THE AFFECTIONS AND THE PASSIONS: THEIR ECONOMIC LOGIC

by

Jack Hirshleifer
University of California, Los Angeles

Working Paper No. 652
Department of Economics
University of California, Los Angeles
March 1992

THE AFFECTIONS AND THE PASSIONS: THEIR ECONOMIC LOGIC

Abstract

Two types of emotions can be distinguished: the affections (stable patterns of malevolence or benevolence toward particular others) and the passions (action-dependent responses to friendly/unfriendly acts). Either type may serve to induce cooperation from a self-interested party, subject to several limitations. Shakespeare's "King Lear" and O. Henry's "Gift of the magi" are examined as instances where benevolence fails to elicit cooperation. To meet the necessary condition for evolutionary survival, in competition with rational self-interested players, the emotional party must benefit not only in "feel good" utility terms but in actual material payoff. SILVER RULE, as an instance of reactive passionate behavior related to the familiar TIT FOR TAT behavior, is shown to lose out in evolutionary competition against RATIONAL play unless the SILVER RULE player, whenever he has the first move, acts as a RATIONAL player would.

THE AFFECTIONS AND THE PASSIONS: THEIR ECONOMIC LOGIC

In standard economic models, human action is driven solely by material self-interest. Real-world behavior, we all know, often violates this principle. Yet the most obvious violation, parental self-sacrifice on behalf of offspring, can fairly easily be explained away as responding to a longer-run concept of self-interest. After all, it is only through direct or collateral descendants that our genes can be propagated into the more remote future (see, e.g., Hamilton [1964], Dawkins [1989]). But there are other violations, notably willingness to sacrifice one's life for an unrelated individual or even for an abstract cause, that cannot be so easily disposed of. I will be attempting to explain some of these in terms of the economic logic of the emotions.[1]

As used here, the emotions encompass all those internalized impulses that may, to a greater or lesser extent, lead a person to over-ride his or her material self-interest. The sub-category I call the affections are relatively stable patterns of concern for the well-being of particular others. These concerns may be benevolent (as in maternal love) or malevolent (e.g., xenophobia).[2] Actions motivated by the affections, while running counter to strict self-interest, need not be inconsistent with rationality -- only that the impacts upon friends and enemies as well as upon oneself are factored into the cost-benefit calculation. Under the influence of what I call the passions, in contrast, the actor does not

---

[1]Hirshleifer [1987] represented an earlier effort on my part to deal with these issues. Frank [1988] is a more extended treatment of somewhat parallel ideas.

[2]My earlier paper (Hirshleifer [1987]) used the word "sentiments" for what are here called "affections." The latter term seems to capture the meaning more precisely, it being understood that both sympathetic and antipathetic attitudes are included.

calculate but instead "loses control." Outrageous behavior, even on the part of your beloved child, may precipitate a furious reaction -- possibly for the child's own good, possibly not. And on the other hand an act of chivalry, even on the part of an enemy, may trigger unthinking and even imprudently generous reciprocation. In sum: the affections are autonomous, the passions reactive; the affections are stable, the passions situation-dependent; the affections are cool, the passions are hot.

That our motivations incorporate non-self-interested as well as self-interested considerations is not a very novel idea. But the hypothesis here is more specific, to wit, that the emotions are in effect a trick played upon us by Nature. Non-self-interested drives have survived as constituents of the human psyche only because, at a deeper level, they really <u>are</u> consistent with self-interest. Paradoxically, our material goals are sometimes better achieved when not intentionally pursued.

1. <u>Dealing with a "Rotten Kid" (Becker [1976]</u>

Suppose a loving parent (think of King Lear) is attempting to elicit cooperation from a self-interested "Rotten Kid" (e.g., either of his wicked daughters Goneril and Regan). Figure 1 illustrates how things <u>might</u> have worked out well, even though for Lear they did not.

Thinking specifically of Goneril, the eldest daughter, her income Y_G is plotted on the vertical axis while Lear's income Y_L is scaled along the horizontal axis. U_L is a typical indifference curve for Lear, its negative slope indicating that he attaches positive utility both to his own income and to her income. (In contrast, Goneril's merely self-interested utility is illustrated by the horizontal indifference curve U_G.) By assumption,

Goneril has the first move, choosing an initial income-pair along the opportunity curve QQ'. Thereafter, Lear can respond by transferring income from himself to her on a 1:1 basis.

In the absence of any hope of reward from Lear, Goneril's self-interested optimum would be at point M in the diagram -- yielding the largest income Y_G she can attain along QQ'. But in the light of Lear's known benevolence, Goneril does better choosing point J where the joint income $Y_G + Y_L$ is maximized. Lear will then make the indicated transfer to her, moving northwest along the 135^o "transfer line" TT' to his tangency optimum at point A. Goneril is evidently better off at A than at M. In the terminology of E.O. Wilson [1978], Lear's "hard core" (unconditional) altruism has elicited "soft core" (pragmatic) altruism on her part.[3]

The crucial point for our purposes is that Lear's benevolence not only helps Goneril but redounds to his own benefit. And this is true not just in terms of Lear's psychic or "feel good" satisfaction but <u>even in a strictly material sense</u>: Lear's actual income Y_L (and not just his utility U_L) is greater at point A than at point M. This epitomizes my thesis. Benevolence has survived as part of the human psyche because it may improve our prospects in evolutionary competition, allowing us to maximize our material advantages in circumstances where simple selfish pursuit of them will not. In effect, Lear's benevolence guarantees an implicit mutually advantageous contract with Goneril, a contract that would not otherwise be enforceable.

---

[3]Very much in the tradition of Adam Smith, Wilson was arguing that benevolent "hard core" altruism is a weaker force than "soft core" altruism (enlightened self-interest) in eliciting cooperation over large social groupings. The emphasis here is quite different, pointing to the fact that hard core and soft core altruism often <u>complement</u> one another.

However, this analysis applies only under certain limiting conditions. And specifically, only if Goneril has the first move while Lear has an ex-post opportunity to make a compensating transfer of income. Where Lear went wrong was in voluntarily making his transfers to Regan and Goneril ex-ante rather than ex-post. Having nothing left to gain from him thereafter, as merely pragmatic altruists Regan and Goneril had no reason to reciprocate his unilateral gifts to them.

One other qualification: Lear's benevolent motivation must be sufficiently strong. A somewhat different preference pattern for Lear is illustrated by a typical alternative indifference curve U'_L (dashed in the diagram). Once again the slope is negative, indicating that his own income and Goneril's income both enter positively into Lear's utility function. But the magnitude of Lear's benevolence is less than before: specifically, if Goneril were to choose the same family-income-maximizing point J along QQ', Lear would now make a positive but smaller voluntary ex-post transfer ending up at his tangency point B. Goneril now observes that her income at point B if she cooperates will be less than what she can achieve on her own at point M. Thus Lear's "deficient" benevolence leads to an outcome M where both parties are worse off than at A.

Another point of interest: supposing that Lear's benevolence is indeed inadequate as just illustrated, would some degree of benevolence on Goneril's part help repair the deficiency so that the two find it advantageous to cooperate after all? In Figure 2, a typical indifference curve U_G for Goneril is shown as also having the negative slope that reflects a degree of benevolence. In the situation pictured Goneril would be motivated to cooperate, since she achieves higher utility at Lear's ex-

post tangency point at B as compared with the non-cooperative outcome at point M. So once again there is a mutual utility gain from cooperation. But Goneril's utility gain here is only of the "feel good" type -- it masks an actual material loss. It is Lear who now absorbs the material advantage of the cooperation between the two.[4] Thus, in general, benevolence on the part of the first-mover (offspring) is less materially profitable to the donor than benevolence on the part of second-mover (parent). We would expect, in consequence, that evolution will select more for parental rather than offspring benevolence.[5] This is of course broadly consistent with what can be observed among animals and humans.

What of the passions? We have seen that only strong positive affections can elicit mutually advantageous cooperation,[6] and such sentiments are almost necessarily limited to a few "significant others." However, only rarely can we limit our interactions to a precious few. In contrast with the affections, the passions are responses to others' actions or behavior, and as such need not be directed only to parties having a close relation with us. We can get as furiously angry at a stranger who wrongs us

---

[4]Strictly speaking, without hope of reciprocation Goneril's initial move would take her in Figure 2 to a point M' (not shown) where the opportunity curve QQ' is tangent to one of her indifference curves. Such an M' point might or might not be inferior to point B in terms of Goneril's utility. If M' is inferior to B, Goneril will choose J along QQ' and thereby elicit some transfers from Lear; otherwise, she will stay at M' and forego any transfers. Either way, both B and M' yield Goneril a lower material return than point M.

[5]On the reasonable assumption that, at critical phases of the generational life cycle, parents are more in a position to reward or punish offspring than the reverse.

[6]As for the negative affections, malevolence (as reflected by willingness to sacrifice income to hurt another) is generally not an effective technique for eliciting cooperation (Hirshleifer [1987].

as when the same fault is committed by a loved one.

Figure 3 illustrates how a particular pattern of passions, taking the form of what I will call an Anger/Gratitude Response curve (AGR), can serve to elicit cooperation. I now drop the Goneril vs. Lear metaphor and speak only of a self-interested and dispassionate first-mover ("First") facing a passionate second-mover ("Second"), the respective incomes being Y_F and Y_S. In the anger region, Second's AGR demonstrates that he is uncontrollably impelled to punish F for choosing an initial income-pair too far from the aggregate-income-maximizing point J. Punishment here is assumed to take the form of Second accepting a loss of a unit of income in order to impose a corresponding loss on First. Hence the "deprivation lines" D_1, D_2, ... all have 45^o slopes. As F's choice along QQ' approaches point J, S's anger is appeased and eventually his gratitude region is entered. Here Second willingly transfers income to First, hence the "transfer lines" T_1, T_2, ... have 135^o slopes as in Figures 1 and 2. The crucial feature making cooperation possible is the upward slope of the AGR curve in Figure 3, reflecting the fact that the second-mover is "appeasable" -- his anger weakens and his gratitude strengthens as the first-mover becomes more cooperative.[7]

To summarize, under certain special assumptions -- as to the available opportunities, the protocol of interaction (e.g., who moves first), and the patterns of affections or passions -- the emotions may serve to guarantee the execution of threats and promises, so that cooperation may be achieved in the absence of external enforcement of contract.

---

[7] A more extended discussion of how the shape of the AGR curve affects the prospects for cooperation is provided in Hirshleifer [1987].

In the remainder of the paper I will be changing the analytical style, using simple game matrices to represent the nature of the relationship between the parties. While involving some loss of detail, this simpler technique allows coverage of a wider range of situations.

2. Prisoners' Dilemma -- Escaping the trap via benevolence

The familiar Prisoners' Dilemma (PD) is only one of a huge number of possible payoff environments, but it will serve to illustrate a number of issues bearing upon the "emotions as a trick" hypothesis.

In Table 1, Matrix 1 is the standard normal form that defines the PD, the choice options on each side being COOPERATE (C) versus DEFECT (D). In the paired (Row,Column) payoffs shown in the various cells, letters later in the alphabet correspond to higher material payoffs: $w < x < y < z$. Matrix 2 is a specific numerical instance. Since strategy D dominates C for each player, no matter who moves first (or if they move simultaneously), it is rational for each to choose D over C -- so that the parties end up at the strategy-pair D,D with payoffs (x,x) instead of the mutually preferred (y,y).

Very much in parallel with the more detailed picture in the previous section, benevolent affections can lead out of the trap. Suppose that the Row player (like the Rotten Kid) is interested simply in the material payoffs shown in Matrix 2, while the Column player (like Lear) has a psychic utility function that incorporates a positive allowance for the partner's payoff. Specifically, letting m and μ be the material and psychic payoffs, we can suppose that:

(1a) $\mu_1 = m_1 + k_1 m_2$

(1b) $\mu_2 = m_2 + k_2 m_1$

where subscripts 1 and 2 identify the Row and Column players respectively and the k_i coefficients serve to index the degree of benevolence.[8]

Each cell of Matrix 3 shows the material payoffs of Table 2 as before, and also the corresponding psychic payoff within brackets -- on the assumptions that $k_1 = 0$ and $k_2 = 3/4$. Now, if the non-benevolent Row player has the first move she will be led, by a calculation in terms of how the benevolent Column player will respond to each of her choices, to choose COOPERATE rather than DEFECT so that the parties end up at the C,C strategy-pair with material payoffs (3,3).[9] This is of course the analog of the solution at point A in Figure 1. Matrix 4 is calculated in terms of a less benevolent Column player with $k_2 = 1/4$. Here the benevolence parameter is so small that DEFECT remains a dominant strategy for Column, hence Row will also choose DEFECT. So, as the analog of point M in Figure 1, the parties remain at the trap outcome with strategy-pair D,D and material payoffs (2,2). Finally, note that if a benevolent Lear-like Column player must move first, in both Matrix 3 and Matrix 4 he ends up badly off indeed -- with

---

[8] It might be objected that standard game theory already defines the payoffs in <u>utility</u> units, so that no additional psychic adjustments are called for. In the present analysis, material and psychic payoffs are both relevant: the parties' intentional calculations run in terms of the psychic payoffs, but it is the material payoffs that determine ultimate success in evolutionary competition.

[9] Row's reasoning goes as follows. If she chooses COOPERATE, Column's psychic payoffs dictate that he will respond with COOPERATE (since 5 1/4 > 4), the self-interested payoff to Row then being 3. If she plays DEFECT, Column will respond with DEFECT (since 3 1/2 > 3), so that Row's payoff is 2. Thus it pays Row to be nice, in order to elicit the desired response from Column. (Technically, this is the "sub-game perfect" equilibrium.)

material payoffs 0 in each case.

While time does not permit further exploration, a wide range of results are obtainable within the Prisoners' Dilemma payoff environment, depending upon the specific values chosen for the material payoffs w,x,y,z and for the benevolence parameters k_1 and k_2. Suffice it to say that -- provided (a) that benevolence is sufficiently powerful and (b) that the benevolent player has the last move -- it is possible to escape the DEFECT-DEFECT trap. Also, somewhat parallel results can be shown in other contexts, for example in the Chicken payoff environment. On the other hand, as has been seen, it is not true that benevolence works in every case. The section following will illustrate a curious failure of two-sided benevolent intentions.

3. "The gift of the magi"

In O. Henry's Christmas story, Della sells her beautiful tresses to buy a platinum fob for Jim's treasured heirloom watch, while Jim sells the watch to purchase an expensive set of combs worthy of Della's hair. What went wrong?[10]

Matrix 5 shows the payoffs for Jim and Della respectively, where the strategy choices for each are BUY GIFT versus DON'T BUY. As before, the first expression in each cell is the return to the Row player (Jim) and the second expression is the return to the Column player (Della). (In this section all the matrix elements are to be interpreted as psychic utilities.) Looking at the upper-left cell, e_J is Jim's "egoistic" utility gain from receiving a gift, allowing for the negative utility he attaches to the fact

---

[10] Several game-theory approaches to the O. Henry story are surveyed in Brams [1991], but the interpretation here is somewhat different.

that Della would have to buy the gift out of the household funds. Similarly, b_J is his "benevolent" utility gain from making the gift, once again allowing for its price. Della's e_D and b_D are defined similarly. The elements in the other cells of Matrix 5 then fall into place immediately.

Considering Jim's decision, the first thing to note is that his choice will depend <u>solely</u> upon whether or not his net benevolent utility gain b_J exceeds zero -- that is, whether the positive psychic utility he attaches to Della's pleasure from the gift exceeds the negative utility associated with the necessary cash outlay. (Since his egoistic e_J utility element appears only in the first column and enters symmetrically in the upper and lower cells thereof, its influence washes out so far as his decision is concerned.) So Jim might be only limitedly benevolent in the sense that his e_J is much higher than his b_J, or he might be super-benevolent in the sense that his e_J is tiny in comparison with b_J -- either way, his decision whether or not to buy a gift for Della will be unaffected. (And of course, corresponding statements apply to Della's decision.)

The situation as it might have appeared initially to Jim and Della is summarized in the numerical payoffs of Matrix 6. These are calculated on the premises that:

(i) each gift costs \$3, and this cost enters negatively unit-for-unit into the psychic utility on each side (\$1 outlay = -1 util);

(ii) each side's egoistic valuation of the gift to self and benevolent valuation of the gift to the other is the util equivalent of \$2.[11]

---

[11] Thus the first number in the upper-left cell of Matrix 6, Jim's payoff -2, is the sum of his $e_J = 2-3$ and $b_J = 2-3$. The other numbers in the matrix are derived similarly.

These are "golden rule" valuations in the sense that each party values a gift received at no more and no less than a gift conferred, both on the benefit and on the cost side. In Matrix 6 DON'T BUY is then dominant for both Jim and Della. The reason is evidently not selfishness, or even inadequate benevolence. As already seen, even "super-benevolence" would not have changed the outcome. Being aware of their straitened financial circumstances, Jim and Della both realize they need to behave "sensibly" -- the gift is uneconomic as compared with the importance of saving cash for more urgent family needs.

The twist in O. Henry's story is that each spouse discovers a way of seemingly imposing the cost only (or mainly) upon him/herself -- by sacrificing something that is treasured largely out of personal vanity (Jim's watch, Della's hair) -- so that the gift can be purchased without drawing upon the family exchequer at all. Della gives up her hair, Jim gives up his watch. In terms of payoffs, let us suppose:

(iii) each side values the self-sacrifice as the equivalent of -1 unit of "egoistic" utility.

Matrix 7 shows the revised payoffs from Jim's point of view (Jim being of course unaware of Della's parallel reasoning). Now BUY GIFT is dominant for him, on the assumption that Della will be behaving as initially envisaged. Thus Jim's choice is based upon the expectation that the outcome will be at the upper-right cell of Matrix 7, where the apparent payoffs are 1,-1. (Della's -1 = 2-3 in this cell is Jim's perception of the payoff she envisages ex ante for this strategy-pair, the negative value being "erroneous" in reflecting her mistaken belief that his gift to her would cost $3 out of the family finances. Ex post, Jim expects, she will be

favorably surprised to find that her utility is 2.) Correspondingly, Della anticipates the outcome will be at the lower-left cell of Matrix 8 with (ex ante) payoffs -1,1.

The final question that arises is whether, after all has been revealed, they are happy with what has transpired. The superficial answer is no, since the values of the gifts have been lost: the combs are useless without the hair, and the fob without the watch. But more profoundly, O. Henry suggests, they really are happy in the realization that each was willing to sacrifice his/her false values -- even if only to indulge the false values of the other.

4. Loss of control -- the SILVER RULE strategy

I now turn from the affections to the passions. There is something of a mismatch in attempting to provide a game-theory interpretation of the passions. Game theory essentially involves choice, whereas the passions suggest emotional compulsions ruling out the ability to choose. Nevertheless, a simple illustration may be illuminating.

In the Prisoners' Dilemma context, suppose that "loss of control" (on the part of the second-mover or reactive player) takes the form of mentally locking him into what I have called SILVER RULE behavior (Hirshleifer [1982], p. 24). To wit, a SILVER RULE (S) player will make a compulsive COOPERATE response (gratitude) to an opponent choosing COOPERATE (C) on her first move, while DEFECT (D) play on the first move forces him to an obligatory DEFECT response (anger).[12]

---

[12]Huang and Wu [1991] offer a somewhat different analysis of what they call "emotional" (i.e., "passionate") behavior. For them, the analog of my SILVER RULE would be pictured by adjusting the payoffs of the second-mover. Specifically, such a player would get an extra psychic return from responding to DEFECT with DEFECT and to COOPERATE with COOPERATE, but apart

The situation is pictured in Matrix 9. Here the first-mover, Row, as a rational player has the usual choice between her C and D strategies. If the second-mover, Column, were also rational, he would always be choosing the dominant strategy D; knowing this, Row would surely choose D as her first move. However, a passionate second-mover is unable to choose rationally but is locked into his S behavior. Assuming the rational Row player is aware of this, she will now choose COOPERATE -- so that the two parties end up with 3,3 rather than 2,2 payoffs.[13]

Since the mutually improved payoffs can be achieved only if Row knows that Column is a passionate player, the proper behavior on the part of Row is more likely to be elicited if Column can emit a reliable signal of his true character. However, a clever rational player may then try to mimic this signal, fooling the first-mover into choosing COOPERATE only to meet a DEFECT response. This possibility raises issues of great interest that cannot be pursued here. I will only note that both rational first-movers

---

from that would still be making a rational strategic choice. I would argue that this approach fails to capture the crucial "loss of control" aspect of passionate behavior.

[13] SILVER RULE has a certain family resemblance to the well-known TIT FOR TAT strategy, as discussed for example in Axelrod [1984]. A TIT FOR TAT player, like a SILVER RULE player, is not supposed to be making a rational choice but instead is locked into a pattern of reaction to the opponent's prior move. However, Axelrod defined TIT FOR TAT in the context of a simultaneous-play multi-round interaction: in each single round the players move simultaneously, but in later rounds their choices depend upon the opponent's previous behavior. In contrast, the alternative "immediate response" interpretation of TIT FOR TAT in Hirshleifer and Martinez Coll [1988] and Martinez Coll and Hirshleifer [1991] is simultaneous-play single-round: the interaction occurs in one round, but the players are symmetrically able to identify the opponent's move in time to make the appropriate response in the very same round. The context assumed in this paper differs again: here the interaction is sequential-play single-round. The timing is asymmetrical -- the first-mover must make her choice in the dark, whereas the second-mover can observe what his opponent does before making his own move.

and passionate second-movers stand to gain if they can detect and punish mimics, so we that we would expect a kind of co-evolution of ability to cheat and the ability to detect cheats.[14]

Setting aside the mimicry issue, the next section addresses a different evolutionary question -- whether SILVER RULE (passionate) players can survive in long-run evolutionary competition against rational players.

5. The evolutionary competition between passion and rationality[15]

The analysis in the previous sections suggested that: (i) when a SILVER RULE (S) player has the last move, ordinary rational players will be induced to behave cooperatively, and (ii) in that case both types of players receive improved payoffs. The question addressed here is whether those conclusions remain valid or need to be modified when rational and SILVER RULE (passionate) behavioral types are in evolutionary competition.

In evolutionary competition, more profitable forms of behavior will tend to gain increased representation in the overall population, while less profitable behaviors will correspondingly suffer diminished representation. Several evolutionary outcomes are possible: one of the types may drive the other to extinction, or there may be an interior equilibrium such that both types survive -- in some fixed stable proportion, or else varying over a stable cycle -- or, finally, the proportions may continually change in a random or chaotic way without any tendency to settle down at all.

Looked at in terms of evolutionary competition, it becomes evident that

---

[14] On the signalling/cheating problem, see for example Trivers [1971], Dawkins [1989], Ch. 6,10 and Frank [1988], Ch. 5.

[15] The evolutionary competition between rational and emotional players is explored, in a somewhat different context, in Guttman [1991].

the previous section represented only a partial analysis, in that the passionate player always had the last move. In large populations where individuals encounter one another randomly, each type should find itself in a position to have the last move about half the time. But since SILVER RULE was defined only as a pattern of reaction, i.e., as second-mover behavior, just how such a player would act when given the first move remains open.

There are three main possibilities. SILVER RULE on the first move may: (1) Always choose COOPERATE; (2) Always choose DEFECT; (3) Always do whatever a RATIONAL player would do. All three variants are worth considering.[16]

To characterize the nature of the evolutionary equilibrium, we can employ the concept called an "evolutionarily stable strategy" (ESS) as defined in Maynard Smith [1976]. Briefly, a strategy s is an ESS in competition with another strategy s' if either of the following holds:

(i) $V(s|\hat{s}) > V(s'|\hat{s})$ or

(ii) $V(s|\hat{s}) = V(s'|\hat{s})$ and $V(s|\hat{s}') > V(s'|\hat{s}')$

Here V symbolizes payoff, while $\hat{s}$ and $\hat{s}'$ are the "vertices" representing states where all of the population are following strategy s or all are following s', respectively. The interpretation is that strategy s is an ESS if: (i) at its own home vertex, s defeats the intruder s', or (ii) supposing s only ties the intruder there, it defeats s' when s becomes the intruder at the home vertex of the s' strategy.

---

[16] One might at first think that, whichever variant is adopted, passionate SILVER RULE players would not be able to survive in competition with rational ones. After all, a rational player could always behave like a SILVER RULE player whenever it is profitable to do so, and behave differently otherwise. Nevertheless, it turns out there are circumstances in which the rational player does not have the opportunity to profitably behave differently from the SILVER RULE player.

Calculating in terms of the underlying Prisoners' Dilemma payoffs of Matrix 2 in Table 1, the ESS results are summarized in Table 2. Let us start with variant s1, where a SILVER RULE player ("he") would choose COOPERATE given the first move. At the $\hat{s}1$ vertex, i.e, where essentially all the population is playing s1, what happens if a RATIONAL (r) player ("she") enters? We have to separately calculate the payoffs for r and for s1 when each has the first move (to be denoted V') and with the second move (to be denoted V"). The overall payoff V to each type will of course be the average of V' and V". It is straightforward to show that:

At the $\hat{s}1$ vertex: $V'(s1|\hat{s}1) = 3$ and $V'(r|\hat{s}1) = 3$

$V"(s1|\hat{s}1) = 3$ and $V"(r|\hat{s}1) = 4$

At this vertex the payoff to s1 when it has the first move is 3, and when it has the second move is 3, so its average payoff at its own vertex is of course $V(s1|\hat{s}1) = 3$.[17] The payoff to r when it has the first move is 3, and when it has the last move is 4, so $V(r|\hat{s}1) = 3.5$.[18] Since neither of the conditions (i) or (ii) holds for s1, that variant of SILVER RULE cannot be an ESS.

To see if r can be an ESS when played against s1, we calculate:

At the $\hat{r}$ vertex: $V'(s1|\hat{r}) = 0$ and $V'(r|\hat{r}) = 2$

$V"(s1|\hat{r}) = 2$ and $V"(r|\hat{r}) = 2$

---

[17] By definition, an s1 player with the first move chooses COOPERATE. At the $\hat{s}1$ vertex every other member of the population is also s1, hence will reply with COOPERATE. Thus, using Matrix 2, the payoff to s1 players, with or without the first move, is 3.

[18] A RATIONAL player with the first move at the $\hat{s}1$ vertex, knowing that every other member of the population is s1, will choose COOPERATE in order to induce a COOPERATE reply (payoff 3). If on the other hand the RATIONAL player has the last move, after the opening COOPERATE by the s1 player, it will respond with DEFECT (payoff 4).

So at the $\hat{r}$ vertex the average payoff to r is $V(r|\hat{r}) = 2$, while the average payoff to s1 is only $V(s1|\hat{r}) = 1$.[19] So condition (i) holds for r. Thus, the RATIONAL strategy is the only ESS, played against the s1 version of SILVER RULE, i.e., the outcome of evolutionary competition between these two types will be an all-RATIONAL population.

For the competition between r and s2, calculating just as before:

At the $\hat{s}2$ vertex: $V'(s2|\hat{s}2) = 2$ and $V'(r|\hat{s}2) = 3$

$V''(s2|\hat{s}2) = 2$ and $V''(r|\hat{s}2) = 2$

Here the payoff to s2 when it has the first move is only 2 (since an s2 player opens with DEFECT, and whomever he encounters will reply in kind). Since the population consists almost exclusively of such s2 players, the same payoff will be received when an s2 player has the second move, so its average payoff at this vertex is of course $V(s2|\hat{s}2) = 2$. The RATIONAL player, in contrast, anticipating the "reply in kind" associated with SILVER RULE, will play COOPERATE on the first move with a payoff of 3 but of course only 2 when it has the last move. Once again, neither of the conditions (i) or (ii) holds for s2, which also cannot be an ESS.

To see if r is an ESS against s2, we calculate:

At the $\hat{r}$ vertex: $V'(s2|\hat{r}) = 2$ and $V'(r|\hat{r}) = 2$

$V''(s2|\hat{r}) = 2$ and $V''(r|\hat{r}) = 2$

So r and s2 are tied at the $\hat{r}$ vertex. But, we saw above, r defeats

---

[19] At the $\hat{r}$ vertex where every other member of the population is RATIONAL, a RATIONAL player with the first move will always choose DEFECT -- knowing that, whatever she does, the reply will be DEFECT. So RATIONAL receives a payoff of 2 at its own vertex, whether or not it has the first move. The s1 player will by definition choose COOPERATE on the first move, with 0 payoff; on the second move, it will necessarily be replying to DEFECT in which case its own response is DEFECT, with payoff 2. So s1 receives an average payoff of 1 at the $\hat{r}$ vertex.

s2 at the $\hat{s}2$ vertex. So once again, RATIONAL is the only ESS. That is, played against the s2 version of SILVER RULE, the evolutionary outcome would be an all-RATIONAL population.

Finally, what happens under the s3 version of SILVER RULE -- where such a player would, given the first move, do whatever a RATIONAL player would do? Here it is intuitively evident that the two strategies would be always tied. Hence there is no ESS, and we can expect the population proportions to drift in an indeterminate way as between the two strategy choices. The point is that here some degree of cooperation would survive. Whereas an all-RATIONAL population would always engage solely in DEFECT-DEFECT play, with payoff 2 on each side, a mixed r,s3 population would engage in a certain fraction of COOPERATE-COOPERATE interactions. At the $\hat{r}$ vertex both types would always choose DEFECT, at the $\hat{s}3$ vertex both would choose COOPERATE. So as the population drifts between these limits, there would be an ever-varying fraction of each type of behavior.

There are of course many ways of extending the analysis. One route would be to consider other forms of "passionate" behavior, for example one in which reward or punishment actions are not limited merely to COOPERATE and DEFECT play.[20] And it would be of interest to consider environments other than Prisoners' Dilemma, and to allow several different passionate behavior types (like the three variants of SILVER RULE) to be simultaneously present in the population. Another route would be to allow for multiple rounds of interaction between the players in any single encounter (as in

---

[20] In Hirshleifer and Martinez Coll [1988], PUNISHER players are introduced who, in effect, confer greater rewards for friendly play and impose larger punishments upon unfriendly play than would be called for by SILVER RULE.

Axelrod [1984]), which would open up an indefinitely large menu of strategies contingent upon the opponent's prior behavior. Still another would be to allow for the extra cost of rational behavior, and for the consequences of error in recognizing opponents' behavior types.

6. Concluding Remark

The real world is characterized by players of varying emotional types, some with fixed and others with more plastic behavior. All of these must have the ability to survive in competition with strictly rational decision-makers. I have illustrated a few of the circumstances supporting such survival, and tried to show how under certain conditions a degree of cooperation will be thereby induced that exceeds what strictly rational players could achieve.

Table 1
PRISONERS' DILEMMA AND OTHER SELECTED MATRICES

Matrix 1
Prisoners' Dilemma
General form

	COOPERATE	DEFECT
COOPERATE	y,y	w,z
DEFECT	z,w	x,x

Matrix 2
Prisoners' Dilemma
Numerical

	COOPERATE	DEFECT
COOPERATE	3,3	0,4
DEFECT	4,0	2,2

Matrix 3
Material and psychic payoffs
($k_2 = 3/4$)

	COOPERATE	DEFECT
COOPERATE	3,3 [3,5.25]	0,4[0,4]
DEFECT	4,0 [4,3]	2,2 [2,3.5]

Matrix 4
Material and psychic payoffs
($k_2 = 1/4$)

	COOPERATE	DEFECT
COOPERATE	3,3 [3,3.75]	0,4[0,4]
DEFECT	4,0 [4,1]	2,2[2,2.5]

Matrix 5
Gift of the magi (I)
General analysis

	BUY GIFT	DON'T BUY
BUY GIFT	e_J+b_J, e_D+b_D	b_J, e_D
DON'T BUY	e_J, b_D	0,0

Matrix 6
Gift of the magi (II)
Apparent situation

	BUY GIFT	DON'T BUY
BUY GIFT	-2,-2	-1,-1
DON'T BUY	-1,-1	0,0

Matrix 7
Gift of the magi (III)
Jim's re-calculation

	BUY GIFT	DON'T BUY
BUY GIFT	0,-2	1,-1
DON'T BUY	-1,-1	0,0

Matrix 8
Gift of the magi (IV)
Della's re-calculation

	BUY GIFT	DON'T BUY
BUY GIFT	-2,0	-1,-1
DON'T BUY	-1,1	0,0

Matrix 9
Prisoners' Dilemma
Rational play vs. SILVER RULE

	C	D	S
C	3,3	0,4	3,3
D	4,0	2,2	2,2

Table 2

EVOLUTIONARILY STABLE STRATEGY (ESS) OUTCOMES

RATIONAL (r) versus three variants of SILVER RULE (s1,s2,s3)*

(based on payoffs of Matrix 2)

r versus s1: r is an ESS (the equilibrium population is all-RATIONAL)

r versus s2: r is an ESS (the equilibrium population is all-RATIONAL)

r versus s3: there is no ESS (the strategy proportions drift randomly)

*DEFINITIONS:

s1: If SILVER RULE moves first, choose COOPERATE

s2: If SILVER RULE moves first, choose DEFECT

s3: If SILVER RULE moves first, choose whatever a RATIONAL player would choose

REFERENCES

Axelrod, Robert, 1984, The Evolution of Cooperation (New York: Basic Books).

Becker, Gary S., 1976, "Altruism, egoism, and genetic fitness; Economics and sociobiology," Journal of Economic Literature, 14:817-828.

Brams, Steven J., 1991, "Games theory and literature," C.V. Starr Center for Applied Economics, New York Univ., Economic Research Reports RR# 91-29.

Dawkins, Richard, 1989, The Selfish Gene, 2nd. ed. (New York: Oxford Univ. Press).

Frank, Robert H, 1988, Passions within Reason: The Strategic Role of the Emotions (New York: W. W. Norton & Co).

Guttman, Joel M. 1991, "Rational actors, tit-for-tat types, and the evolution of cooperation," Dept. of Economics, Bar-Ilan Univ.

Hamilton, W.D., 1964, "The genetical evolution of social behavior, I," Journal of Theoretical Biology, 7:1-17.

Hirshleifer, Jack, 1982, "Evolutionary models in economics and law: Cooperation versus conflict strategies," Research in Law and Economics, 4:1-60.

Hirshleifer, Jack, 1987, "On the emotions as guarantors of threats and punishments," in John Dupre', ed., The Latest on the Best: Essays in Evolution and Optimality (Cambridge MA: MIT Press).

Hirshleifer, Jack and Juan Carlos Martinez Coll, 1988, "What strategies can support the evolutionary emergence of cooperation?" Journal of Conflict Resolution, 32:367-38.

Huang, Peter H. and Ho-Mou Wu, 1991, "Emotional responses in litigation," Economics Dept., Tulane Univ.

Martinez Coll, Juan Carlos and Jack Hirshleifer, 1991, "The limits of

reciprocity," Rationality and Society, 3:35-64.

Maynard Smith, John, 1976, "Evolution and the Theory of Games," American Scientist, 64:41-45.

Trivers, Robert L., 1971, "The evolution of reciprocal altruism," Quarterly Review of Biology, 46:35-58.

Wilson, Edward O., 1978, "Altruism," Harvard Magazine, 81:23-28.

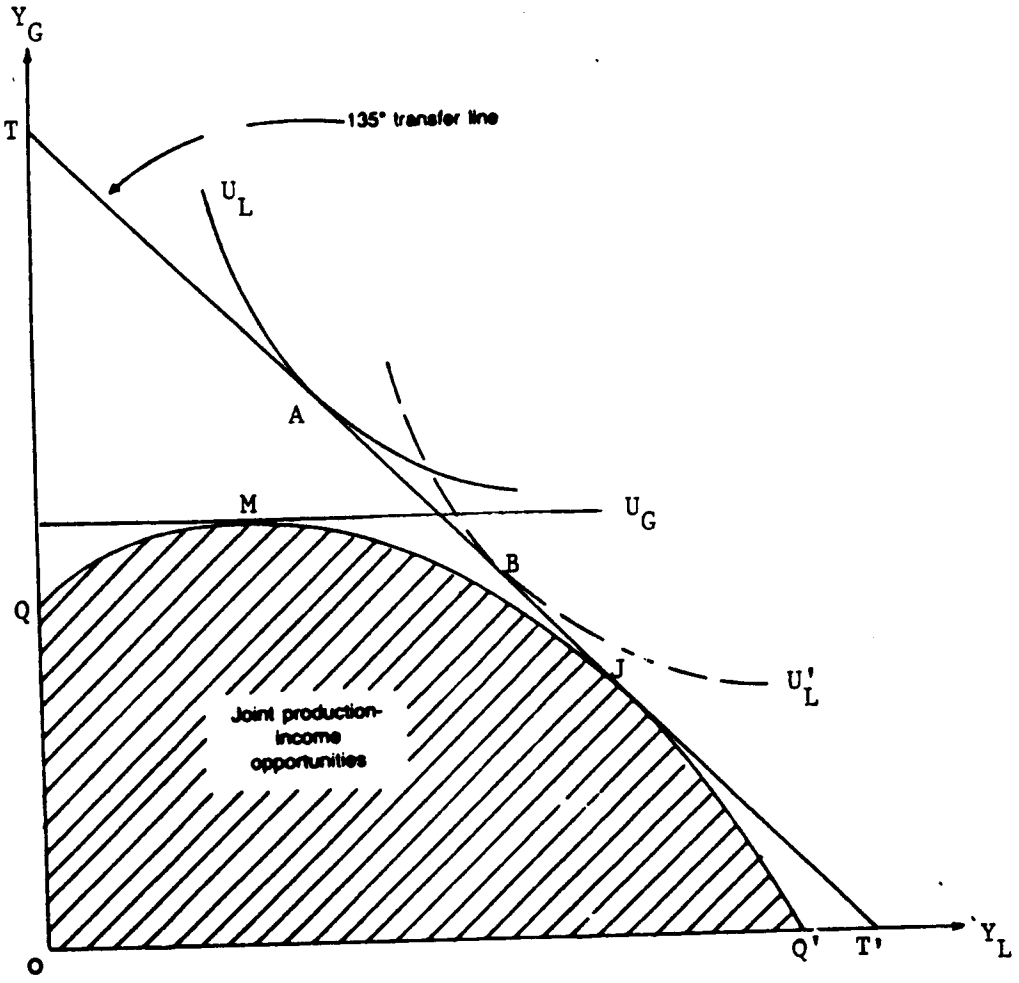

Figure 1

The "rotten kid": Lear versus Goneril

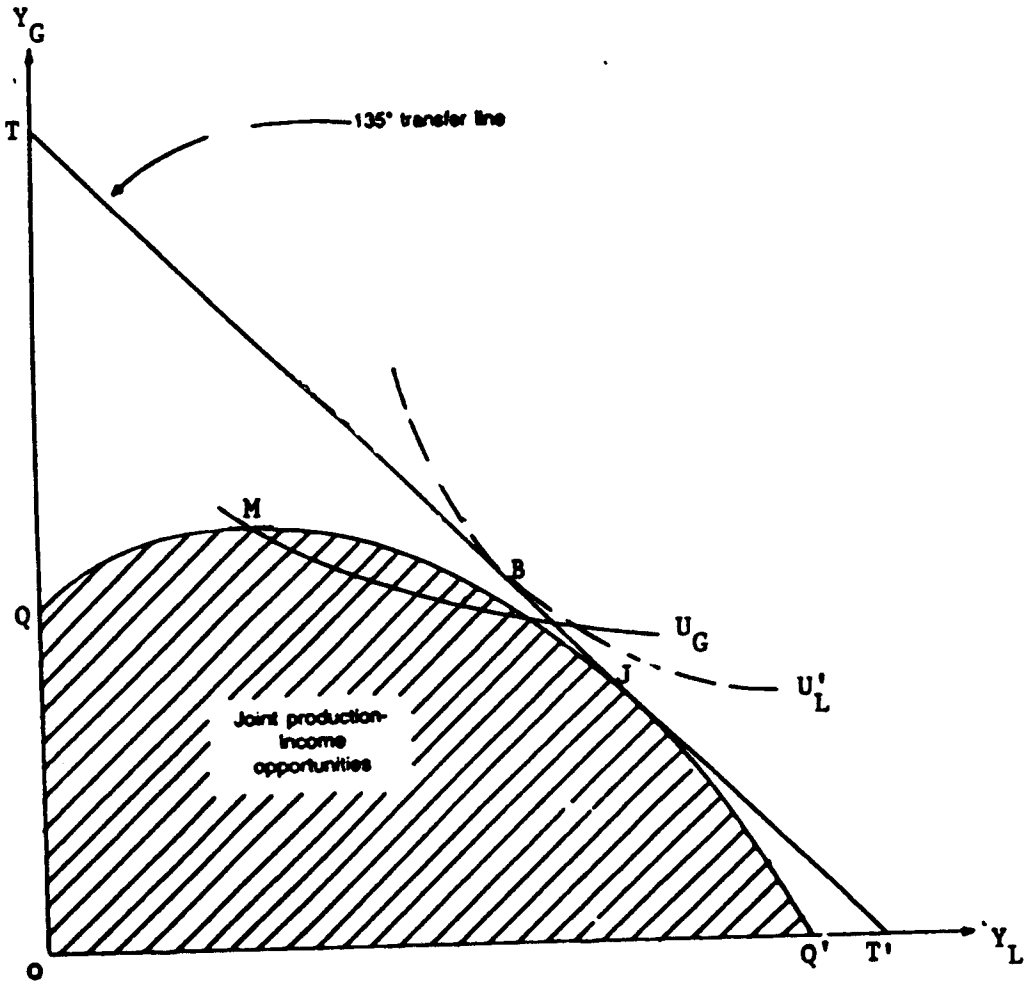

Figure 2

Mutual benevolence: Advantage to the second-mover

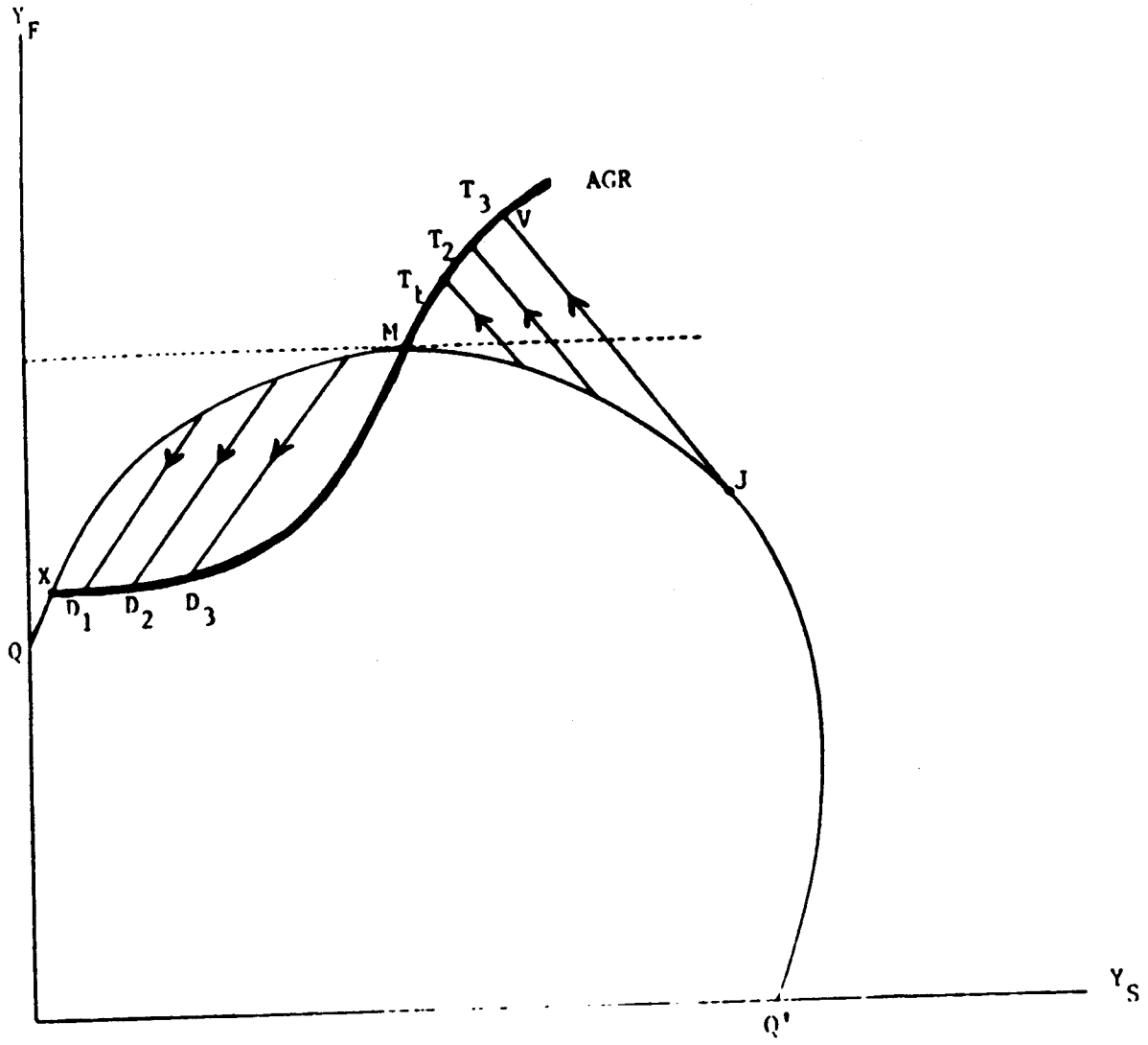

Figure 3

The Anger/Gratitude Response curve -- threat and promise