How Does Selfishness Affect Well-Being?

Igal Milchtaich*

CORE, Université catholique de Louvain,

B-1348 Louvain-la-Neuve, Belgium

OCTOBER 1999

Is selfishness always a bad thing, in the sense that people can only be better off when everyone is concerned with the well-being of others as well as with his own, or are there situations in which altruism can actually make things worse for all people involved? This paper tackles this question in the context of two-person symmetric games, which are modified by making the payoff of each player a weighted average of that player's true payoff and the true payoff of the other player. The exogenously given degree of selfishness, which determines the weight a player attaches to his own payoff, is the same for both players. It is shown that it is not impossible for the equilibrium payoffs in the modified game to be lower than every equilibrium payoff in the original game. For example, in a symmetric Cournot duopoly competition, if the two firms move only halfway towards monopoly then their profits may be lower than those of both a monopolist and a duopolist. However, this can happen only if the symmetric equilibria in the original game are, in the appropriate sense, unstable. Thus, the effect of selfishness on the symmetric equilibrium payoffs in a symmetric two-person game depends crucially on the stability of these equilibria.

Keywords: altruism, symmetric games, Cournot duopoly, evolutionary stability

JEL Classification: C7, D7

---

* Present address: Department of Economics, Bar-Ilan University, Ramat Gan 52900, Israel.
E-mail: milchti@mail.biu.ac.il.

1. Introduction

It may seem reasonable to argue that if all people in a homogenous society were willing to think less about themselves and more about the well-being of others then everyone in the society were better off. Of course, that would create an incentive problem: If the way other people in the society treat a person is not casually related with the way that person treats them, then why should a person care about others? But if the problem caused by the temptation to free-ride were somehow solved, if for example people were educated to see consideration for others as a moral imperative rather than as a matter of strategic choice, then wouldn't these people be better off than people living in a society in which everyone only cares about his own interests? The somewhat surprising answer is "not necessarily." Even in a symmetric two-person interaction, if both persons increase the weight they attach to the other person's payoff at the expense of their own payoff then the equilibrium payoffs of both players may actually decline. A number of such examples are given in this paper.

This state of affairs raises the following question: Under what circumstances does altruism increase the equilibrium payoffs, and selfishness decreases them, and when is the opposite true? Answering this question is the main goal of this paper. It should be emphasized that the question posed is not what makes people (or animals) behave altruistically in situations in which they can expect no future return and when no sanctions are imposed on those who behave selfishly. A satisfying general answer to this important question is yet to be given. The question asked here is a more modest one. It involves the effect of altruism on material payoffs, assuming that all people care equally about the well-being of others.

This emphasis on the material consequences of moral attitudes, rather than on the rational beyond or the evolutionary origins of these attitudes, is what distinguishes this paper from most other papers that deal with questions of selfishness and altruism, or papers that assume that people's preferences over action profiles need not be consistent with maximizing own payoffs. Among the papers that can be classified as belonging to that literature are Güth and Yaari (1992), Rabin (1993), and Güth (1995).

The work of Bernheim and Stark (1988; see also Stark 1989, 1995) is more closely related to the present one. These authors show that, in the context of two-person interactions, more altruism is not always better. Specifically, they consider an asymmetric two-period model in which each of two individuals, *A* and *B*, is endowed with a fixed amount of some good, corn, which he can consume either in period 1 or in period 2, or partially in both. In addition, individual *A* only can transfer some or all of her remaining corn in period 2 to individual *B*. The utility each individual gets from his own consumption in each period is a strictly concave function of the quantity of corn he consumes in that period, and the total utility derived from own consumption is the sum of the utilities in the two periods. Individual *A* also gets a fraction γ of the utility of individual *B* in both periods. For a certain range of γ values, there is a subgame-perfect equilibrium in which individual *A* consumes more than half of her corn in period 1, and the rest in period 2. She does that in order to prevent individual *B* from consuming a large portion of his own corn in period 1 with the intent of freeloading in period 2. (If *A* had enough remaining corn, she would in that case share it with *B* in period 2.) As γ increases, *A*'s consumption pattern becomes even more inefficient, that is, she consumes a larger part of her corn in period 1. The reason is that, as *A* becomes more altruistic towards *B*, she must take more extreme measures in order to make freeloading unattractive to *B*. The consumption pattern of individual *B* does not change with γ, and remains efficient throughout.

Bernheim and Stark (1988) also announce a positive result that (roughly) says that, for a generic two-person strategic-form game with finite sets of actions, when altruism is sufficiently strong there is a pure-strategy equilibrium which nearly maximizes the payoffs of each of the two players. (These payoffs are not, however, equal to the players' payoffs in the original game but are rather weighted averages of the original payoffs, with coefficients that reflect the intensity of altruism.) The question of when does a moderate degree of altruism have a positive effect on the equilibrium payoffs and when does it have a negative effect on them (as in the above example) is, however, left open by these authors.

The main result of the present paper is that, in symmetric two-person games, there is a strong link between the manner in which selfishness and altruism affect the symmetric equilibrium payoffs and the stability of these equilibria. Specifically, if both persons are increasing the weight they attach to the other person's payoff then the symmetric equilibrium payoffs can decrease only if the original equilibrium is, in some sense, unstable. This general "law" manifests itself in different ways in different classes of games. In a game with a finite number of pure strategies "unstable" means "evolutionarily unstable." In games, such as the Cournot duopoly game, in which strategies are real numbers, an equilibrium is unstable if a small deviation from the equilibrium by one player creates an incentive for the other player to deviate from it even more, and in the same direction.

A noteworthy aspect of these results is that they employ the notions of "stability" and "evolution," which may suggest that the effect of selfishness on the symmetric equilibrium payoffs has something to do with underlying dynamics of the system. The model, however, specifies no such dynamics. The question posed here only involves (comparative) statics: the relation between the set of equilibria in one society and the set of equilibria in another society that differs from the first in the (fixed and universal) degree to which people care about others.

In the special case of symmetric 2×2 games, where each of the two players can only randomize between two actions, whether or not selfishness decreases the symmetric equilibrium payoffs is determined by a single parameter. This parameter is the difference between each player's expected payoff when both players coordinate on the same action, and choose the first or the second action with equal probabilities, and each player's expected payoff when the players independently randomize fifty-fifty between the two actions. (Mathematically, it is proportional to the difference between the sum of the diagonal entries and the sum of the off-diagonal entries of the payoff matrix.) The sign of this difference determines whether the unique completely mixed symmetric equilibrium (if the game has one) is stable. Hence, by the main result, it also determines how altruism affects the symmetric equilibrium payoffs.

The rest of the paper is organized as follows. The next section introduces the setup that is used for formulating and studying the question outlined above. That question is spelled out in more detail in Section 3. In Section 4, an example is given that shows how altruism can affect the profits of two firms that are involved in a symmetric quantity competition in a way that runes counter to one's intuition: When each firm is seeking to maximize a weighted average of the profits (which is, effectively, a step towards monopoly), the profits of both firms are lower than they are when each firm only cares about its own profits. Mutual altruism is thus not necessarily a good thing for these firms. The general analysis of the effect of selfishness—and altruism—on the symmetric equilibrium payoffs in symmetric two-person games starts in Section 5. The first general result asserts that selfishness does not increase the symmetric equilibrium payoffs (and altruism is does desirable) in every game in which coordination has a negative effect on payoffs. This result is then specialized to the case of symmetric Cournot duopoly and, in Section 6, to symmetric $n \times n$ games (in which each player can only randomize among n actions). The connection between the effect of selfishness on the symmetric equilibrium payoffs and the stability of these equilibria is made in Section 7. This connection is first established for general symmetric two-person games and then, with more detail, for $n \times n$ games and games with real strategies. The question of when is the relation between selfishness and the symmetric equilibrium payoffs monotonic, so that equilibrium payoffs are neither locally maximized nor locally minimized by an intermediate degree of selfishness, is studied in the Appendix.

2. The setup

Consider a symmetric two-person game Γ, in which the set of strategies can be either finite or infinite and the payoff of player 1 is an arbitrary function $M(x, y)$ of the strategies x and y played by players 1 and 2, respectively. Symmetry means that the payoff of player 2 is given by $M(y, x)$, and the average payoff $M_0(x, y)$ of the two players is thus equal to $1/2\,[M(x, y) + M(y, x)]$. The perceived payoff of either player is an affine combination of the true payoff of that player and the average payoff of the two players, with both players attaching the same weight $s \geq 0$ to their own payoff.

This parameter s, which is determined exogenously, will be called the degree of selfishness. The perceived payoff $M_s(x, y)$ of player 1 is thus equal to $s\,M(x, y) + (1–s)\,M_0(x, y)$, and the perceived payoff of player 2 is $M_s(y, x)$. The case $s = 1$ represents complete selfishness: the perceived payoffs are equal to the true payoffs. The case $s = 0$ represents complete unselfishness: the perceived payoffs are both equal to the average payoff. The case $s > 1$ represents "super selfishness," that can be interpreted as envy. Indeed, $M_s(x, y) = M(x, y) + 1/2\,(1–s)\,[M(y, x) – M(x, y)]$, so the degree of selfishness also reflects the manner in which the perceived payoff of each player is affected by the difference between the true payoffs of the two players. The perceived payoff of an envious player decreases when the difference between the other player's payoff and the player's own payoff increases, even if the latter payoff does not itself change. The coefficient $1/2\,(1–s)$ may be referred to as the players' degree of altruism. This terminology is justified by the fact that, expressed as an affine combination of the true payoffs of the two players, the perceived payoff of player 1 is $(1+s)\,M(x, y) + 1/2\,(1–s)\,M(y, x)$.[1]

The symmetric two-person game Γ_s in which the payoff function of player 1 is given by M_s will be called the perceived game. Γ itself will be referred to as the true game. Notice that $\Gamma_1 = \Gamma$, and that $(\Gamma_s)_t = \Gamma_{st}$ for every (nonnegative) s and t. Hence, iterating the above procedure would not produce new payoff functions. A useful formula, which is used in a number of places in this paper, is

$$(1) \qquad (s–t)\,M_r + (t–r)\,M_s = (s–r)\,M_t,$$

for every r, s, and t.

[1] Stark (1995, p.16) considers a variant of this setup in which the perceived payoff of each player is an affine combination of that player's true payoff and the perceived payoff of the other player. Specifically, the perceived payoff of player 1 is implicitly defined by the equation $M_s(x, y) = 1/2\,[(1+s)\,M(x, y) + (1–s)\,M_s(y, x)]$. Substituting a similar expression for $M_s(y, x)$ and solving for $M_s(x, y)$, this definition gives $M_s(x, y) = [1/(3–s)]\,[2\,M(x, y) + (1–s)\,M(y, x)]$. It is readily observed that the difference between this expression for the perceived payoff and the one given above merely corresponds to an (increasing) transformation of the degree of selfishness: substituting $(1+s)/(3–s)$ for s in the above expression gives the variant presented here.

3. The questions

When considering the effects of selfishness, a distinction must generally be made between the true, or objective, effects and those that are only due to the fact that each player perceives the other player's payoff as partly contributing to his own. It is, however, not difficult to see that this distinction is unnecessary as long as both players are playing the same strategy: in this case, the expected payoffs in Γ_s and in Γ are the same. Thus, we observe the following:

Basic Observation. For every strategy x, $M_s(x, x) = M(x, x)$. Also, for every two strategies x and y, $M_s(x, x) + M_s(y, y) - M_s(x, y) - M_s(y, x) = M(x, x) + M(y, y) - M(x, y) - M(y, x)$.

In particular, the equilibrium payoffs in every symmetric equilibrium (x, x) in the perceived game Γ_s are the same as they would have been in the true game Γ. Of course, (x, x) is generally not an equilibrium in Γ. It is therefore our basic observation that makes the following question meaningful: How do the symmetric equilibria in the perceived game compare with the symmetric equilibria in the true game? In particular, when does selfishness decrease the symmetric equilibrium payoffs, and when does it increase them?

It is well known that, for example, every symmetric $n \times n$ game (in which the set of strategies is the unit simplex in $\boldsymbol{R}^n$ and the payoff function is bilinear) has a symmetric equilibrium. However, asymmetric equilibria, when they exist, can be as important as the symmetric equilibria or more important. The symmetric 2×2 mis-coordination game in which both players get 0 if they choose the same action and 1 if they don't provides a trivial example. Here, the asymmetric equilibrium payoffs Pareto dominate the symmetric equilibrium payoffs. Why, then, should we restrict attention to symmetric equilibria? The reason for this is threefold. First, when the perceived payoffs of the two players are not equal, they are also different from the respective true payoffs. As noted above, this makes their interpretation problematic. Second, the fact that, in every symmetric equilibrium, the payoffs of the two players are equal guarantees that these equilibria can always be Pareto ranked; thus the sense in which one such equilibrium is

better than the other is never ambiguous. Third, if what we have in mind are randomly matched pairs drawn from some large population of identical individuals, rather than two concrete players, then we are naturally led into considering the notion of a symmetric equilibrium strategy, played by everyone in the population, rather than that of an equilibrium strategy profile, which generally requires some unmodeled asymmetry in the interaction, that determines which (mixed) strategy each of the two players is going to play.

At first sight, it may seem almost tautological that symmetric equilibrium payoffs can only increase when players become less selfish. At least, it may seem that in the extreme case in which both players seek to maximize total, rather than own, payoffs, symmetric equilibrium payoffs must be higher than at the other extreme in which each player is only concerned with his own payoff. Indeed, we have the following easy result:

Proposition 3.0. If Γ is a symmetric $n \times n$ game then $\max_x M(x, x)$ is a symmetric equilibrium payoff in Γ_0.

Proof. If x is a strategy at which the above maximum is attained then, for every strategy y, $0 \geq d/d\varepsilon\,|_{\varepsilon = 0_+}\, M((1-\varepsilon)\, x + \varepsilon\, y,\, (1-\varepsilon)\, x + \varepsilon\, y) = 2\,[M_0(y, x) - M_0(x, x)]$. This proves that x is a symmetric equilibrium strategy in Γ_0. ■

Proposition 3.0 does not, however, tell the whole story. For one thing, it only refers to the case of complete unselfishness. Hence it leaves open the question of how do the symmetric equilibrium payoffs change when there is a less drastic departure from complete selfishness. And then it only tells us something about the highest symmetric equilibrium payoff in Γ_0. It does not tell us anything on how a particular symmetric equilibrium in Γ changes when selfishness is continuously decreased, possibly all the way down to zero.

4. An example: quantity competition

Consider a symmetric Cournot duopoly game, in which firm 1 and firm 2 simultaneously decide how much of the same product they are going to produce. For both firms, the cost of producing quantity q is $C(q)$. Market-clearing price is determined as a function $P(Q)$ of the total output $Q = q_1 + q_2$, where q_i is the output of firm i. The difference $q_1 P(Q) - C(q_1)$ between the revenue and the production cost is the profit $M(q_1, q_2)$ of firm 1. The profit of firm 2 is $M(q_2, q_1)$. For a given degree of selfishness s, the perceived profit $M_s(q_1, q_2)$ of firm 1 is

$$1/2\ \{[(1+s)\, q_1 + (1-s)\, q_2]\ P(Q) - (1+s)\, C(q_1) - (1-s)\, C(q_2)\}$$

and the perceived profit of firm 2 is $M_s(q_2, q_1)$.

Can the symmetric equilibrium profits when the payoffs of the firms are given by M_s be less than the symmetric equilibrium profits when the true game is being played? Surprisingly, the answer to this question is "yes." True, setting $s = 0$ effectively turns the duopoly into a monopoly, and this can only increase the firms' profits. But if selfishness is only partially eliminated then the result can be very different.

Take the price (or inverse demand) function $P(Q) = 1/Q^{2.05} + 1.025$, for example, and suppose that the cost function satisfies $C(q) = q$ for $q \leq 2$ and rapidly increases beyond 2. It can be shown that, in this case, the perceived game has a unique equilibrium for every $0.525 < s < 1.05$ and that this equilibrium is symmetric. The corresponding equilibrium profit is an increasing function of s. This can be seen in Fig. 1, where the perceived profit of a firm as a function of its own output is shown for three different values of s. In all three cases, the output of the other firm in kept fixed at the respective equilibrium output level. Note that the equilibrium output level (which, for each of the three values of s, is the point at which the respective graph peaks) increases as s decreases. Correspondingly, the equilibrium profits decrease.

There is one sense in which this example is different from the (usually more abstract) other examples in this paper. The "perceived" payoff of a firm may actually be realized. For example, if the people who take decisions for one firm also hold shares in another firm, then they can directly benefit from the profits of both firms. In general, however,

the utility that a player gets from the payoff of another player will be interpreted as an imaginary one. The question that we ask is, What happens when each player is behaving as if he is interested in maximizing an affine combination of the two payoffs, rather than his own payoff? Whether or not the perceived payoff represents real gains for the player is (in light of our Basic Observation) immaterial.

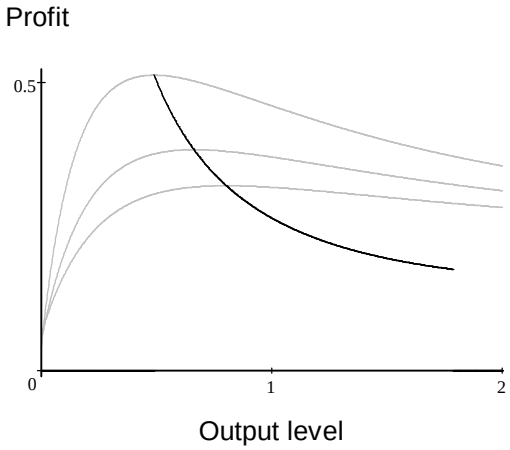

Fig. 1. The perceived profit of a firm as a function of the quantity it produces, in the symmetric Cournot duopoly game described in the text, when the output level of the other firm is fixed at the equilibrium output level. The upper gray line corresponds to the true game ($s = 1$). The middle and lower gray lines correspond to the perceived games with $s = .96$ and $s = .92$, respectively. The black line shows the equilibrium output level and (the true as well as perceived) profit for every $0.525 < s \leq 1$.

5. General analysis

Say that selfishness weakly decreases well-being in a symmetric two-person game Γ if, for every $s \geq 0$ and every $t > s$, every symmetric equilibrium payoff in Γ_s is not lower than any symmetric equilibrium payoff in Γ_t. Say that selfishness decreases well-being in Γ if, for every $s \geq 0$ and every $t \geq s$, every symmetric equilibrium strategy x in Γ_s yields a (strictly) higher equilibrium payoff than every symmetric equilibrium strategy y in Γ_t such that $y \neq x$. If selfishness decreases well-being in Γ then, for every s, Γ_s has at most one symmetric equilibrium. That equilibrium may, however, be the same for every s.

Proposition 5.0. If, for every two strategies x and y, $M(x, x) + M(y, y) - M(x, y) - M(y, x) \leq 0$ then selfishness weakly decreases well-being in Γ. If strict inequality holds whenever $x \neq y$ then selfishness decreases well-being in Γ.

Proof. It is straightforward to check that, for every x, y, s, and t,

$$(2) \qquad (t-s)\,[M(y, y) - M(x, x)] = 2t\,[M_s(y, x) - M_s(x, x)] + 2s\,[M_t(x, y) - M_t(y, y)] + \\ + (s+t)\,[M(x, x) + M(y, y) - M(x, y) - M(y, x)].$$

Suppose that x is a symmetric equilibrium strategy in Γ_s and y is a symmetric equilibrium strategy in Γ_t. Then the first two terms on the right-hand side of (2) are nonpositive. If $M(x, x) + M(y, y) - M(x, y) - M(y, x)$ is also nonpositive then the whole right hand-side is nonpositive, and therefore $M(x, x) \geq M(y, y)$ if $t > s$. Similarly, if that expression in negative and $t > s$ then $M(x, x) > M(y, y)$. The last sentence remains true if the condition $t > s$ is replaced by $t \geq s$. For if $s = t$ then (since x and y are now both symmetric equilibrium strategies in Γ_s) $M_s(y, x) - M_s(x, x) \leq 0 \leq M_s(y, y) - M_s(x, y)$, and therefore $M(x, x) + M(y, y) - M(x, y) - M(y, x) = [M_s(y, y) - M_s(x, y)] - [M_s(y, x) - M_s(x, x)] \geq 0$. ■

The inequality in Proposition 5.0 can be interpreted as expressing a negative effect of coordination on the players' payoffs. Indeed, the expression $M(x, x) + M(y, y) - M(x, y) - M(y, x)$ is equal to four times the difference between the expected payoff of each player when the players coordinate on the same strategy and jointly randomize fifty-fifty between x and y, and the expected payoff of each player when the players independently randomize between the two strategies. If that expression is negative (nonpositive) for every two distinct strategies x and y, then we will say that coordination decreases (respectively, weakly decreases) well-being in Γ. A further justification for this terminology is provided by the fact that if $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$ ($n \geq 2$) is any finite list of distinct strategies, and each of the two players chooses strategy $x^{(i)}$ with positive probability p_i ($i = 1, 2, \ldots, n$), then the difference between the expected payoff of each player when the players always choose the same strategy and the expected payoff when the choices of strategies are independent is equal to $1/2 \sum^{n}_{i,j=1} p_i\, p_i\, [M(x^{(i)}, x^{(i)}) + M(x^{(j)}, x^{(j)}) - M(x^{(i)}, x^{(j)}) - M(x^{(j)}, x^{(i)})]$. This difference is

hence negative (nonpositive) if coordination decreases (respectively, weakly decreases) well-being in Γ. Proposition 5.0 can thus be rewritten more succinctly as follows.

Corollary 5.1. Selfishness decreases (weakly decreases) well-being in every symmetric two-person game in which coordination decreases (respectively, weakly decreases) well-being.

Suppose now that Γ is a game with real strategies, that is, the set of strategies in this game is a subset of the real line. The payoff function $M(x, y)$ in such a game is said to be submodular if $M(x', y') - M(x, y') \leq M(x', y) - M(x, y)$ whenever $x \leq x'$ and $y \leq y'$. The payoff function is strictly submodular if strict inequality holds whenever $x < x'$ and $y < y'$. Submodularity (strict submodularity) of the payoff function is a sufficient condition for coordination to weakly decrease (respectively, decrease) well-being in Γ. If the set of strategies is an interval (either finite or infinite) and the payoff function $M(x, y)$ is twice continuously differentiable, then that function is submodular if and only if its mixed partial derivative is (everywhere) nonpositive. When applied to the above Cournot duopoly example, the submodularity condition has a simple interpretation: it says that the increase (or the negative of the decrease) in firm 1's revenue when that firm increases its output from x to $x' > x$ is never higher when the output level of firm 2 is high than when it's low. The corresponding differential condition is $\partial^2/\partial q_1 \partial q_2 \, [q_1 P(q_1+q_2)] \leq 0$; in words: marginal revenue is a nonincreasing function of the output level of the other firm. A weaker, and simpler, condition, that also implies that coordination (and, hence, selfishness) weakly decreases well-being in a given Cournot duopoly game, is given in the following Corollary to Proposition 5.0. When satisfied by the price function P, this condition guarantees that the equilibrium profits do not behave in the way they do in the example presented in Section 4. It is interesting to note that all of the conditions mentioned here do not involve the cost function C.

Corollary 5.2. Selfishness weakly decreases well-being (and hence symmetric equilibrium profits cannot decrease as the firms move towards monopoly) in every symmetric Cournot duopoly game in which total revenue is a concave function of the

total output. Selfishness decreases well-being in every symmetric Cournot duopoly game in which total revenue is a strictly concave function of the total output.

Proof. This follows immediately from Proposition 5.0 and the fact that, for every x and y, $M(x, x) + M(y, y) - M(x, y) - M(y, x) = 1/2\ [2x\ P(2x) + 2y\ P(2y)] - (x+y)\ P(x+y)$. ■

6. Symmetric $n \times n$ games

In a symmetric $n \times n$ (or symmetric bimatrix) game, each of the two players has a finite number n of possible actions. A (mixed) strategy $x = (x_1, x_2, \ldots, x_n)$ specifies the probability x_i with which a player chooses the ith action, $i = 1, 2, \ldots, n$. The strategy that assigns positive probability to the ith action only will simply be denoted by i. Such a strategy is said to be pure. A strategy that assigns positive probability to each of the n actions is said to be completely mixed. The payoff function M in a symmetric $n \times n$ game Γ is completely specified by the $n \times n$ payoff matrix $M(i, j)$ $(i, j = 1, 2, \ldots, n)$. In the special case of symmetric $n \times n$ games, Proposition 5.0 takes the following form:

Proposition 6.0. Selfishness decreases (weakly decreases) well-being in every symmetric $n \times n$ game in which the quadratic form

$$(3) \qquad F(\xi_1, \xi_2, \ldots, \xi_{n-1}) = \sum_{i,j=1}^{n-1} [\, M(i, j) + M(n, n) - M(i, n) - M(n, j)] \, \xi_i \, \xi_j$$

is negative definite (respectively, negative semidefinite).

Proof. For every two strategies, $x = (x_1, x_2, \ldots, x_n)$ and $y = (y_1, y_2, \ldots, y_n)$, $M(x, x) + M(y, y) - M(x, y) - M(y, x)$ is equal to $\sum_{i,j=1}^{n} M(i, j)\ \xi_i\ \xi_j$, where $\xi_i = x_i - y_i$. Since $\sum_{i=1}^{n} \xi_i = 0$, this sum is equal to the right-hand side of (3). ■

The quadratic form F is particularly simple when $n = 2$. It is, in this case, negative definite or negative semidefinite precisely when $M(1, 1) + M(2, 2) - M(1, 2) - M(2, 1)$ is, respectively, negative or nonpositive. Therefore, the effect of selfishness on the symmetric equilibrium payoff depends, in this case, on whether the sum of the two

diagonal entries of the payoff matrix is greater or smaller than the sum of the two off-diagonal entries.

A paradigmatic example of a symmetric 2×2 game in which selfishness has a negative effect on the symmetric equilibrium payoffs is the Hawk–Dove game, or Game of Chicken ((a) in Fig. 2). The reason why the Hawk–Dove game exhibits this behavior is not difficult to understand. This game represents a dispute over a single, indivisible object. Therefore, the more altruistic the players are, the more they are inclined to play "dove" and bear the risk of losing the object to the other player. Consequentially, the probability of getting the Pareto dominated outcome, at equilibrium, decreases with *s*. Another example of a symmetric 2×2 game in which the symmetric equilibrium payoffs are unique and increase when the degree of selfishness decreases is the Battle-of-the-Sexes game, in its symmetric form ((b) in Fig. 2). In this game, as in the previous one, the sum of the off-diagonal entries of the payoff matrix exceeds the sum of the diagonal entries.

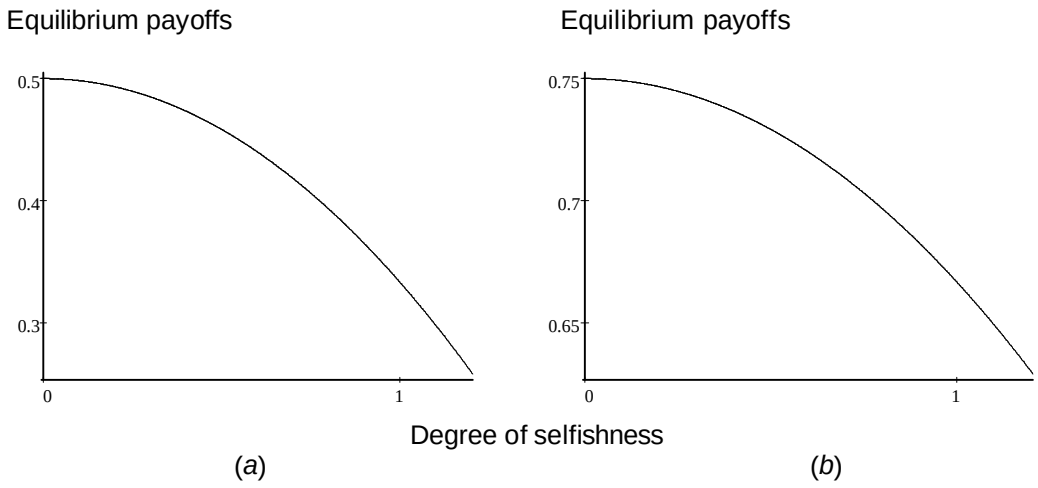

Fig. 2. Symmetric equilibrium payoffs as a function of the degree of selfishness in (*a*) the Hawk–Dove game with payoff matrix $\begin{bmatrix} -1 & 1 \\ 0 & 1/2 \end{bmatrix}$ and (*b*) the Battle-of-the-Sexes game with payoff matrix $\begin{bmatrix} 0 & 1 \\ 2 & 0 \end{bmatrix}$. (Each payoff matrix specifies the payoffs of the row player; the payoffs of the column player are given by the transposed matrix.)

The effect of selfishness on the symmetric equilibrium payoffs in the Prisoner's Dilemma presented in Fig. 3(a) is markedly different. For values of *s* close to 0 or to 1 the symmetric equilibria are pure and yield the socially efficient and inefficient outcomes, respectively. However, for intermediate values of *s* there is also a completely mixed symmetric equilibrium, and the corresponding equilibrium payoffs increase with *s*. Fig. 3(b) shows another Prisoner's Dilemma, in which the behavior of the symmetric equilibrium payoffs is more like those in the Hawk–Dove game. In the first Prisoner's Dilemma, the sum of the diagonal entries exceeds the sum of the off-diagonal entries. In the second Prisoner's Dilemma, it's the other way around.

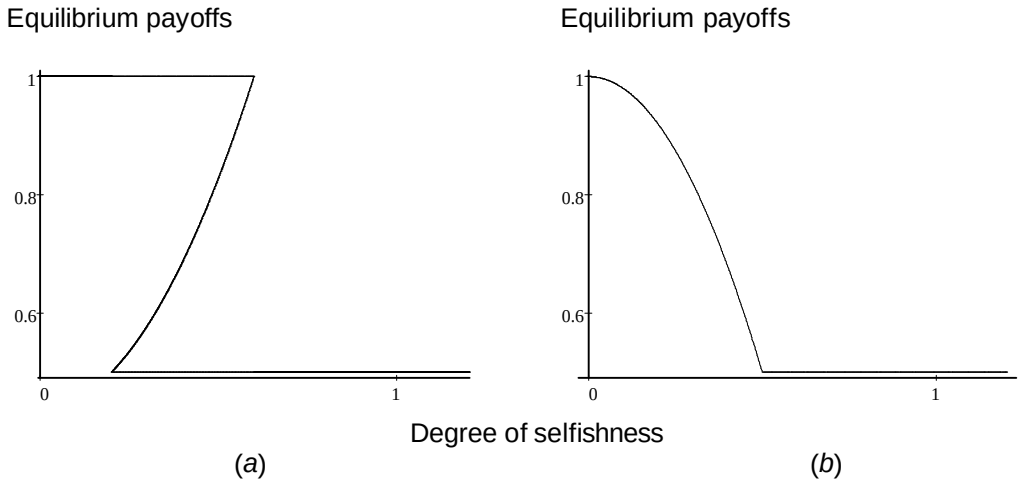

Fig. 3. Two kinds of Prisoner's Dilemma: (*a*) with payoff matrix $\begin{bmatrix} 1/2 & 5/4 \\ 0 & 1 \end{bmatrix}$ and (*b*) with payoff matrix $\begin{bmatrix} 1/2 & 2 \\ 0 & 1 \end{bmatrix}$.

The following proposition shows that a symmetric 2×2 game can always be classified as belonging either to the class of games represented by the games in Fig. 2 or to the class of games represented by the game in Fig. 3(a), or as a border case. It belongs to the first class if the quadratic form *F* is negative definite, to the second class if *F* is positive definite, and is a border case if *F* is identically zero.

Proposition 6.1. Let Γ be a symmetric 2×2 game with payoff matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$, and let Δ denote the difference $(a + d) - (b + c)$.

1) If $\Delta < 0$ then, for every s, the perceived game Γ_s has a unique symmetric equilibrium. The corresponding equilibrium payoffs are determined as a continuous, nonincreasing function of s.

2) If $\Delta > 0$ then the highest symmetric equilibrium payoff in Γ_s is a nonincreasing function of s. This payoff always corresponds to an equilibrium in pure strategies. The set of s values such that Γ_s also has a completely mixed symmetric equilibrium is connected. For each such s there is a unique completely mixed symmetric equilibrium in Γ_s, and the corresponding equilibrium payoffs are determined as a continuous, nondecreasing function of s.

3) If $\Delta = 0$ then either $\max(a, d)$ is the unique symmetric equilibrium payoff in Γ_s, for every s, or there is some $s_0 > 0$ such that $\max(a, d)$ is the unique symmetric equilibrium payoff when $s < s_0$, $\min(a, d)$ is the unique symmetric equilibrium payoff when $s > s_0$, and for $s = s_0$ the set of symmetric equilibrium payoffs in Γ_s is the interval $[\min(a, d), \max(a, d)]$.

Proof. The proof is, for the most part, straightforward. If Γ_s has a completely mixed symmetric equilibrium, and if $\Delta \neq 0$, then direct computation shows that the equilibrium payoff is equal to $s^2(b-c)^2/(4\Delta)$ plus some constant that does not depend on s. ■

By Proposition 6.1, the highest symmetric equilibrium payoff in every symmetric 2×2 game is a nonincreasing function of s. This, however, is not true for general symmetric $n \times n$ games. Consider, for example, the symmetric 3×3 game with payoff matrix

$$\begin{bmatrix} 0 & 1/2 & -1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \end{bmatrix}.$$

It can be shown that, in this example, the perceived game has a unique equilibrium for every $s > 1/3$, and that this equilibrium is symmetric and completely mixed. The corresponding equilibrium payoff is equal to $8s^2/(1-121s^2)$, and is thus an increasing function of the degree of selfishness.

7. Stability

The conditions in Proposition 5.0 are global ones, in the sense that they do not refer to any specific equilibrium in the true or in the perceived game. Correspondingly, they allow us to compare every symmetric equilibrium payoff in Γ_s with every symmetric equilibrium payoff in Γ_t, for every s and t. We now turn our attention to local comparisons: between a particular symmetric equilibrium in Γ_s and those symmetric equilibria in Γ_t that are close to it in some given topology on the set of strategies. As we will see, whether or not such a symmetric equilibrium in Γ_t yields higher payoffs than the given symmetric equilibrium in Γ_s depends critically on whether or not the latter equilibrium is stable.

For given symmetric two-person game Γ and degree of selfishness s, say that a symmetric equilibrium strategy x in Γ_s is <u>weakly stable</u> if $M_s(x, y) \geq M_s(y, y)$ for every strategy y in some neighborhood of x. Say that x is <u>stable</u> if $M_s(x, y) > M_s(y, y)$ for every $y \neq x$ in such a neighborhood. (Note that, in the latter case, x has a neighborhood where it is the only symmetric equilibrium strategy in Γ_s.) Say that selfishness <u>locally weakly decreases well-being</u> at x if there is a neighborhood of x such that, for every strategy y in` that neighborhood and for every $t > s$ such that y is a symmetric equilibrium strategy in Γ_t, $M(x, x) \geq M(y, y)$. Say that selfishness <u>locally decreases well-being</u> at x if there is a neighborhood of x such that, for every strategy $y \neq x$ in that neighborhood and for every $t \geq s$ such that y is a symmetric equilibrium strategy in Γ_t, $M(x, x) > M(y, y)$. Note that, in these definitions, "locally" only refers to strategies; the degree of selfishness t may or may not be close to s.

If coordination decreases (weakly decreases) well-being in Γ then, for every degree of selfishness s, every symmetric equilibrium strategy x in Γ_s is stable (respectively,

weakly stable). This follows from the fact that, for every strategy y, $M_s(x, y) = M_s(y, y) + [M_s(x, x) - M_s(y, x)] - [M_s(x, x) + M_s(y, y) - M_s(x, y) - M_s(y, x)] \geq M_s(y, y) - [M(x, x) + M(y, y) - M(x, y) - M(y, x)]$. Proposition 5.0 and Corollary 5.1, which show a connection between the effect of coordination and the effect of selfishness on well-being, are therefore immediate corollaries of the following result, which establishes a connection between stability and the local effect of selfishness on well-being.

Theorem 7.0. Selfishness locally decreases (locally weakly decreases) well-being at every symmetric equilibrium strategy that is stable (respectively, weakly stable).

Proof. Let Γ be a symmetric two-person game, s and $t \geq s$ degrees of selfishness, x a symmetric equilibrium strategy in Γ_s, and y a symmetric equilibrium strategy in Γ_t. It follows from Eq. (2) that $(t-s)[M(y, y) - M(x, x)] = 2s[M_t(x, y) - M_t(y, y)] + (t-s)[M_s(y, x) - M_s(x, x)] + (s+t)[M_s(y, y) - M_s(x, y)]$. If x is weakly stable and y is close enough to it then the right-hand side of this equation is nonpositive. If, in addition, x is stable and $y \neq x$ then the right-hand side is negative (because $s + t > 0$. This follows from the fact that, in this case, y cannot be a symmetric equilibrium strategy in Γ_s, and therefore $s \neq t$.) ■

One sense in which a stable (weakly stable) symmetric equilibrium strategy is "stable" is the following. If both players are playing that strategy, then neither of them has an incentive to deviate. And if both players are playing some other strategy, which is close to the symmetric equilibrium strategy in question, then each of them has an incentive to deviate (respectively, does not have an incentive not to deviate) and play the original symmetric equilibrium strategy instead. When there is more structure to the set of strategies than just a topology on it, then it may be possible to say more about the interpretation of stability and about the relation between stability and the local effect of selfishness on well-being. Two cases will now be considered.

7.1. Symmetric n×n games

We will say that a strategy y can weakly invade a given strategy x in a symmetric $n \times n$ game Γ if $M(y, x) = M(x, x)$ and $M(y, y) \geq M(x, y)$. If the last condition holds with strict inequality then we will say that y can invade x in Γ. A symmetric equilibrium strategy x is called an evolutionarily stable strategy (ESS) if no strategy $y \neq x$ can weakly invade it, and a neutrally stable strategy (NSS) if no strategy y can invade it. In the case under consideration these notions are, in fact, equivalent to the notions of stability and weak stability, respectively, as defined above: In every symmetric $n \times n$ game, a symmetric equilibrium strategy is evolutionarily stable if and only if it is stable, and neutrally stable if and only if it is weakly stable (see van Damme, 1987; Bomze and Weibull, 1995). However, this characterization of stable symmetric equilibrium strategies in terms of uninvadability also hints to an extension of Theorem 7.0 that holds for every symmetric equilibrium strategy: A nearby strategy, which is a symmetric equilibrium strategy in the perceived game corresponding to a higher (lower) degree of selfishness, yields a higher (respectively, lower) equilibrium payoff than the given strategy x if and only if it can invade x.

Theorem 7.1. Let Γ be a symmetric $n \times n$ game, s a degree of selfishness, and x a symmetric equilibrium strategy in Γ_s. Then there is a neighborhood of x such that, for every y in that neighborhood and for every t such that y is a symmetric equilibrium strategy in Γ_t, the product

(4) $\quad (t - s)\,[M(y, y) - M(x, x)]$

is positive if and only if y can invade x in Γ_s. In addition, if y can weakly invade x in Γ_s then (4) is nonnegative, and if y cannot even weakly invade x in Γ_s and $t > 0$ then (4) is negative.

Proof. The set of all strategies whose carrier[2] contains the carrier of x constitutes an open neighborhood of x. Let y be in that neighborhood, and let t be a degree of selfishness such that y is a symmetric equilibrium strategy in Γ_t. For every strategy z

[2] The carrier of a (mixed) strategy y is defined as the set of all pure strategies i such that $y_i > 0$.

whose carrier is contained in the carrier of y, $M_t(z, y) = M_t(y, y)$. In particular, $M_t(x, y) = M_t(y, y)$. It follows that, for every strategy z as above,

(5) $\quad M(x, x) - M_t(z, x) = [M_t(x, x) - M_t(x, y)] - [M_t(z, x) - M_t(z, y)] \leq (1+t)\ \varepsilon,$

where $\varepsilon = \max_i [|M(i, x) - M(i, y)| + |M(x, i) - M(y, i)|]$.

Suppose, first, that $M_s(y, x) = M_s(x, x)$. It then follows from Eq. (2) that $(t\ \ s)\ [M(y, y) - M(x, x)] = (s+t)\ [M_s(y, y) - M_s(x, y)]$. The right-hand side of this equation has the same sign (positive, negative, or zero) as $M_s(y, y) - M_s(x, y)$. For if $M_s(y, y) \neq M_s(x, y)$ then (since $M_t(y, y) = M_t(x, y)$) $s \neq t$, and hence $s + t > 0$. By definition, $M_s(y, y) - M_s(x, y) \geq 0$ if and only if y can weakly invade x in Γ_s, and strict inequality holds if and only if y can invade x in Γ_s. This concludes the proof in the case in which $M_s(y, x) = M_s(x, x)$.

In the rest of this proof we will assume that $M_s(y, x) < M_s(x, x)$. As this assumption implies that y cannot weakly invade x, it suffices to show that the product (4) is nonpositive, and is moreover negative if $t > 0$, provided that ε is close enough to zero.

The inequality $M_s(y, x) < M_s(x, x)$ implies that there is some pure strategy i_0 in the carrier of y, and some $\varepsilon_0 > 0$ that only depends on s and on x, such that $M_s(i_0, x) < M_s(x, x) - \varepsilon_0$. Since $M_s(x, x) - M_s(i_0, x) = [M(x, x) - M_t(i_0, x)] + [M_t(i_0, x) - M_s(i_0, x)]$, it follows from (5) that

(6) $\quad \varepsilon_0 < (1+t)\ \varepsilon + |t-s|\ m,$

where $m = \max_{i, j} |M(i, j)|$. Provided that $\varepsilon < \varepsilon_0/(1+s)$, it follows from (6) that $t \neq s$.

Suppose that $t < s$. If $M_0(y, x) < M(x, x)$ then there is some pure strategy i_1 in the carrier of y, and some $\varepsilon_1 > 0$ that only depends on x, such that $M_0(i_1, x) < M(x, x) - \varepsilon_1$. It follows, in this case, from (1) that $s\ M_t(i_1, x) = (s-t)\ M_0(i_1, x) + t\ M_s(i_1, x) < s\ M(x, x) - (s-t)\varepsilon_1$, and therefore $(s-t)\varepsilon_1/s < M(x, x) - M_t(i_1, x) \leq (1+t)\ \varepsilon$ by (5). Then, by (6) and the assumption $t < s$, $\varepsilon_0 < (1 + m\ s/\varepsilon_1)\ (1+s)\ \varepsilon$. This inequality does not, however, hold if ε is close enough to zero. We may therefore assume that $M_0(y, x) \geq M(x, x)$. Then, $t\ [M_s(y, x) - M_s(x, x)] + s\ [M_t(x, y) - M_t(y, y)] + (s+t)\ [M(x, x) - M_0(y, x)] \leq 0$, with strict inequality if $t > 0$. It follows from Eq. (2) that the left-hand

side of this inequality is equal to $(-s)\,[M(y, y) - M(x, x)]$. Since this expression has the same sign as (4), that product is also nonpositive, and is negative if $t > 0$.

Suppose now that $t > s$. If $M(y, x) = M(x, y)$ then $M_s(y, x) = M_t(x, y)$, and hence $M(y, y) = M_t(x, y) = M_s(y, x) < M(x, x)$. Therefore, in this case, (4) is negative. If $M(y, x) \neq M(x, y)$ then there are two pure strategies, i_2 and j_2, in the carrier of y, and some $\varepsilon_2 > 0$ that only depends on x, such that $M(i_2, x) - M(x, i_2) > M(j_2, x) - M(x, j_2) + 2\varepsilon_2$, and hence $M(i_2, y) - M(y, i_2) > M(j_2, y) - M(y, j_2) + 2\varepsilon_2 - 2\varepsilon$. On the other hand, $(1+t)\,[M(i_2, y) - M(j_2, y)] + (1-t)\,[M(y, i_2) - M(y, j_2)] = 2\,M_t(i_2, y) - 2\,M_t(j_2, y) = 0$. Therefore, $(1+t)(\varepsilon_2 - \varepsilon) + [M(y, i_2) - M(y, j_2)] < 0$. Provided that $\varepsilon < \varepsilon_2/2$, we get that $1 + t < r$, where $r = 4m/\varepsilon_2$.

If $M_r(y, x) < M(x, x)$ then there is some pure strategy i_3 in the carrier of y, and some $\varepsilon_3 > 0$ that only depends on x, such that $M_r(i_3, x) < M(x, x) - \varepsilon_3$. It follows, in this case, from (1) that $(r-s)\,M_t(i_3, x) = (t-s)\,M_r(i_3, x) + (r-t)\,M_s(i_3, x) < (r-s)\,M(x, x) - (t-s)\varepsilon_3$, and therefore $(t-s)\varepsilon_3/(r-s) < M(x, x) - M_t(i_3, x) \leq (1+t)\,\varepsilon$ by (5). Then, by (6) and $1 + t < r$, $\varepsilon_0 < (1 + m\,(r-s)/\varepsilon_3)\,r\varepsilon$. This inequality does not, however, hold if ε is close enough to zero. We may therefore assume that $M_r(y, x) \geq M(x, x)$. Hence, $(t+r)\,[M_s(y, x) - M_s(x, x)] + (t-r)\,[M_t(x, y) - M_t(y, y)] + (s+t)\,[M(x, x) - M_r(y, x)] < 0$. It is not difficult to check that the left-hand side of this inequality is equal to $(r-s)\,[M(y, y) - M(x, x)]$. Since this expression has the same sign as (4), that product is also negative. ■

7.2. Games with real strategies

Since selfishness does not locally decrease well-being in the example of symmetric Cournot duopoly game given in Section 4, it follows from Theorem 7.0 that the symmetric equilibrium strategies in that example are not stable. The following discussion will shed more light on the sense in which the equilibria in that example are unstable.

Let Γ be a symmetric two-person game with real strategies (in which the set of strategies is a subset of the real line) and let (x, x) be a symmetric equilibrium in Γ. If x

belongs to the interior of the set of strategies and M is twice continuously differentiable in a neighborhood of (x, x) then, since the expression $M(y, x)$ is maximized by setting $y = x$, $\partial M/\partial x = 0$ and $\partial^2 M/\partial x^2 \leq 0$ at the point (x, x). If, in addition, the symmetric equilibrium strategy x is weakly stable then the expression $M(y, y) - M(x, y)$ is locally maximized by setting $y = x$, and therefore $\partial^2 M/\partial x^2 + 2\,\partial^2 M/\partial x \partial y \leq 0$ at (x, x). It follows, in this case, that, at the equilibrium point (x, x),

(7) $\partial^2 M/\partial x^2 + \partial^2 M/\partial x \partial y$

is nonpositive. Conversely, suppose that (7) is positive at (x, x). Then that symmetric equilibrium is "unstable" in a very intuitive sense. Specifically, suppose that $\partial^2 M/\partial x^2$ is different from zero (and is hence negative) at the equilibrium point (x, x). It then follows from the implicit function theorem that there is a continuously differentiable function f from some neighborhood of x to the set of strategies such that $M(f(y), y) \geq M(z, y)$ for every y and z in that neighborhood, with strict inequality for $z \neq f(y)$. Thus, the strategy $f(y)$, if it belongs to the neighborhood, is the unique local best response of player 1 to the strategy y played by player 2. Obviously, $f(x) = x$. The derivative of f at the point x is equal to $-[\partial^2 M/\partial x \partial y(x, x)]/[\partial^2 M/\partial x^2(x, x)]$. Therefore, if (7) is positive at (x, x) then $df/dy > 1$ at x. Hence, if player 2 makes a small deviation from the equilibrium then the (local, at least) best response of player 1 is to make an even greater deviation in the same direction. If players are myopic, and react to each other's strategy by moving towards their best response to it, then it is to be expected in this case that any small, accidental deviation from the equilibrium would lead to the players drifting away from it in an ever accelerating pace. For concrete examples of such "myopic" dynamics see, for example, Okuguchi (1976).

In the above Cournot duopoly example, $\partial^2 M_s/\partial x^2 + \partial^2 M_s/\partial x \partial y > 0$ at the unique symmetric equilibrium of the true or the perceived game for every $0.525 < s < 1.05$. Therefore, for each of these degrees of selfishness, the reaction curve of firm 1 (which is the locus of all strategy profiles (x, y) such that producing quantity x maximizes the perceived profit of firm 1 when the output level of firm 2 is y) is upward-sloping near the equilibrium point, where it intersects the forty five-degree line, and passes below that line to the right of the intersection point. If one firm deviates from the equilibrium

by either increasing or decreasing its output level by a small amount, then the best response for the other firm is to increase or decrease, respectively, its output level even more. Theorem 7.0 suggests that it is this instability which is responsible for the fact that, in this example, equilibrium profits drop initially, rather than increase, as the firms move towards monopoly. The next theorem proves that this is indeed the case. This theorem links the local effect of selfishness on well-being in games with real strategies directly with the sign of (7).

A symmetric equilibrium strategy x in a symmetric two-person game Γ with real strategies will be said to be analytic if it belongs to the interior of the set of strategies, the payoff function M is twice continuously differentiable in a neighborhood of (x, x), and there are constants $a_1, b_1, a_2, b_2, \ldots$ such that, for every y in some neighborhood of x, $M(y, y) - M(x, x) = a_1 (y-x) + a_2 (y-x)^2 + \ldots$ and $\partial M/\partial x(y, y) = b_1 (y-x) + b_2 (y-x)^2 + \ldots$. If $a_1 = a_2 = \ldots = 0$ then $d/dy\,[M(y, y)] = 0$ in some neighborhood of x. Conversely, if $a_n \neq 0$ for at least one n then there is some neighborhood of x in which $d/dy\,[M(y, y)] \neq 0$ for every $y \neq x$.

Theorem 7.2. Let Γ be a symmetric two-person game with real strategies, s a degree of selfishness, and x an analytic symmetric equilibrium strategy in Γ_s. Suppose that in every neighborhood of x there is at least one strategy y such that $M(y, y) \neq M(x, x)$. Then there is a neighborhood of x such that, for every $y \neq x$ in that neighborhood and for every degree of selfishness t such that y is a symmetric equilibrium strategy in Γ_t,

$$(t-s)\,[M(y, y) - M(x, x)]$$

has the same sign (positive, negative, or zero) as

$$\partial^2 M_s/\partial x^2 + \partial^2 M_s/\partial x \partial y,$$

where the partial derivatives are computed at the point (y, y).

Proof. It follows from the assumptions, and the identity $d/dy\,[M(y, y)] = 2\,\partial M_0/\partial x(y, y)$, that there is a convex neighborhood of x in which $\partial M_0/\partial x(y, y) \neq 0$ for every $y \neq x$. Let y be a strategy in that neighborhood. By the mean value theorem, there is some $0 < \theta < 1$ such that $M(y, y) - M(x, x) = 2\,(y-x)\,\partial M_0/\partial x(\xi, \xi)$, where

$\xi = (1-\theta)\, x + \theta\, y$. Therefore, the difference $M(y, y) - M(x, x)$ has the same sign as $(y-x)\, \partial M_0/\partial x(y, y)$. If $y \neq x$, and is a symmetric equilibrium strategy in Γ_t, for some degree of selfishness t (which must be different from zero, for otherwise $\partial M_0/\partial x(y, y) = 0$), then it follows from (1) that $(s-t)\, \partial M_0/\partial x + t\, \partial M_s/\partial x = s\, \partial M_t/\partial x = 0$ at the point (y, y). Then, $(t-s)\, [M(y, y) - M(x, x)]$ has the same sign as $(y-x)\, \partial M_s/\partial x(y, y)$, and hence (by the mean value theorem and the fact that $\partial M_s/\partial x(x, x) = 0$) the same sign as $\partial^2 M_s/\partial x^2(\xi', \xi') + \partial^2 M_s/\partial x \partial y(\xi', \xi')$, where $\xi' = (1-\theta')\, x + \theta'\, y$ for some $0 < \theta' < 1$. Now, since x is analytic, there is a convex neighborhood of x such that either $d/dz\, [\partial M_s/\partial x(z, z)] = 0$ for every z in that neighborhood, or inequality holds for every $z \neq x$ in it. In both cases, $\partial^2 M_s/\partial x^2 + \partial^2 M_s/\partial x \partial y$ has the same sign at (y, y) and at (ξ', ξ'), provided that y belongs to the neighborhood. ■

Appendix: Monotonicity

The equilibrium payoffs corresponding to a completely mixed symmetric equilibrium in a symmetric 2×2 game can either increase or decrease when the degree of selfishness decreases. However, by Proposition 6.1, in any given game the trend cannot reverse itself as s decreases (or increases). This is also true in symmetric 3×3 games. But in symmetric 4×4 games it is possible, for example, for the payoffs corresponding to a unique completely mixed symmetric equilibrium to decrease initially as s decreases from 1 and then to increase again as s approaches 0. This raises the question, What conditions on a symmetric $n \times n$ game Γ guarantee monotonicity of the completely mixed symmetric equilibrium payoffs?

Let s be a degree of selfishness and x a completely mixed symmetric equilibrium strategy in Γ_s. If the payoff matrix $M_s(i, j)$ is nonsingular then $M_s(x, x) \neq 0$, and for every t close enough to s the unique solution $y = (y_1, y_2, \ldots, y_n)$ of the system

$$(8) \qquad \sum_{i=1}^{n} M_s(i, j)\, y_i = M_t(x, j) \qquad (j = 1, 2, \ldots, n)$$

is a completely mixed strategy (because $\sum^n_{i=1} y_i = [\sum^n_{i=1} M_s(i, x)\, y_i]/M_s(x, x) = M_t(x, x)/M_s(x, x) = 1$). The first part of the following theorem shows that a sufficient condition for this strategy to be a symmetric equilibrium strategy in Γ_t is

$$M(i, j) - M(i, 1) - M(1, j) = M(j, i) - M(j, 1) - M(1, i)$$

for every $1 \le i, j \le n$. This condition is equivalent to the matrix given on the right-hand side of (3) being symmetric. It will hence be referred to as the symmetry condition.

Proposition A.0. Let Γ be a symmetric $n \times n$ game satisfying the symmetry condition, s and t degrees of selfishness, and x a completely mixed symmetric equilibrium strategy in Γ_s. Then the following propositions hold:

1) A strategy y satisfies (8) if and only if it is a symmetric equilibrium strategy in Γ_t and every strategy is a best response to it. In particular, every completely mixed symmetric equilibrium strategy y in Γ_t satisfies (8).

2) If y is a strategy that satisfies (8) then, for every $r \ne t$ between s and t, $[(t-r)/(t-s)]\, x + [(r-s)/(t-s)]\, y$ is a completely mixed symmetric equilibrium strategy in Γ_r. The corresponding equilibrium payoff is equal to $M(x, x) + [(r^2-s^2)/(t^2-s^2)]\, [M(y, y) - M(x, x)]$, and is thus a quadratic, monotone function of the degree of selfishness r.

Proof. It is not difficult to check that, for every strategy y and every pure strategy j, $M_t(j, y) = M_s(j, x) + 1/2\ (s+t)\ [M(x, j) - M(j, x) - M(y, j) + M(j, y)] + M_s(y, j) - M_t(x, j)$. The symmetry condition implies that the expression in square brackets is equal to $M(x, 1) - M(1, x) - M(y, 1) + M(1, y)$, and is thus independent of j. Since x is a completely mixed symmetric equilibrium strategy in Γ_s, $M_s(j, x)$ also does not depend on j. Hence, a strategy y is a symmetric equilibrium strategy in Γ_t and every strategy is a best response to it, if and only if the difference $M_s(y, j) - M_t(x, j)$ does not depend on j. Since $M_s(y, x) - M(x, x) = 0$, this condition is equivalent to (8).

If y satisfies (8) then, for every $0 \le \theta < 1$, $M_s((1-\theta)\, x + \theta\, y, j) = M_{(1-\theta)\, s\, +\, \theta\, t}(x, j)$ ($j = 1, 2, \ldots, n$), and hence the completely mixed strategy $z = (1-\theta)\, x + \theta\, y$ is (by the first part of the proof) a symmetric equilibrium strategy in Γ_r, where $r = (1-\theta)\, s + \theta\, t$.

Direct computation shows that the equilibrium payoff $M(z, z)$ is equal to $M(x, x) + \theta\,[M(y, y) - M(x, x)] - \theta\,(1-\theta)\,[M(x, x) + M(y, y) - M(x, y) - M(y, x)]$. Since, by (2), $(s+t)\,[M(x, x) + M(y, y) - M(x, y) - M(y, x)] = (t-s)\,[M(y, y) - M(x, x)]$, this gives $(t^2-s^2)\,M(z, z) = (t^2-s^2)\,M(x, x) + (r^2-s^2)\,[M(y, y) - M(x, x)]$. ■

Every symmetric 2×2 game trivially satisfies the symmetry condition. Proposition A.0 hence explains, from a more general point of view, the monotonicity of the payoffs corresponding to the completely mixed symmetric equilibria in the games in Fig. 1 and Fig. 2 and the particular shapes of the graphs describing the symmetric equilibrium payoffs in these games. It remains to find out what other games satisfy the symmetry condition. Recall that a game with a finite number of players is called a potential game (Monderer and Shapley, 1996) if the payoff function of each player can be written as the sum of two functions, such that the first function (called the potential) is the same for all players and the second function only depends on the strategies played by the other players.

Proposition A.1. The symmetry condition holds if and only if the symmetric $n \times n$ game Γ is a potential game.

Proof. An $n \times n$ matrix P is a potential for Γ if and only if P is symmetric and there exists a vector b such that $M(i, j) = P(i, j) + b(j)$ for $i, j = 1, 2, \ldots, n$. It is straightforward to check that if such a matrix exists then the symmetry condition holds. Conversely, $M(i, j)$ can always be written as $P(i, j) + [M(1, j) - M(j, 1)]$, where $P(i, j) = M(i, j) - M(1, j) + M(j, 1)$. If the symmetry condition holds then $P(i, j) = P(j, i)$, and P is then a potential for Γ. ■

References

Bomze, I. M., and Weilbull, J. W. (1995). Does neutral stability imply Lyapunov stability? *Games and Economic Behavior* **11**, 173–192.

Bernheim, B. D., and Stark, O. (1988). Altruism within the family reconsidered: Do nice guys finish last? *American Economic Review* **78**, 1034–1045.

Güth, W. (1995). An evolutionary approach to explaining cooperation behavior by reciprocal incentives. *International Journal of Game Theory* **24**, 323–344.

Güth, W., and Yaari, M. E. (1992). Explaining reciprocal behavior in simple strategic games: an evolutionary approach. In (Ulrich Witt, ed.) *Explaining process and change: approaches to evolutionary economics*. Ann Harbor: Michigan Univ. Press.

Monderer, D., and Shapley, L. S. (1996). Potential games. *Games and Economic Behavior* **14**, 124–143.

Okuguchi, K. (1976). *Expectations and stability in oligopoly models.* Lecture Notes in Economics and Mathematical Systems 138. Berlin: Springer-Verlag.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *The American Economic Review* **83**, 1281–1302.

Stark, O. (1989). Altruism and the quality of life. *American Economic Review* (Papers and Proceedings) **79**, 86–90.

Stark, O. (1995). *Altruism and Beyond.* Cambridge: Cambridge Univ. Press.

van Damme, E. (1987). *Stability and perfection of Nash equilibria.* Berlin: Springer-Verlag.