ON THE EMOTIONS AS GUARANTORS OF THREATS AND PROMISES

by

Jack Hirshleifer

University of California, Los Angeles

UCLA Dept. of Economics
Working Paper #337
August 1984

ON THE EMOTIONS AS GUARANTORS OF THREATS AND PROMISES

The role of the passions or emotions in supporting civil society has been discussed by social theorists and moral philosophers since earliest times. Adam Smith in particular, whose name is more usually associated with the claim that an economic system may function effectively even when men act only in accordance with calculated self-interest, was actually very concerned with aspects of human nature that set limits upon the pursuit of self-interest:

> Nature, when she formed man for society, endowed him with an original desire to please, and an original aversion to offend his brethren. She taught him to feel pleasure in their favourable, and pain in their unfavourable regard. She rendered their approbation most flattering and most agreeable to him for its own sake; and their disapprobation most mortifying and most offensive.[1]

More intriguing than his appreciation of the force of "positive" emotions like benevolence and sympathy are Smith's insights into how even the more dubious passions — pride, vanity, and ambition — may promote the interests of society.[2] This point had of course already been made by Mandeville in The Fable of the Bees [1714]. But Smith comes closer to my theme here in his argument that these "negative" sentiments can be most socially useful precisely when they drive people to undertake activities beyond the bounds of pragmatic self-interest. Thus, of a "poor man's son...visited with ambition" Adam Smith says:

> For this purpose he makes his court to all mankind; he serves those whom he hates, and is obsequious to those whom he despises. Through the whole of his life he pursues the idea of a certain artificial and elegant repose which he may never arrive at, for

---

[1] Smith, Moral Sentiments, p. 212.

[2] On this see especially Coase [1956], pp. 536, 542-543.

> which he sacrifices a real tranquillity that is at all times in his power, and which, if in the extremity of old age he should at last attain to it, he will find to be in no respect preferable to that humble security and contentment which he had abandoned for it.[3]
>
> And it is well that nature imposes upon us in this manner. It is this deception which arouses and keeps in continual motion the industry of mankind. It is this which first prompted them to cultivate the ground, to build houses, to found cities and commonwealths, and to invent and improve all the sciences and arts, which ennoble and embellish human life.[4,5]

The point I will be making is somewhat different, however. I agree that the emotions, positive or negative, can indeed be socially useful in driving a person to act beyond the bounds of pragmatic self-interest. But paradoxically the consequence is not necessarily adverse for the person himself. He can sometimes best further his self-interest by not intending to pursue it. Methodologically speaking also, I will be advancing somewhat beyond the classical discussions referred to above, in providing a systematic analysis of the precise ways in which different emotional sets may promote or subvert socially advantageous arrangements. In attempting to do this I will be following a lead by Becker [1976], who demonstrated how "altruism" can, in effect, force cooperation upon a completely selfish partner (the "rotten kid theorem"). I will try to show more generally here how, and up to what limits, positive or negative emotions can serve a constructive role as guarantors of threats or promises in social interactions.

---

[3] Moral Sentiments, p. 300.

[4] Ibid., p. 303.

[5] Compare also the Talmudic argument that even the evil impulse instilled by God in man is "very good":

> Were it not for that impulse, a man would not build a house, marry a wife, beget children or conduct business affairs.--Cohen [1949].

If a mutually desired objective is to be achieved, it is often necessary that one or more of the parties forego the opportunity to reap a self-interested gain. Intelligence permits reasoning beings to "look around the corner," to visualize the advantages of not pursuing immediate self-interest. But the problem of securing the necessary coordination of actions remains. A "meeting of the minds" — a contract, to use that term in its broadest sense — does not generally suffice; some method of enforcing (or otherwise guaranteeing) performance is generally required.[6]

The most obvious method of enforcement is through the legal-judicial system. But it has long been appreciated that in some cases contracts may be self-enforcing (Macaulay [1963]), the key point being that fear of losing profitable future business with a trading partner may suffice to deter defection here and now. This topic has recently been studied by Telser (1980), using the analytical model of the repeated-play Prisoners' Dilemma ("supergame" theory, in the standard jargon). The difficulty is the well-known incentive to defect at the last round of play. This obstacle can be overcome, so that the contract becomes self-enforcing, when the number of rounds is infinite or at least if there is a sufficiently high probability of play always continuing for another round (see also Luce and Raiffa [1957], p. 102).

In contrast with this line of discussion, I will be dealing solely with single-round games. But my analysis will not be limited to Prisoners'

---

[6]Sometimes a meeting of the minds can suffice, as in adoption of a convention which no-one has any incentive to violate. Agreeing to drive on the right (or left), or to meet under the clock at Grand Central Station, are possible examples (on this see Schelling [1978], Ch. 3). I have shown that such cordination tends automatically to emerge under particular types of game payoff matrices (Hirshleifer [1982], p. 14.)

Dilemma, which is only one of a number of distinct payoff environments combining a mutual gain from cooperation with a self-interested motive to defect. In the situations to be considered the possibility of enforcement stems from an assumed asymmetric game protocol, such that one of the players "has the last word" (Hirshleifer [1977]) and thus is potentially in a position to confer reward or punishment. Since offers of reward or punishment are contingent strategies, we are in the realm of "metagames" in the jargon of the trade (Howard [1971], Thompson and Faith [1981]).

In what follows I will first briefly discuss the nature of threats vs. promises in different payoff environments. Then I provide an explicit analysis for several different categories of emotions. Finally, I will speculate upon the possible reasons why emotions, and other limitations upon "rationality," have survived as part of the human constitution.

I

An individual who makes a threat or promise is pledging to respond in a contingent way to another's actions, with the goal of influencing the other party's choice. The intended effect would presumably always be to the threatener's advantage, though not necessarily to the other's disadvantage. The only point worthy of special note is that a promise or threat must be <u>to do something that the individual would not otherwise be motivated to do</u>. That is what distinguishes these pledges from mere forecasts, however informative, of one's likely responses to another's actions (see also Schelling [1960], Ch. 2).

Matrix 1 illustrates a Prisoners' Dilemma payoff environment, where 4 represents the highest and 1 the lowest of the ordinally ranked returns to each of the players. Here and throughout, unless indicated to the contrary

Matrix 1

Prisoners' Dilemma

	LOYAL	DEFECT
LOYAL	3,3#	1,4
DEFECT	4,1	2,2*

Matrix 1A

Expanded Prisoners' Dilemma

	LOYAL	DEFECT
LOYAL	3,3	1,4
DEFECT	4,1	2,2
TIT FOR TAT	3,3	2,2
REVERSAL	4,1	1,4

Matrix 2

Chicken

	DOVE	HAWK
DOVE	3,3#	2,4*
HAWK	4,2**	1,1

the Column player moves first so that it is the Row player who "has the last word." It is then possible to define contingent strategies for Row. Let the two elementary strategies available to each player be LOYAL and DEFECT. Row's additional contingent strategies can be termed TIT FOR TAT (play LOYAL in response to LOYAL, and DEFECT in response to DEFECT) and REVERSAL (play DEFECT in response to LOYAL, and LOYAL in response to DEFECT). It is standard to represent these contingent strategies by additional rows in the game's normal form, as in the "expanded" Matrix 1A. However, for my purposes it suffices to deal directly with the underlying Matrix 1, where it is easy enough to visualize the effects of threats or promises on the part of the Row player.

With Column having the first move suppose that Row — even though he may contemplate or even intend a contingent strategy like TIT FOR TAT — cannot guarantee to Column that he will be following it. Then, returning to Matrix 1, Column will surely choose DEFECT. (For, his choice of LOYAL would be responded to by Row's DEFECT, leaving Column with his worst payoff of 1.) Row will of course answer Column's DEFECT with his own DEFECT. Thus the parties end up at the cell of Matrix 1 marked with a star; the payoffs are 2 each — their next-to-worst outcomes in each case.[7] This is of course the traditional "trap" equilibrium of Prisoners' Dilemma,[8] a seemingly paradoxical result

---

[7]It might be thought that this analysis is over-elaborate, in that DEFECT is by inspection a "dominant" strategy for each player given the payoffs of Matrix 1. But dominance arguments must be used with great care in sequential-play protocols. It is easy to demonstrate underlying payoff matrices where the first-mover would want to employ a "dominated" strategy, since he can thereby influence the other player's responding choice.

[8]The starred cell is also the unique "Nash equilibrium" of the usual (simultaneous move) Prisoners' Dilemma. However, we are dealing here with a sequential-move rather than simultaneous-move game. It would be possible, using the appropriate expanded matrix in each case, to extend the Nash equilibrium idea to cover sequential-move games but I will not in fact be using that solution concept.

since, by cooperation, the parties could have reaped their next-to-best payoff of 3 each.

This Pareto-superior 3,3 outcome (marked with a # in Matrix 1) would indeed be achieved if Row could guarantee abiding by a promise to play LOYAL in response to LOYAL on Column's part. Column, if he could rely upon that promise, then would play LOYAL on the first move. What makes Row's utterance a promise (rather than a forecast) is that he offers to not do what he is in fact motivated to do when his turn comes up. In promising LOYAL Row engages himself to confer a benefit upon Column, at a cost to himself, should the latter first choose LOYAL himself. But that Row will play DEFECT in response to DEFECT is only a forecast and not a threat, since Row would be doing that anyway.

To illustrate an actual threat, now consider the almost equally famous payoff environment known as the game of Chicken (Matrix 2). Holding to the ornithological metaphor, the two strategies can be called HAWK and DOVE. If Row cannot guarantee performance of a contingent strategy then Column, having the first move, must inspect Row's best response to each of his choices. Evidently, Column's best first move is the "less cooperative" HAWK strategy, leading to an asymmetrical 2,4 outcome in his favor (starred in the matrix.) In order to influence Column to change his strategy, Row would have to guarantee execution of a threat to play HAWK in response to HAWK. Once again, that is an engagement to do something at a cost to himself, but now the action imposes a loss on Column. If Column bows to the threat and plays DOVE instead, Row now plays HAWK leading to the reversed payoff outcome 4,2 in Row's favor (double-starred in the matrix).

On the other hand, Row might sweeten the deal by issuing a threat-and-promise. That is, in addition to the aforesaid threat to play HAWK in

response to HAWK, he could promise to forego part of his gain and respond to DOVE with DOVE. Then, the parties could achieve the symmetrical 3,3 outcome (marked # in Matrix 2). It might seem puzzling that, if Row's threat is solid enough to work, he would ever reduce his gain by combining it with a promise. But throughout this paper we are contemplating the possibility that parties might, owing to emotional limits upon self-interested rationality, not necessarily do what it is in their material advantage to do. In such circumstances, threat-and-promise might conceivably trigger a less hostile response than threat alone.

II

In this section I will examine the consequences of two different scaled classes of emotion: (1) the malevolence/benevolence spectrum and (2) the anger/gratitude spectrum. The first category is action-independent, the second is action-dependent. I will continue to employ the assumption throughout that one player always has the first move, in fact, let us call him First. The other player, Second, is therefore the one who may be in a position to confer reward or punishment.

Benevolence/Malevolence

Figures 1a and 1b illustrate benevolence and malevolence on the part of the last-mover, Second. Thus it is Second's indifference curves that are being pictured, on axes representing the two parties' material incomes Y_S and Y_F. In Fig. 1a Y_S and Y_F are both goods for a benevolent Second, hence his indifference curves have the normal slope and curvature. In Fig. 1b his own income Y_S remains of course a good for Second but now Y_F is a bad from his point of view. Hence the abnormal indifference curve map in Fig. 1b.

I will also be assuming that Second always has the power to transfer income to First, or else to deprive him of income, if he chooses — in each case, at a cost to himself. Specifically, in the transfer mode Second can increase First's income Y_F by giving up his own income Y_S on a 1:1 basis. (Note the dashed 135° "transfer lines" TT in Fig. 1a, where the arrows attached indicate the direction of movement.) In the deprivation mode, Second can reduce First's income but again only by incurring an assumed equal cost himself — indicated by the dashed 45° "deprivation lines" DD in Fig. 1b. (Once again, the arrows indicate the possible direction of movement.)

In Fig. 1a the WEP curve (a "Wealth Expansion Path") connects all the tangencies of Second's benevolent indifference curves with the 135° transfer lines. It follows immediately that, provided the starting position is to the <u>southeast</u> of the WEP curve (that is, provided he is relatively well off and First relatively poorly off to begin with), Second will always transfer exactly that amount of income leading him to a final solution along WEP. Similarly, in Fig. 1b the WEP curve connects all the tangencies of Second's malevolent indifference curves with the 45° deprivation lines. Here Second will always impose enough deprivation to end up along WEP, provided the starting position is to the <u>northeast</u>. (That is, provided both parties are relatively well endowed with income to begin with. Second must be reasonably wealthy in order to afford the cost of inflicting the deprivation, while First must be well off to begin with else the initial situation would not be intolerable to a malevolent Second.)

Another property of the WEP curves will be important in what follows. In Fig. 1a WEP has positive slope throughout. This indicates that, as a benevolent Second grows wealthier, he will want to choose positions involving greater Y_S and greater Y_F <u>both</u>. I.e., as he becomes richer he wants to

end up with more income for himself and more income for the object of his benevolence. This represents a special though reasonable assumption, that Y_S and Y_F are both "normal" or "superior" goods from Second's point of view. By a corresponding argument, in Fig. 1b the curve WEP has negative slope. As Second becomes richer, he prefers to so arrange matters as to have more income himself while leaving the target of his malevolence worse off.

Now we turn to the decisions available to the first-mover. It will be assumed that a joint productive opportunity boundary like QQ' in Fig. 2 always exists, and that First has the sole choice of the productive arrangements to be made -- i.e., he determines the point to be chosen along QQ'. Furthermore, First throughout is assumed to be a merely self-interested individual, neither benevolent nor malevolent. It then follows that, in the absence of any anticipated reaction from Second, First would always simply prefer the most northerly position along QQ' -- point M in the diagram. As it happens, the joint productive arrangements represented by point M do generate some income for Second as well, but this is a merely incidental fact from First's selfish point of view.

But now note the position of Second's (solid) Wealth Expansion Path WEP in Fig. 2. This indicates that, beyond point X where WEP cuts QQ', Second regards himself as wealthy enough to display some benevolence toward First. Thus, First, should he choose his short-sightedly selfish optimum M, would end up doing even better -- since Second would benevolently transfer enough income to him along the transfer line T_1T_1 to end up at point N which lies to the north of M. But that is still not the best that a <u>far-sightedly</u> selfish First could achieve. In fact, it will be evident from the diagram that First should choose point J, where QQ' is tangent to the highest attainable 135° transfer line T_2T_2. True, in the short run, First will have

sacrificed income on behalf of Second. But the latter will then benevolently transfer enough of his enlarged income to First so as to achieve a final indifference-curve tangency at point A. From the selfish First player's point of view, A is better than (north of) N. In fact, A represents the largest income attainable by First under the conditions assumed.

That however is not the surprising part of this result. The real point is, not only does "enlightened self-interest" lead the selfish first-mover to a better outcome thanks to a predictable benefit from Second's benevolence, but the latter gains as well! And the gain to Second is not merely in terms of his psychic satisfaction from seeing First better off. Even in the crassly material sense, the benevolent second-mover himself has gained from his generosity. His own material income Y_S is greater at point A than it would have been otherwise, i.e., had the selfish First player chosen point M along QQ'.

What has happened here is that Second's benevolence guaranteed an implicit "promise" to reward First, a necessary condition for securing cooperation from a merely self-interested individual. In language sometimes used by biologists, the second-mover's "hard-core altruism" has served to elicit the first-mover's pragmatic or "soft-core altruism".[9]

Another aspect of the constructive role played by benevolence is brought out if we look at the alternative (dashed) Wealth Expansion Path WEP' in Fig. 2. This curve is associated with a different possible set of preferences for

---

[9]The terminology is due to E.O. Wilson (1978). However, the point made here diverges from Wilson's. He was concerned to contrast the weakness of "hard-core altruism" (benevolence) compared to pragmatic "soft-core altruism" (enlightened self-interest) as organizers of cooperation in large social units. The analysis here indicates that the two factors may sometimes complement one another in a socially useful way.

Second, suggested by the dashed indifference curves U'. These alternative preferences are qualitatively similar in that they also represent benevolent tastes, but now to a quantitatively lesser degree. In these altered circumstances, if First were to choose the jointly cooperative position J along QQ' as before, Second would then transfer a smaller amount to him — so that the parties would end up at position B along WEP' rather than at A along WEP. But, for the selfish First player, position B is inferior to his short-sightedly selfish optimum M. The conclusion: far-sighted or enlightened self-interest may not suffice to achieve mutual improvement, absent a sufficiently strong degree of benevolence on the part of the player having the last move.

Of course, in the latter situation Second could still promise to provide First with a sufficient reward to induce cooperation, say by pledging to choose position A if First selects J along QQ'. But by assumption there is nothing to guarantee Second's promise — save his own benevolence, which is here inadequate. In fact, there is a critical threshold degree of benevolence on the part of Second needed to elicit cooperation from First. To wit, to provide enough inducement for a far-sightedly selfish First to choose point J, Second's benevolent Wealth Expansion Path must cut the transfer line T_2T_2 north of its intersection with the dotted line through M. Thus, lots of love may do the trick where a little love achieves nothing at all!

If benevolence can serve as guarantor of a promise by the party having the last word, can malevolence serve as a guarantor of a threat? It surely can, but the overall effect is not to elicit cooperation, if we hold to the assumption that the WEP curve in Fig. 1b has negative slope throughout. (I.e., the richer Second gets, the more he can afford to and want to spend on depriving First.) In Fig. 3, then, should the first-mover short-sightedly

choose the most northerly position M along QQ', he will have empowered a malevolent Second to impoverish him to the degree represented by the final position L. If First were instead to choose the efficient productive point J, he would end up still worse off at K. Evidently a selfish First does best to choose point X, where he is initially poorer but where Second is also so poor that he cannot afford to (or at least does not care to) incur costs to impose any further deprivation upon him. Thus malevolence which guarantees execution of a threat has an effect, leading First to choose X instead of M. But the consequence is that both parties are worse off, in terms of psychic satisfaction and in terms of material income as well.

Gratitude/Anger

I now turn to the action-dependent emotional scale ranging from gratitude at the positive end to anger at the negative end of the spectrum. The question as before will be whether these emotions can guarantee execution of threats or promises and thereby promote achievement of mutually beneficial solutions.

Allowing for the possibility of malevolence and benevolence represented a departure from the economist's standard assumption of self-interested behavior, but did not otherwise do violence to the picture of "economic man" as a rational utilitarian calculator. An individual who values other people's income (positively or negatively) as well as his own can nevertheless coolly go about his business of calculating a preferred final outcome in the light of his given preferences and opportunities.[10] But I am now introducing a much more serious departure from the standard assumptions of the economist. To

[10]Such models of interpersonal consumption preferences have been employed by a number of economists including Boulding [1962], Ch. 2, and Becker [1971].

wit, the idea that an individual can be passionate — in the sense of "losing control" and doing what he doesn't really want to do (on this see Schelling [1980]). Or an alternative (and my preferred) interpretation would be that what he wants to do need not depend only upon the final outcome in the utilitarian sense — i.e., strictly upon the ultimate distribution of incomes between the two parties -- but rather may be action-dependent. An income distribution that could be tolerable as an accidental or random event, for example, might lead to violent revolt if seen to be the result of conscious choice on the part of another economic agent. Common observation tells us that, whatever the textbooks assume, such behavior is in fact very important in the make-up of normal human beings. I will be showing that, at least in certain circumstances, such non-utilitarian behavior makes ultimate utilitarian sense!

Since preferences are now assumed action-dependent, we can no longer postulate a fixed preference map defined simply over the parties' incomes Y_F and Y_S. Nevertheless, it is possible to place plausible restrictions upon how a second-mover influenced by anger/gratitude motivations would respond to choices by a first-mover. (By analogy with the foregoing, I assume here that only Second is subject to anger/gratitude emotions, First remaining throughout a coolly calculating and self-interested "economic man.") In Fig. 4, Second's responses to First's choice along QQ' are summarized diagrammatically by his Anger/Gratitude Response Curve AGR. This curve might take on a number of possible shapes, subject to the restrictions that: (1) increasing wealth empowers Second to spend more either on transferring income to or else withdrawing income from First, but (2) the more "cooperative" is First's behavior (i.e., the more his choice along QQ' approaches point J), the less is Second's anger and/or the greater is Second's gratitude — with

the obvious implication for his willingness to confer benefit or injury upon the other.[11]

In the picture of Fig. 4, the "effective anger" region of Second's AGR curve extends from point X to point M along the opportunity boundary QQ'. (If First were to choose any point to the west of X along QQ', Second might be even angrier still, but too impoverished to do anything about it.) As First's choice hypothetically shifts to the east of X so as to provide Second with more income, the latter can increasingly afford to react in an angry way but becomes decreasingly inclined to do so. Beyond the crossover at M, Second enters his "effective gratitude" region. Here as his income grows thanks to First's making a more cooperative productive decision, Second is increasingly more able to react (now in a positive grateful way) and is also increasingly inclined to do so. Thus in the effective anger region, between X and M, the gap between QQ' and AGR first widens and then narrows. But in the effective gratitude region beyond the crossover at M the gap increasingly widens as Second's income increases. This type of situation, it may be noted, corresponds to the threat-and-promise contingent behavior mentioned in Section I above.[12]

---

[11]While there is some diagrammatic similarity between the AGR and the previous WEP curves, they stand logically on quite a different footing. An AGR curve could also be interpreted as a locus of tangencies of transfer lines (or deprivation lines) with the second-mover's indifference curves. But, the crucial point is, owing to the emotional action-dependent effects, Second's entire indifference-curve map changes in response to First's choice along QQ'.

[12]Of course a person capable only of anger can only guarantee execution of threats; a person capable only of gratitude is similarly restricted to promises. Having such a narrowed repertory of behaviors may be less costly, but restricts the contingent strategies available.

The outcome of the pattern pictured in Fig. 4 is at point V. The efficient productive solution is achieved (at J on QQ') plus a redistribution thereafter such that both parties are better off in comparison with First's simple selfish solution at point M. But this is only one of quite a number of possible consequences of an anger-gratitude scaled reaction pattern.

Two other possibilities are pictured in Fig. 5. Here, in contrast to Fig. 4, anger is the dominant emotion influencing Second's behavior. However, this can have very different implications as illustrated by the alternative AGR curves pictured — the (solid) AGR° versus the (dashed) AGR'. In each case, by assumption here the crossover point along QQ' lies to the east of the efficient productive solution J; Second does not give up his wrath until he is very well treated indeed. If Second's emotional set corresponds to AGR°, the best choice for First along X°K° is at point K°. Second's threat has worked so effectively as to force First to make a big sacrifice — going not merely beyond First's simple selfish optimum M but even beyond the jointly efficient outcome. If Second's AGR' curve is applicable, on the other hand, First will prefer point X'. Second's threat does not work at all. Or rather it works, but to the disadvantage of both parties, in view of the steps First will take to escape the threat. X' is inferior for both, not merely in comparison with what could have been achieved by Second's making a transfer after First chooses the jointly efficient outcome J, but even in comparison with First's simple selfish outcome M.

What is the difference between the two cases? AGR' represents a relatively "small" threat in comparison with AGR° -- it does not take hold at all until point X' rather than X°, and it inflicts smaller punishment throughout the relevant range. This feature is not the essential, however. In fact, Fig. 6 provides a comparison where the smaller threat represented by

AGR' is more effective than the bigger threat represented by AGR°. (More effective both in terms of eliciting a more favorable joint outcome, and in providing a higher income to Second as threatener.) As will be evident from inspection of these diagrams, the key feature is not the magnitude of the threat but rather the overall slope of the AGR curve. Or, to be slightly more specific (since the AGR may change from positive to negative slope or vice versa along its length) the key feature is the location of the most northerly point along the AGR curve, since this is what a coolly calculating First can foresee ending up with.

While many other pictures can be drawn illustrating interesting possibilities, the following points summarize the key considerations:

1. The rationally selfish first-mover will want and (assuming that he has the relevant information) be able to achieve the most northerly point along the AGR curve which represents Second's scaled emotional response to his behavior.

2. Second's AGR curve will typically take off at some interior point X along QQ', where he first has income enough to be able to indulge his anger. The AGR curve may have positive or negative slope, but will eventually cross QQ' again at a point where First's move has become generous enough to appease Second's anger.

3. If anger is Second's dominating emotion, reflected by an AGR curve that does not cut QQ' until some point K to the east of the efficient point J, then the solution for First — the most northerly point along AGR (unless M is higher still) — will tend to be toward one extreme or the other, depending upon whether the slope of the AGR is predominantly positive or negative. If the slope is predominantly positive, there are typically positive marginal payoffs to First of appeasement, of bowing to the threat. Note that it does not matter so much how heavy a punishment Second might impose in aggregate,

but only how his reaction responds, on the margin, to First's concessions. A positively-sloped AGR function thus tends toward to an excellent outcome for Second, with First appeasing him heavily. If on the other hand the slope of the AGR is predominantly negative, First typically suffers negative marginal payoffs from making concessions. Here the typical outcome is toward the other extreme, at point X where Second is left too poor to inflict punishment. Of course, here First will end up poor as well.

4. But if gratitude is Second's dominating emotion, reflected by an AGR crossing QQ' to the west of point J, the efficient solution tends to be achieved — provided again that the AGR has positive slope. Both parties then end up better off than at First's simple selfish optimum. (But should the AGR have negative slope, at least over part of its range, this could make for a less cooperative solution.)

III

Biologists and anthropologists have been long concerned with explaining the great gap between human intellectual capacities and those of the nonhuman primates, an advance that appears to go beyond the adaptive environmental requirements of primitive man (Washburn [1960], Rose [1980], Staddon [1981]). But equally mysterious, perhaps, is the survival of those limitations upon self-interested rationality we term the passions or emotions. In The Expression of the Emotions, Charles Darwin emphasized the universality of these emotions over the human species, and also their continuity with behaviors visible among animals. Of course, there would be no special need to explain survival of the passions if they are only "imperfections". Since the development of any trait involves some energetic cost, or at least some opportunity cost in terms of other capacities that have to be sacrificed, we

do not need to explain why all living beings are not unlimitedly fecund, powerful beyond measure, or as speedy as light. But we ought not prejudge the question, as to whether the observed limitations upon the human ability to pursue self-interested rationality are really no more than imperfections — might not these seeming disabilities actually be functional?

The thrust of the argument here has been that certain patterns of environmental payoffs to interpersonal cooperative opportunities can make retention of a capacity for emotion materially profitable. In this paper I distinguished between the action-independent versus the action-dependent passions. The first category was illustrated by non-self-interested motivations (the malevolence/benevolence spectrum), the second by impairment of the ability to calculate owing to reactive "loss of control" (the anger-gratitude spectrum). Given that one party in a social interaction is a selfish and perfectly rational calculator, it turns out that there are circumstances in which it is indeed profitable for the other to diverge from self-interested rationality — in accordance with one or both of these emotional scales.

As a related point, evolutionary biologists have also been concerned to explain the survival of "altruism" on the one hand and "spite" on the other. Despite the psychological connotations of these terms, the interpretation in the standard literature has been entirely operational rather than motivational: "altruism" is taken to mean acting so as to help another organism, at a cost to oneself, while "spite" refers to incurring a cost so as to injure another. Since evolutionary success is a selfish criterion,[13] the biological literature has attempted to explore the different patterns by which proximate

---

[13]"Natural selection will never produce in a being any structure more injurious than beneficial to that being, for natural selection acts solely by and for the good of each." Darwin, Origin of Species, Ch. 6.

sacrifice, the cost incurred to benefit or to injure another organism, can ultimately pay off through some indirect route. (E.g., in the case of "altruism", if one's beneficiaries are kin (Hamilton [1964]), or if an other-benefiting act leads to adequate reciprocation (Trivers [1971]). This paper, in contrast, directly attacks the problem of motivation. (Hence it is not aplicable to lower organisms lacking the capacity for emotions.) What it examines are possible mechanisms whereby individuals may be led to supply the reward or punishment (the positive or negative reciprocation) that make certain forms of social cooperation possible. The vague terms "altruism" and "spite" are inadequate to describe the subtly different forms that these reciprocations may take.

In summary, the models analyzed here represent a special but illuminating case. It was assumed that, in a given payoff environment, a merely self-interested and accurately calculating party has the first move, after which a possibly emotion-influenced agent makes his choice in response. When it comes to the action-independent emotions -- the malevolence/benevolence spectrum -- benevolence can serve to guarantee a promise, but malevolence is not generally effective in guaranteeing execution of a threat. (A curious yet important point is that just a little benevolence may not work either — a generosity threshold must be overcome.) As for the action-dependent reactions, the anger/gratitude scale, gratitude has effects rather parallel with benevolence, but anger tends to be more effective than malevolence in securing cooperation. The key reason for the difference is that malevolence becomes a more powerful force as the responding party's wealth increases, a factor which strongly inhibits any desire on the part of the first-mover to enrich the other by cooperation. But anger erodes as the first-mover's choice shifts toward cooperation, hence the net effect may provide the needed inducement for

achieving a mutually beneficial arrangement. The most general and interesting conclusions are: (i) that absence of self-interest can pay off even measured in terms of material selfish gain and, a parallel but quite distinct point, (ii) that the loss of control which makes calculated behavior impossible can be more profitable than calculated optimization. (It follows that a coolly calculating individual might more or less successfully pretend to be driven by passion-dominated responses. Furthermore, in this pretence he need not be seeking a merely selfish goal!)

Analytically, this paper demonstrates the not-so-paradoxical fact that it is possible to analyze, in terms of effects upon rationally calculated self-interest, the consequences of non-self-interested motivations and of limitations upon the ability to calculate. The economist must go beyond the assumption of "economic man" precisely because of the economic advantage of not behaving like economic man — an advantage that presumably explains why the world is not populated solely by economic men.

I would like also to indicate some of the limitations of the analysis, which suggest a number of directions for generalization:

1. The analysis here does not pretend to explain all the behaviors and attitudes we think of as emotions, but only certain of these which are alleged to help guarantee the execution of contracts. Other types of emotions also serve important functions, for example, fear which makes us flee danger or romantic love which helps us win mates, but are not relevant for the purpose at hand. Coming closer to the topic here, there are still other sentiments that bear upon the kinds of interpersonal transactions a group of individuals can arrive at, but which I have not studied — envy, pride, and shame are among them.

2. I have dealt only with one-time interactions. There is already a considerable literature dealing with repeated interactions, in which refusal to continue a business relationship with a defecting trading partner may alone suffice to enforce a contract. Nevertheless, I suspect the anger/gratitude response pattern is an important additional factor, providing a degree of extra support where mere refusal of future business cannot carry all the weight of maintaining a social relationship.

3. The postulated division of responsibility, whereby the first-mover has free choice of productive arrangements while the second-mover can only react by a reward or deprivation response, is of course a special assumption. But I do not believe it is unduly restrictive. We could equally well imagine the first-mover as making some preliminary choice that narrows down the productive or other options available, after which the second-mover makes a finalizing decision, possibly including productive aspects as well. The assumed special shapes of the functions could also be easily generalized. If the productive opportunity locus were to become less concave-downward (i.e., if QQ' in the diagrams were to approach a linear or even a convex-downward shape), the potential mutual gain via productive coordination would be reduced. Cooperation will then, other things equal, become less likely.[14] And similarly, one could allow for different "exchange rates" in the second-mover's ability to transfer or to deprive the other party, with predictable consequences for the efficacy of second-mover's ability to exert influence upon the other.

4. I considered only interactions between a self-interested coolly-calculating party on the one hand, and an emotion-driven individual on the other, where the latter has the last move. If the time-sequence were

---

[14]See Friedman [1980].

reversed, what I have called "threat" and "promise" would evidently not be successful.

5. Finally, the analysis here hints at a more daring suggestion. Leadership importantly involves the function of conferring reward or punishment upon other members of the society, i.e., serving as second-mover in many social interactions. While the role of leader is often simply seized by the strongest and/or cleverest, there also is commonly some degree of popular consent involved. Might it not be the case that the "inspirational" or "charismatic" quality that we look for in leaders is an extraordinary capacity to transcend self-interested motivations, to be passionately driven by action-dependent, non-utilitarian goals? There are of course merely self-interested, calculating princes who would follow Machiavelli's advice:

> Therefore, a prudent ruler ought not to keep faith when by so doing it would be against his interest, and when the reasons which made him bind himself no longer exist. -- The Prince, XVIII

But such a ruler would likely find it difficult to elicit from citizens that extra measure of devoted cooperation upon which the survival of his regime may depend. So, even the most Machiavellian of princes is likely at least to simulate the possession of genuine action-independent or action-dependent passions -- what Boulding [1969] has called the heroic ethic. Of course, I need hardly add, a nation is not always better off for having heroic leaders.

REFERENCES

Becker, Gary S., "A theory of social interactions," J. POLIT. ECON., v. 82 (Nov./Dec. 1971), pp. 1063-1093.

_____, "Altruism, egoism, and genetic fitness: Economics and sociobiology," J. ECON. LIT., v. 14 (Sept. 1976), pp. 817-828.

Boulding, Kenneth E., Conflict and Defense (New York: Harper & Row, 1962).

_____, "Economics as a moral science," AM. ECON. REV., v. 59 (March 1969).

Coase, R.H., "Adam Smith's view of man," J. LAW & ECON., v. 19 (Oct. 1976).

Cohen, A., Everyman's Talmud, New American ed. (New York: E.P. Dutton & Co., 1949).

Darwin, Charles, The Origin of Species by Means of Natural Selection, 6th ed., 1872 (1st ed., 1859).

_____, The Expression of the Emotions in Man and Animals (1872).

Friedman, David, "Many, few, one: Social harmony and the shrunken choice set," AM. ECON. REV., v. 70 (March 1980).

Hamilton, W.D., "The genetical evolution of social behavior, I," J. THEOR. BIOL., v. 7 (1962), pp. 1-17.

Hirshleifer, J., "Shakespeare vs. Becker on altruism: The importance of having the last word," J. ECON. LIT., v. 15 (June 1977), pp. 500-502.

_____, "Evolutionary models in economics and law: Cooperation versus conflict strategies," RES. IN LAW & ECON., v. 4 (1982), pp. 1-60.

Howard, Nigel, Paradoxes of Rationality (Cambridge, MA: MIT Press, 1971).

Luce, R. Duncan and Howard Raiffa, Games and Decisions (New York: Wiley, 1957).

Macaulay, Stewart, "Non-contractual relations in business," AM. SOCIOL. REV., v. 28 (Feb. 1963), pp. 55-67.

Rose, Michael, The mental arms race amplifier," HUMAN ECOLOGY, v. 8 (1980), pp. 285-293.

Schelling, Thomas C., The Strategy of Conflict (London: Oxford University Press, 1960).

_____, Micromotives and Macrobehavior (New York: Norton, 1978).

_____, "The intimate contest for self-command," THE PUBLIC INTEREST, No. 60 (Sept. 1980).

Smith, Adam, The Theory of Moral Sentiments, E.G. West ed. (Indianapolis: Liberty Classics, 1969).

Staddon, J.E.R., "On a possible relation between cultural transmission and genetical evolution," in P.P.G. Bateson and P.H. Klopfer, eds., Advantages of Diversity, v. 4 (New York: Plenum, 1981), pp. 135-145.

Telser, L.G., "A theory of self-enforcing agreements," J. OF BUS., v. 53 (Jan. 1980), pp. 27-45.

Thompson, Earl A. and Roger L. Faith, "A pure theory of strategic behavior and social institutions," AM. ECON. REV., v. 71 (June 1981), pp. 366-380.

Trivers, Robert L., "The evolution of reciprocal altruism," QUART. REV. BIOL., v. 46 (March 1971), pp. 35-58.

Washburn, "Tools and human evolution," SC. AM., v. 203 (1960), pp. 63-75.

Wilson, Edward O., "Altruism," HARVARD MAG., v. 81 (Nov.-Dec. 1978).

Second is benevolent

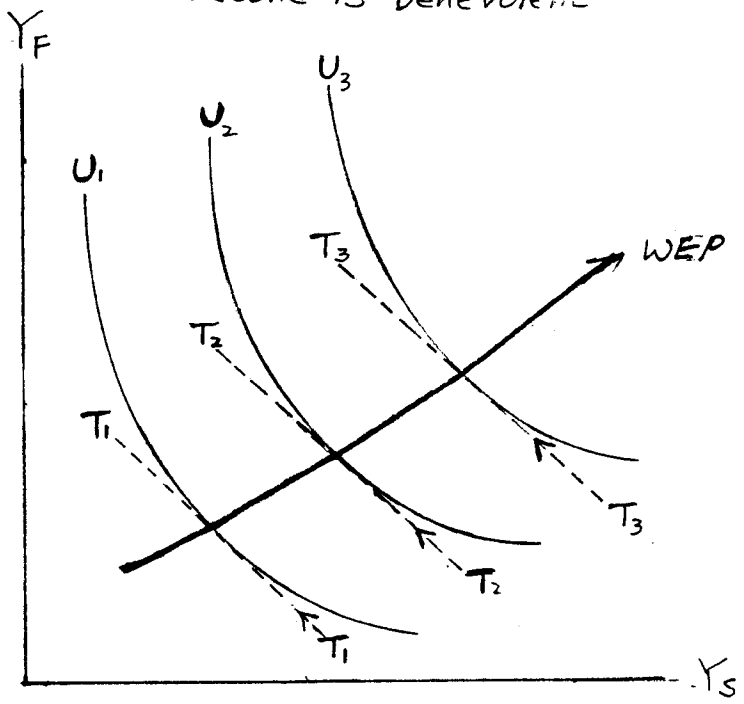

Figure. 1a

Second is malevolent

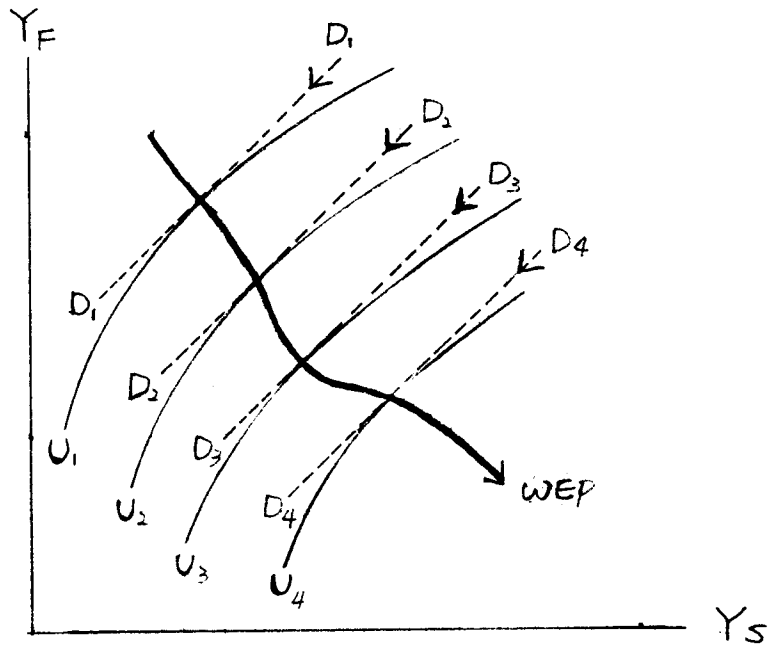

Figure. 1b

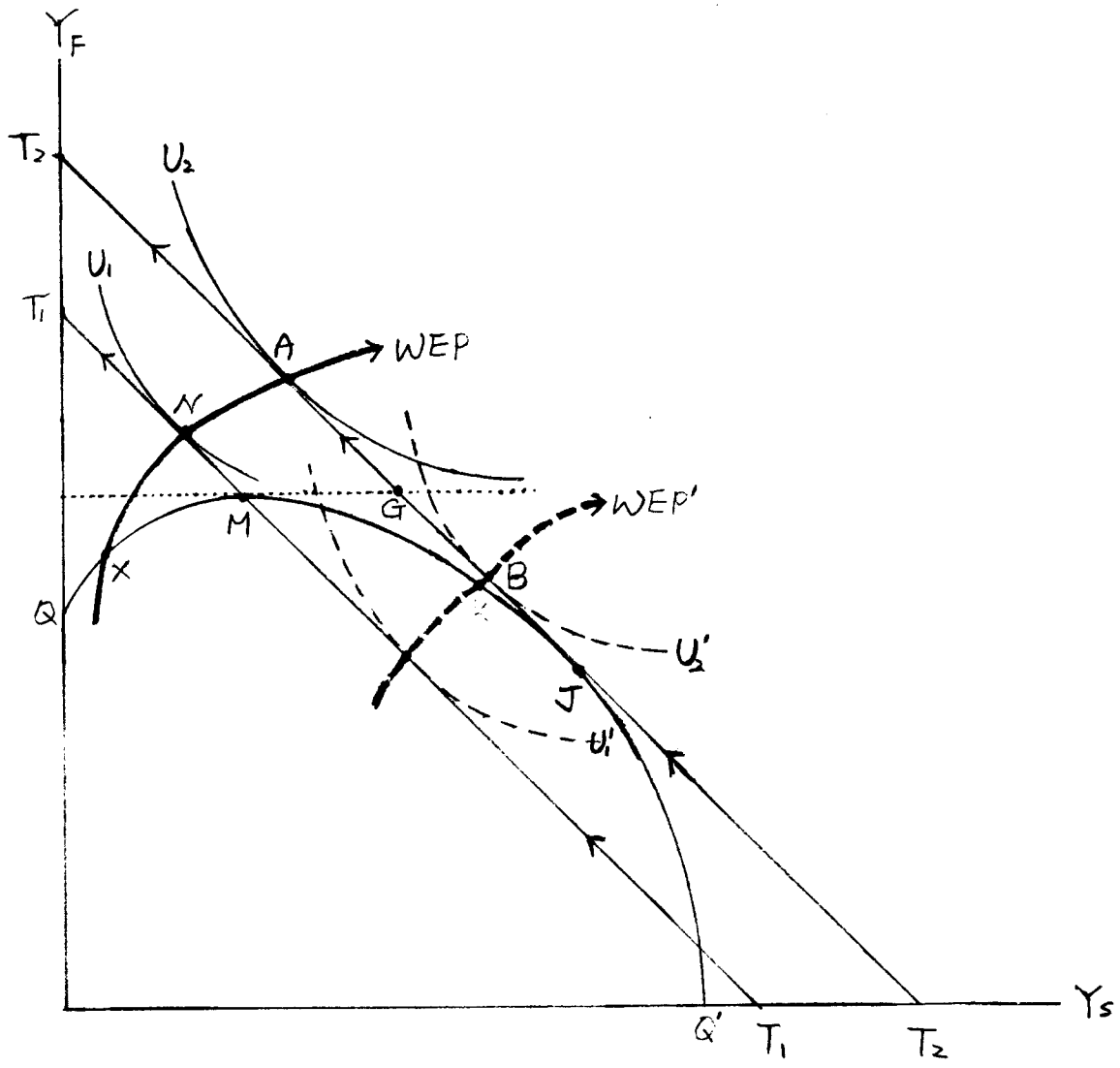

Figure. 2

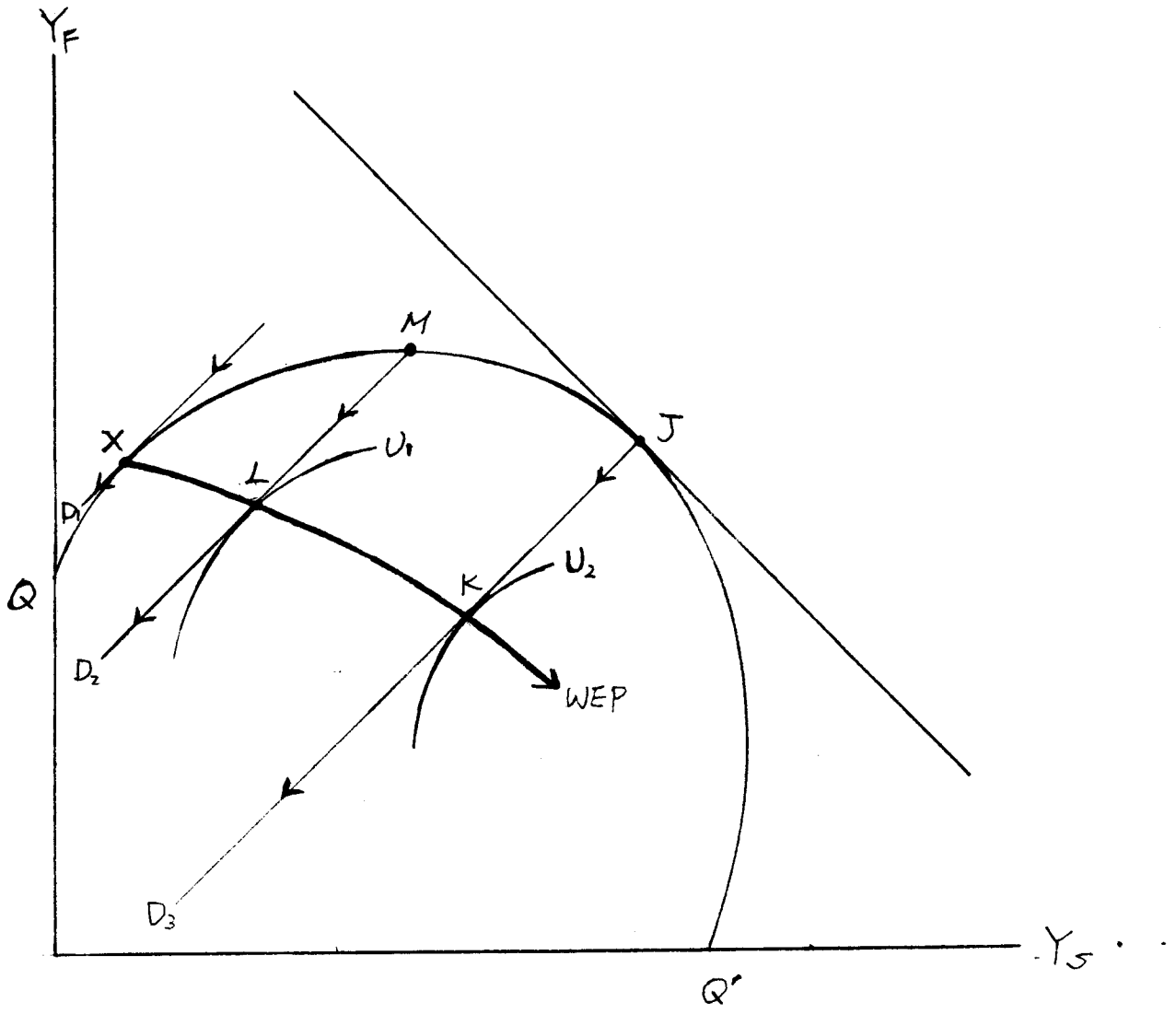

Figure. 3

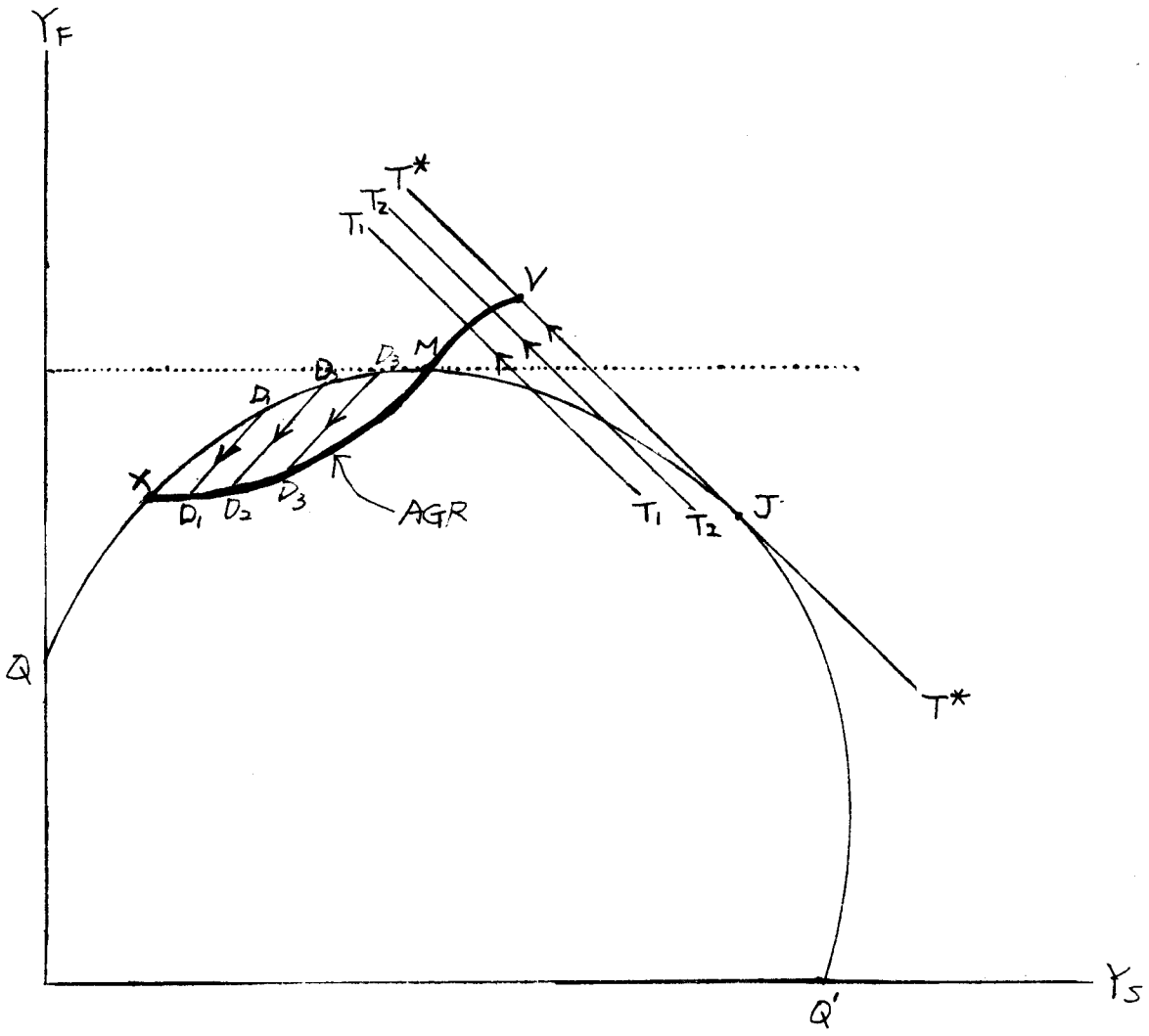

Figure. 4

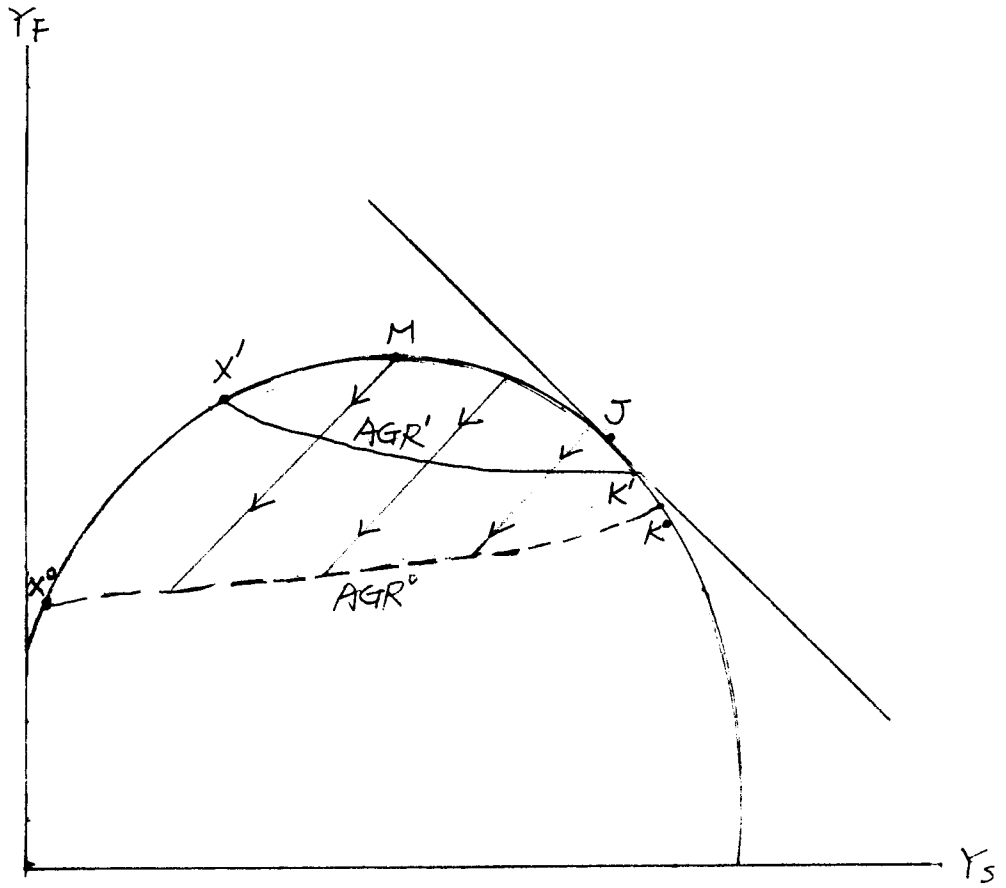

Figure 5

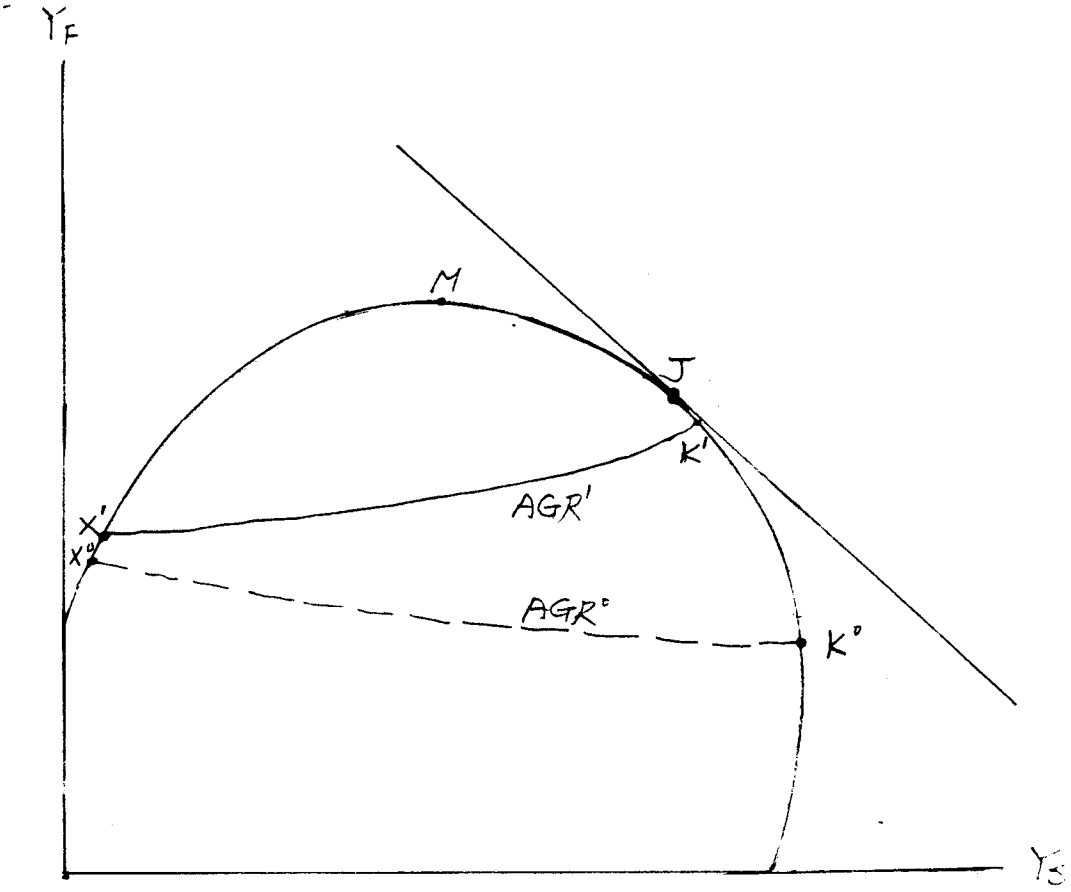

Figure 6