x_2)\right) \tag{17}$$

and B's will be $F(y, x)$.

Thus we have constructed a symmetric game. For this game, we can define the two symmetric Golden rules, as well as inclusive fitness and personal fitness just as we did for symmetric games. A person whose neighbor pursues strategy y will choose x so as to maximize:

$$L(x, y) = F(x, y) + F(y, x) \tag{18}$$

if he follows the Love-thy-neighbor rule and will choose x to maximize

$$K(x, y) = F(x, x) \tag{19}$$

if he follows the Kan tian rule.

Inclusive fitness is defined as:

$$H(x, y) = F(x, y) + rF(y, x) \tag{20}$$

and personal fitness as:

$$V(x, y) = rF(x, x) + (1 - r)F(x, y). \tag{21}$$

It is useful to note that Equations 17 and 20 imply that $H(x, y)$ can be decomposed as follows:

$$H(x, y) = \frac{1}{2}H^1(x_1, y_2) + \frac{1}{2}H^2(y_1, x_2) \tag{22}$$

where

$$H^1(x_1, y_2) = F^1(x_1, y_2) + rF^2(x_1, y_2) \tag{23}$$

and

$$H^2(y_1, x_2) = F^2(y_1, x_2) + rF^1(y_1, x_2). \tag{24}$$

Equilibrium and Linkage

In the case of symmetric games, we found that stable monomorphic equilibria must be symmetric Nash equilibrium of the game in players act as if their payoffs are given by the personal fitness function V. In the case of asymmetric games, we get two different answers, depending on what we assume about the genetics that determine strategies.

We claim the following result:

Proposition 3 *If a single genetic locus controls one's actions in both roles, then the stable monomorphic equilibria must be symmetric Nash equilibria of the game in which players payoffs are given by the personal fitness function $V(x, y)$. If the genes that control one's strategy when one is cast in different roles are unlinked, so that mutations at one gene are uncorrelated with those at the other, then the stable monomorphic equilibria must be symmetric Nash equilibria of the game with inclusive fitness $H(x, y)$.*

The first assertion is very easily proved. If a single genetic locus controls one's actions in both roles, then the argument that was used to show that stable monomorphic equilibria must be an ESS for personal fitness maximizers applies here as well.

Now suppose that the genes that control an individual's behavior in the two roles are not close together on the same chromosome, so that mutations occurring at one locus are uncorrelated with those at the other. Consider a monomorphic population in which everyone uses the strategy $\bar{x} = (\bar{x}_1, \bar{x}_2)$. Suppose that a mutation occurs in the gene that controls the strategy played in role 1, such that individuals with the mutant gene use strategy $x = (x_1, \bar{x}_2)$. We consider the average fitness of carriers of the mutant gene. With probability 1/2, a carrier of the mutant gene will be called to play role 1 and will have fitness $F(x_1, \bar{x}_2)$. With probability 1/2, the carrier of the mutant gene will play role 2. Given that a mutant individual is called to play role 2, with probability r, the other player, who is called to play role 1 is also a mutant and plays x_1 and a probability $1 - r$ that the other player is a normal and plays $\bar{x}_1$. It follows that the expected fitness of a mutant that plays x_1 is

$$\frac{1}{2}F(x_1, \bar{x}_2) + \frac{1}{2}\left(rF(x_1, \bar{x}_2) + (1-r)F(\bar{x}_1, \bar{x}_2)\right). \tag{25}$$

This expression in turn is equal to

$$\frac{1}{2}\left(H^1(x_1, \bar{x}_2) + (1-r)F(\bar{x}_1, \bar{x}_2)\right) \tag{26}$$

From Equation 26 it follows that a population of $\bar{x}$ strategists can be invaded by a mutant gene for the strategy $(x_1, \bar{x}_2)$ if $H^1(x_1, \bar{x}_2) > H^1(\bar{x}_1, \bar{x}_2)$. Therefore a necessary condition for a population of $\bar{x}$-strategists to be a monomorphic equilibrium is that $\bar{x}_1$ maximizes $H^1(x_1, \bar{x}_2)$ over all possible values of x_2.

Similar reasoning shows that a population of $\bar{x}$ strategists can be invaded by a mutant gene that changes the strategy that is used in role 2 from $\bar{x}_2$ to x_2 if $H^2(\bar{x}_1, x_2) > H^2(\bar{x}_1, \bar{x}_2)$. Therefore for a population of $\bar{x}$-strategists to be a monomorphic equilibrium, it is also necessary that $\bar{x}_2$ maximizes $H^2(\bar{x}_1, x_2)$ over all possible values of x_2.

We note that

$$H(x, \bar{x}) = H^1(x_1, \bar{x}_2) + H^2(\bar{x}_1, x_2) \tag{27}$$

We have seen that $\bar{x}$ is a monomorphic equilibrium only if $\bar{x}_1$ maximizes $H^1(x_1, \bar{x}_2)$ and $\bar{x}_2$ maximizes $H^2(\bar{x}_1, x_2)$. But if these conditions are satisfied, then x maximizes $H(x, \bar{x})$. Therefore a monomorphic equilibrium

strategy $\bar{x}$ must be a symmetric Nash equilibrium for the game with payoff function $H(x, y)$.

References

[1] Theodore C. Bergstrom. On the evolution of altruistic ethical rules for siblings. *American Economic Review*, 85(1):58–81, 1995.

[2] Scott A. Boorman and Paul R. Levitt. *The Genetics of Altruism*. Academic Press, New York, 1980.

[3] L. L. Cavalli-Sforza and M.W. Feldman. Darwinian selection and "altruism". *Theoretical Population Biology*, 14:268–280, 1978.

[4] M. W. Feldman and L. L. Cavalli-Sforza. Further remarks on Darwinian selection and "altruism". *Theoretical Population Biology*, 19:251–260, 1981.

[5] Alan Grafen. The hawk-dove game played between relatives. *Animal Behaviour*, 27(3):905–907, 1979.

[6] W.D. Hamilton. The genetical evolution of social behavior, parts i and ii. *Journal of Theoretical Biology*, 7:1–52, 1964.

[7] W.G.S. Hines and John Maynard Smith. Games between relatives. *Journal of Theoretical Biology*, 79:19–30, 1979.

[8] Immanuel Kant. *Foundations of the Metaphysics of Morals*. Bobbs-Merrill, Indianapolis, 1969.

[9] Diogenes Laertius. *The lives and opinions of eminent philosophers*. G. Bell, London, 1915.

[10] John Maynard Smith. Optimization theory in evolution. *Annual Review of Ecology and Systematics*, 9:31–56, 1978.

[11] John Maynard Smith. *Evolution and the Theory of Games*. Cambridge University Press, Cambridge, U.K., 1982.

[12] John Maynard Smith and G.R. Price. The logic of animal conflict. *Nature*, 246:15–18, 1973.

[13] R.B. Parkinson. *The Tale of the Eloquent Peasant*. The Griffith Institute, Oxford, 1991.

[14] Lao Tze. *Lao Tze's Treatise on the response of the Tao: T'ai-shang Kan-Ying P'ien: a contemporary translation of the most popular Taoist book in China.* AltaMira Press, Walnut Creek, CA, 2003.

Appendix

In evolutionary models, a stable monomorphic equilibrium is a population in which individuals are all of a single genotype and no mutant gene will reproduce more rapidly than those of the equilibrium type. The special feature of genes that regulate interaction with kin is that there is a significant chance that someone who inherits a rare mutant gene regulating this behavior will have a relative who also has this gene. For sexual diploid siblings, for example, if an individual has a rare dominant gene for treating his sibling unusually well, or unusually badly, there is a probability of 1/2 that one's sibling will also share this gene. More generally, in symmetric games between relatives, whose coefficient of relationship is r, the probability that someone with a gene for a mutant behavior towards this relative will find this behavior reciprocated with probability r.

Consider a population in which individuals play a symmetric game with a relative whose coefficient of relatedness is r and where the fitness payoff to playing x when the relative plays y is $F(x, y)$. Let us define

$$V(x, y) = rF(x, x) + (1 - r)F(x, y) \tag{28}$$

If the normal population plays strategy $\bar{x}$ in this game, then a mutant who plays x in a game with a relative of relatedness r will find this strategy reciprocated with probability r from and will encounter the normal strategy $\bar{x}$ with probability $1 - r$. Thus the expected payoff to the mutant will be $V(x, \bar{x})$. A population of individuals playing $\bar{x}$ will be an equilibrium only if no mutant types do better than the normal $\bar{x}$ type. This will be the case only if

$$V(\bar{x}, \bar{x}) \geq V(x, \bar{x}) \tag{29}$$

for all possible strategies x.