Analyse & Kritik 22/2000 (© Lucius & Lucius, Stuttgart) p. 98–113

Bernd Lahno

In Defense of Moderate Envy

Abstract: In contrast to Axelrod's advice "don't be envious" it is argued that the emotion of envy may enhance cooperation. TIT FOR TAT does exhibit a certain degree of envy. But, it does so in inconsistent ways. Two variants of TIT FOR TAT are introduced and their strategic properties are analyzed. Both generate the very same actual play as TIT FOR TAT in a computer tournament without noise. However, if noise is introduced they display some greater degree of stability. This is due to the fact that they form, in a prisoner's dilemma supergame with suitable parameters, an equilibrium with themselves that is subgame perfect or (in case of the first strategy) close to subgame perfect. It is additionally argued that these strategies are exceptionally clear and comprehensible to others in that they conform to well known real live behavior patterns.

1. Introduction

Robert Axelrod's first behavioral advice for individuals who want to succeed in iterated prisoner's dilemma situations is: "Don't be envious" (110).[1] On closer examination of what Axelrod means by envy, however, the advice offered appears questionable. TIT FOR TAT, the most successful strategy in the computer tournaments conducted by Axelrod, is by no means incompatible with a certain degree of envy. I shall argue that the success of TIT FOR TAT is even in part dependent on this characteristic. I shall suggest a variant of TIT FOR TAT that represents the idea of a player being consistently motivated by moderate envy. This strategy shares the advantages of TIT FOR TAT. At the same time it is more in line with basic human emotions, and it offers greater robustness in dealing with accidental mistakes in the decision process.

Various authors have pointed out an apparent disparity between the assumptions on which Axelrod's model is based and the general human trait of envy (Homans 1985; Campbell 1986; Gowa 1986). In view of our general human disposition towards envy it was in particular argued that prisoner's

[1] All unspecified page citations refer to Axelrod 1984.

dilemma situations take on the form of so-called "Game[s] of Difference" (Taylor 1987) in which the obstacles to cooperation are much more severe.[2] This line of argument seeks to differentiate between objective payoffs and the subjective utilities affecting individual decisions.

From an evolutionary point of view this argument does not seem particularly convincing. In comparison to a standard rational choice approach, the evolutionary perspective has the distinct advantage of being more or less independent of the difficulties that emerge in determining subjective utility functions. Within the evolutionary approach payoffs are a measure of effective evolutionary success, which is objectively determined within a given context. Perceived differences in payoffs may have motivational force but they are not part of the objective payoffs that in the last resort determine evolutionary success.

In the following, I shall treat envy as an emotional disposition that affects individual decision-making. Its 'utility' or 'rationale' is determined by the success of the behavior caused by it, where success itself is measured independently of envy by objectively given payoffs.

1. "Don't be envious"?

To Axelrod, 'envy' is the tendency to compare continuously one's own achievement with the achievement of others. An envious person will only be satisfied if his own fortune at least equals that of his counterpart. He will therefore constantly try to prevent others from gaining a relative advantage over him.[3]

Axelrod justifies his advice not to be envious basically with two different lines of argument. The first is based on the results of his computer tournaments. The most successful strategies of the tournaments were mainly characterized by two properties: they were 'nice' and 'forgiving'. A strategy is called "nice" if following this strategy a player is never the first to defect; it is called "forgiving" if it is not the case that a player following this strategy always defects after his opponent once defected. Strategies characterized by these two features do not aim at gaining a one-sided advantage. Being nice they aim at cooperation for the mutual benefit of both parties. As they are forgiving, such strategies also remain cooperative towards opponents who occasionally try to gain an unfair advantage over others. Since the results of his tournaments show that niceness and forgiveness are prerequisites for success in iterated prisoner's dilemma games under various plausible circumstances,

[2] Gowa 1986. Behr 1981 investigated how the results of computer tournaments changed if the success of a strategy was only measured by how it performed with regard to its respective opponent.

[3] This, of course, is a rather restrictive concept of envy. For a broader account of the emotion of envy see Elster 1999.

Axelrod concludes that it will generally be favorable to seek success jointly with rather than at the expense of the respective partner.

Axelrod's second argument is based on a strategic analysis. One can only be more successful than one's opponent in a prisoner's dilemma situation if one defects. "But defection leads to more defection and to mutual punishment" (111). The attempt to be better than the other party in iterated prisoner's dilemma situations is therefore self-defeating. Individual success is directly dependent on the partner's success under such conditions. If one player achieves more than his opponent does, this indicates for Axelrod only that they both achieved less than they could have. The party who gets more than his opponent is therefore not a truly successful party. To be truly successful one rather has to provide an incentive for the partner to be as cooperative as possible. The fact that success is dependent on a cooperative and not a competitive attitude is also proved by a noteworthy characteristic of the most successful strategy of the tournaments. A player who uses TIT FOR TAT can never attain more points than his direct opponent can.

The case seems to be well argued up to this point. However, taking into account Axelrod's own evolutionary approach a puzzle arises: *If it really is better not to be envious, why are we?*

Successful strategies will increase their population share in the long run, whereas strategies that are not successful will be driven out. The evolutionary process will only come to an (relative) end, if (almost) all strategies are roughly equally successful. Thus, if Axelrod is right, then in view of the long history of the development of human behavior one should expect to encounter instances of serious envy only very rarely. However, as a matter of fact the opposite is true. Moreover, if we assume that a relative equilibrium has been achieved in the evolutionary process, then abstinence from envy will not lead to any improvement for the individual. Still, it might be that our evolutionary development is in the very beginning and a stable equilibrium has yet to be achieved. In this case only Axelrod's advice can claim some plausibility. But even if we grant this, we should not forget—and Axelrod himself stresses the point—that no strategy can be optimal in iterated prisoner's dilemma situations irrespective of the strategies of other parties. Even the renunciation of competition and competitiveness is not advantageous in every case.

There is a more general question lurking in the background here. Envy is an emotion and emotions determine the way in which we perceive the world (de Sousa 1978). They affect our understanding of which aims are to be pursued and the suitable means of so doing. Emotions determine our behavior in that they affect our preferences and our convictions. While they may evolve in specific circumstances, their effects are seldom restricted to specific situations. Therefore for the evolutionary success of an envious disposition the consequences of envy in prisoner's dilemma situations alone are hardly decisive.

We may not be able to shift our behavioral gears arbitrarily and thus not be as free in choosing strategies as is often assumed in evolutionary game theory. With regard to a specific behavioral problem, our choices may be governed by restrictions that are the result of more general dispositions. Therefore we might not be able to discriminate perfectly between situations and to develop behavior patterns that are optimal for every situation. This does not contradict the basic premises of evolutionary theory. For the evolutionary success of those characteristics that determine our behavior in this way is measured by their positive or negative consequences over all.

I can put this general problem safely aside here since I shall argue next that, contrary to what Axelrod says but in line with his results, and basically consistent with Axelrod's use of the term, envy is in fact advantageous for cooperation in iterated prisoner's dilemma games. We should nurture the feeling rather than fight it. According to Axelrod, "[...] envy leads to attempts to rectify any advantage the other player has attained" (111). This is true. Axelrod further argues that this leads to unnecessary and self-destructive defection. This is untrue. If envious feelings subside when the balance has been redressed, envy strengthens cooperation. Only if one party knows that the other party is not willing to accept one-sided exploitation will there be a sufficient incentive to cooperate. In this sense, TIT FOR TAT is also 'envious'—because it can be provoked. However, TIT FOR TAT is not consistently envious!

2. Envy, Fairness and Retribution

I have a character trait that I am loath to admit to—even though I believe that I share it with many other people and that, all in all, possessing it is not without advantage. In some situations I am overcome by a strong emotion with regard to fellow humans that is barely justifiable on grounds of individual rationality. For example, when I drive to work in the morning, the traffic regularly jams at the exit of the three-lane motorway over a distance of several hundred meters, thus seriously impeding progress. However, some motorists also wishing to use the exit are 'cleverer' than I am. They use the central lane to quickly jump the queue and then, when a favorable opportunity arises, join the exit lane. These motorists are the regular object of my negative emotions. I catch myself being at great pains to close the gap between me and the car in front so as to prevent anyone from pushing in. Other motorists behave in a similar way. This often results in a long queue of 'clever' drivers along the central lane, all waiting to be let in with their car indicators flashing. This naturally leads to the traffic also jamming along the central lane and in extreme cases along the entire motorway. Yet the strategy employed by the clever drivers is always successful as they reach their goal more quickly than those who—like me—nicely join the queue in the exit lane. I am powerless to prevent this. But I know that if I could prevent it or if I could cause the 'culprits' some inconvenience without any great effort (e.g. slowing down at the lights to make them stop while I get through), I would. My emotions would certainly encourage me to do so.

My emotional reaction is almost impossible to justify from a rational point of view. While it is true to say that the sum of clever motorists causes the jam along the right-hand lane to get worse and that my progress is impeded—perhaps their joint behavior is even the entire cause of the congestion—yet none of the clever drivers actually individually causes the traffic jam. None could actually do anything to change my negative situation by altering his individual behavior. The disadvantage that I suffer as a result of the single behavior of the driver whom I prevent from joining the exit lane is so minor that I would in other situations be quite prepared to accept a similar disadvantage uncomplainingly, if only out of politeness (e.g. letting someone go through a door first). There is yet more: my reaction is not even geared towards diminishing my harm, its sole purpose is to inconvenience the other driver in a similar manner. I am even prepared to suffer further inconvenience to attain my goal.

My emotional reaction corresponds to what Axelrod calls 'envy'. This feeling results from the realization that the other driver has gained an unfair advantage—an advantage that I have been denied—and aims at redressing the balance. The only thing that counts in this situation is relative advantage and not absolute well-being. As I can do nothing to change my situation, I want all the others to be as badly off too. Whether 'envy' really is the proper word for this emotion is questionable as it is dependent on specific conditions. While I am aware that the individual clever driver is not responsible for my situation, I still consider his behavior to be unfair (because this basic sort of behavior is responsible for my situation). The fact that I think such behavior is unfair is decisive for my emotional reaction. For instance, I react quite differently when an ambulance forces me to stop and I do not experience a comparable feeling when a fast Porsche overtakes me (although I can sometimes be a little envious in another sense when that happens). My reaction is not caused by another person's advantage per se, but only by what I perceive to be an unfair advantage. Envy, by contrast, is often viewed as a negative emotional reaction to another's *absolute* advantage that emerges independent of considerations of fairness.

The fact that my emotion is geared towards redressing the balance may make it safe to assume that it is actually driven by the quest for fairness. However, such a characterization would be far too general. Strangely enough, this feeling only appears when *others* act unfairly. If I myself act in a socially damaging way and profit from doing so, I do not always feel obliged to compensate those that suffer as a result of my behavior. The feeling presented here only demands fairness up to a certain degree, namely to the degree to which every (unfair) advantage of *another person* compared to mine calls for compensation.

Another suggestion to define the emotion would be to call it a desire for revenge or retribution. But this also only applies in part. Retribution is a feeling that is caused by having suffered damage and is directed against the person who caused the damage. But in the present case, there is no *single* person to put the blame on. After all, this is clearly a (negative) feeling of retribution in the following sense: first of all, it is an emotional reaction to a certain behavior that is looked upon as being damaging or bad and secondly, the feeling is directed towards causing some damage to the perpetrator or the perpetrators of the behavior.[4]

The feeling characterized here exhibits some of the characteristics of envy as commonly understood. It is focussing on the other party's relative advantage without considering one's own absolute personal benefit. Just like the desire for fairness, the feeling also aims at redressing the balance. And like retribution, it entails passing judgement on the morality of the other person's actions and aims at harming this other person. In the sense in which Axelrod generally defines 'envy' as an attitude, which is exclusively concerned with the relative advantage of others and not with one's own absolute personal benefit, this is a (mild) form of envy. However now, Axelrod's advice "Don't be envious" should be viewed with caution. His own personal investigations have shown that behavior determined by such feelings of envy can be very successful in situations that resemble iterated prisoner's dilemma games.

3. A Variant of TIT FOR TAT

Imagine a prisoner's dilemma supergame in which the players have a fixed discount parameter ω. The stagegame is symmetrical, the payoffs are indicated in the conventional way by the parameters in the table at Figure 1:

		Player 2	
		C	D
Player 1	C	RR	ST
	D	TS	PP

$T > R > P > S$

[4] Various authors (above all Mackie 1982 and Westermark 1932) have noted the important influence of retribution on our social behaviour as a whole and especially on our behaviour in moral contexts. Axelrod confirms this special influence by highlighting the 'provocability' of successful strategies.

Player 1 understands that reciprocal cooperation leads to a mutually beneficial and (Pareto-) efficient result. He is prepared to contribute towards such a result. Other than that, player 1 will be influenced by feelings of the above-mentioned kind, i.e. he will try to immediately balance out any advantage that a potential opponent may gain. This suggests the following strategy:

> MODERATE ENVY: Defect at time t if and only if your opponent defected more often than you did in the preceding stagegames up to time t.

Like TIT FOR TAT, this strategy is 'nice', i.e. somebody following MODERATE ENVY is never the first to defect. Again like when using TIT FOR TAT, he is never out to gain a one-sided advantage. In a supergame, MODERATE ENVY can end up with a lower but can never achieve a higher payoff than any opponent. MODERATE ENVY is provocable in the sense of Axelrod's proposition 4 (62), but the strategy is, just like TIT FOR TAT, also prepared to return to mutual cooperation after a defection, i.e. it is forgiving (36). MODERATE ENVY really does possess all of TIT FOR TAT's characteristics that Axelrod emphasizes as being favorable, it is, e.g., also maximally discriminating (66).

All this is not very surprising as MODERATE ENVY normally prescribes the same behavior as TIT FOR TAT. The two strategies are indistinguishable during the regular run of a supergame. More precisely:

> If S is any supergame strategy for the iterated prisoner's dilemma, the combination of strategies (S, MODERATE ENVY) follows the same path through the game as the combination (S, TIT FOR TAT).

To prove this, one needs only to recall that, within a finite number of rounds of play, a TIT FOR TAT player never defects more often than his opponent and at most defects one time less. A simple inductive argument shows that after regular play, a TIT FOR TAT player will have defected precisely once less than his opponent in the game up to any point in time t (inclusive) if and only if the opponent defected at time t. The TIT FOR TAT player will therefore make the same decision as a MODERATE ENVY player at every point in time $t + 1(t \geq 0)$. As both strategies prescribe cooperation at $t = 1$, TIT FOR TAT and MODERATE ENVY players will both act the same throughout the entire game with regular play.

As a MODERATE ENVY player consistently behaves in a cooperative manner towards a player adopting the same strategy, it also follows that MODERATE ENVY remains collectively stable (56) under the same conditions as TIT FOR TAT. Thus the following ensues (cf. 59, 218):

The strategy profile (MODERATE ENVY, MODERATE ENVY) is a Nash-equilibrium in the prisoner's dilemma supergame with discount factor ω if and only if:

$$\omega \leq max\left\{\frac{T-R}{T-P}, \frac{T-R}{R-S}\right\}.$$

One could now suspect that MODERATE ENVY is only another name for TIT FOR TAT and that the strategies are in actual fact identical. However, this is not the case.

4. The Reward of Fairness

It should be noted that MODERATE ENVY's behavior is only identical to that of TIT FOR TAT given a *regular* play. If (S_1, S_2) is any strategy profile, play shall be considered 'regular' (in accordance with (S_1, S_2)) up to point in time t if player 1 had consistently made his decisions in accordance with S_1 and player 2 in accordance with S_2 up to the point in time t. The play will therefore be regular as long as neither player deviates from his respective strategy.

In the final chapter of his book, Axelrod points out a weakness of TIT FOR TAT that concerns the rules governing the TIT FOR TAT player's behavior after deviations have occurred. Unfavorable echo effects (176) can ensue if the opponent of a TIT FOR TAT player chooses a similar strategy. For instance, let us assume that TIT FOR TAT player 1 meets TIT FOR TAT player 2 and player 1 deviates from his strategy by defecting at time t. TIT FOR TAT prescribes that player 2 must then defect at time $t + 1$, while player 1 is to cooperate. The roles are reversed at time $t + 2$, player 2 cooperates again whereas player 1 reacts to player 2's defection with a further defection. If neither player deviates from his strategy again during the further run of the game, the players will take turns in exploiting each other for the rest of the game.

Axelrod remarks: "TIT FOR TAT is not forgiving enough." However, this does not seem to properly describe the situation. If the opponent reacts to a wrong move with his defection there is nothing to forgive. In such cases the other's right to reparation or compensation should much rather be honored. The operative word is not therefore forgiveness, but rather a certain degree of fairness. TIT FOR TAT cannot supply this degree of fairness because it reacts mechanically to every defection. A player using MODERATE ENVY will decide differently. He will react to every imbalance that is to his disadvantage. As no such disadvantage is apparent if his opponent defects in order to get even, he will continue to cooperate after an 'inadvertent' defection by himself caused defection by the opponent. Echo effects are prevented by a modicum of fairness.

The difference between the behavioral dispositions described by TIT FOR TAT and MODERATE ENVY strategies only manifests itself when occasional behavioral deviations, mistakes or errors, occur. This is not possible in computer tournaments in which decisions are made according to a reliable mechanism. TIT FOR TAT and MODERATE ENVY strategies are indeed indistinguishable under such conditions. In particular, MODERATE ENVY will perform just like TIT FOR TAT in every computer tournament. But there is still an essential difference. This difference concerns the rationale behind the behavior patterns prescribed by the strategies and the stability of such behavior rules when faced with minor behavioral deviations.

Note that in a prisoner's dilemma supergame at every point in time and after any finite history of the game the players face a future that in itself constitutes a prisoner's dilemma supergame identical to the total game. Every subgame in a prisoner's dilemma supergame is itself a prisoner's dilemma supergame. Let us once again review the situation that ensues after one of two players that both generally use TIT FOR TAT mistakenly defects. For the then following subgame, their strategies prescribe the following behavior: the culprit cooperates in the next game and the opponent defects, both thereafter copying one another's behavior. If one treats the subgame as a separate new game, the culprit adopts TIT FOR TAT in this game and his opponent SUSPICIOUS TIT FOR TAT (Boyd/Lorberbaum 1987). These two strategies do not form an equilibrium point in a prisoner's dilemma supergame in which TIT FOR TAT is collectively stable. Thus, at least one of the two players (and in this case both) could do better by adopting a new strategy should the other player stick to his strategy. TIT FOR TAT does not provide a rationally justifiable behavioral rule in case of occasional mistakes. This forms the crux of the weakness that Axelrod identified with regard to TIT FOR TAT. Instead of returning to mutually beneficial cooperation, the players take turns in exploiting each other after a mistake has been made.

MODERATE ENVY seems better able to accommodate occasional mistakes. If a mistake occurs between two players using MODERATE ENVY, it will be best for the injured party to stick to his strategy provided the opponent immediately goes on to play MODERATE ENVY. Under normal conditions, the same is also true in reverse: he who once mistakenly defected does best to go on playing MODERATE ENVY after the incident. But this, of course, depends on what is to be understood by "normal conditions" in this case. I shall turn to this question now. The investigation will give rise to more general considerations about the (evolutionary) stability of supergame strategies.

5. Subgame Perfectness

Consider a prisoner's dilemma supergame with payoff parameters and a common discount factor such that MODERATE ENVY forms an equilibrium if matched with itself. MODERATE ENVY demands cooperation in exactly those cases when a player has in total defected just as often or more often than his opponent in the stage-games preceding the decision. With regard to an application of MODERATE ENVY, any possible situation within a prisoner's dilemma supergame is therefore sufficiently determined for each player by his defection balance. n_i shall determine this balance for a player i. If d_i is the number of his defections in the total number of the preceding stagegames and d_j is the respective number of the opponent, n_i is determined by $n_i = d_i - d_j$. Using the parameter n_i the MODERATE ENVY strategy can also be defined as follows for player i:

> Cooperate if and only if $0 \leq n_i$ applies for parameter n_i as determined by the history of the game so far.

Let us now assume that the balance is strictly positive for a player in any given situation. MODERATE ENVY then requires him to cooperate. If he plays against another player also using MODERATE ENVY he must accept one-sided defections until balance has been attained. Following MODERATE ENVY his expected utility is:

$$E = \frac{1-\omega^{n_i}}{1-\omega}S + \frac{\omega^{n_i}}{1-\omega}R.$$

The only real alternative to MODERATE ENVY is consistent defection. One can easily calculate that a shift to ALL D is unprofitable in situations where the following applies:

$$\omega^{n_i} \geq \frac{P-S}{R-S} \qquad (1)$$

If both parties of a supergame assume that their co-player adopts MODERATE ENVY, then in any situation during the supergame, as long as condition (1) is fulfilled, it is not only best for the injured party to stick to the MODERATE ENVY strategy but also for the one who has profited from afflicting the injury on previous rounds of play.

For example: let the payoff parameters be chosen in accordance with the values of the computer tournament: $T = 5, R = 3, P = 1, S = 0$, and the discount factor amount to $\omega = 0.9$. In this case, (1) is fulfilled for all $n_i \leq 10$. Thus, as long as neither of the players has defected more than ten times in excess of his co-player's defections during the entire course of the preceding game up to a point in time t, it remains favorable for both players to follow

MODERATE ENVY at t. Provided that the respective opponent sticks to MODERATE ENVY, the strategy MODERATE ENVY promises each player the highest possible payoff.

Now, a strategy generally determines a decision for every conceivable situation—including those that cannot occur if the strategy is followed properly. Every supergame strategy therefore induces a strategy on each of the subgames of the supergame. A strategy profile that induces a Nash equilibrium on every subgame of a game is called a subgame perfect equilibrium (Selten 1975).

The criterion of subgame perfectness demands that the equilibrium strategies provide rationally comprehensible orientation for every conceivable development. It results from a consistent application of the principle of strictly future-oriented rationality. Thus, a subgame perfect equilibrium excludes any threat whose actual execution would demand decisions that would be rationally unjustifiable with regard to the expected consequences. What part can such a criterion play in an evolutionary context as considered by Axelrod?

The above considerations provide an answer to this question. Under realistic conditions one must assume that occasional deviations and mistakes will occur even given a relatively stable state in which (almost) all members of a population follow a collectively stable strategy. It is, in fact, only on the basis of such an assumption that some relative degree of stability seems attainable when applying TIT FOR TAT. There is no provocation within a population that consistently and exclusively applies the TIT FOR TAT rule. The provocability of a strategy therefore loses its function. Unconditional cooperation leads to the same result under such conditions as following TIT FOR TAT. But unconditional cooperation can be implemented with less effort. If on the other hand enough players forego their provocability in favor of a more cost-effective alternative, the population becomes prone to fall prey to intrusive defectors—even though TIT FOR TAT is collectively stable in the abstract formal sense.[5]

But if some deviations are to be expected even within a relatively homogeneous population, evolutionary pressure will be put on a strategy that lacks a mechanism to cope optimally with such deviations. The strategy is likely to be superseded by variants that are better equipped to deal with such irregularities. If players do make mistakes, they will not only make one-off mistakes in a supergame. Occasionally players may deviate several times from their strategy during the course of one supergame. A truly stable strategy must also provide optimum answers for such cases. However, there is no general answer to how many and which mistakes are to be expected and in how far such

[5] This type of instability is possible because 'collective stability' is a relatively weak condition and in particular differs from 'evolutionary stability' as defined by Maynard Smith. But, note that there is no evolutionarily stable cooperative strategy in the strong sense in a prisoner's dilemma supergame (Boyd/Lorberbaum 1987; Lorberbaum 1994).

mistakes really can be evolutionarily effective. Strategies that form a subgame perfect equilibrium with themselves are characterized by providing optimum behavior rules to cover any possible situation in a supergame. Strategies of this kind are therefore particularly stable with regard to occasional deviant decisions.[6]

Various authors have discussed the problem of stability in cases of occasional errors.[7] Wu and Axelrod (1995) tested the robustness of several strategies for the prisoner's dilemma supergame, some of which were especially designed to cope with this problem, in a computer tournament 'with noise'. In such a tournament, 'mistakes' are generated by an external random mechanism. For any intended choice there was a 1% chance in the tournament that the opposite choice would in fact be implemented. In an ecological simulation, the most successful strategy turned out to be another variant of TIT FOR TAT, which was first introduced by Robert Sugden (1986). Axelrod and Wu call this strategy CONTRITE TIT FOR TAT. Although it is structurally very simple (being, like TIT FOR TAT, a simple finite automaton), it is not easily described. Assume there are two basic states a player may be in: he is either 'in good standing' or he is not. If (and only if) both players are in good standing and one defects unilaterally, this one looses his good standing. He will regain his good standing as soon as he cooperates once (no matter what his opponent does). Both players start out in good standing. CONTRITE TIT FOR TAT then prescribes: Defect if and only if your opponent is not in good standing.

Like MODERATE ENVY, CONTRITE TIT FOR TAT generates the same actual course of action as TIT FOR TAT if no mistakes occur. So we have the same basic condition of collective stability (in the abstract sense given by Axelrod). If occasional mistakes do occur, CONTRITE TIT FOR TAT reacts quite similarly to MODERATE ENVY;[8] it allows for some compensation in cases of unilateral defection. The main difference between the two strategies lies in their rigor in respect to compensation and fairness. No matter what has taken place before, whether one or two unilateral defections or two hundred have occurred, a player regains his good standing by simply cooperating once, and he does so even if his opponent cooperates as well. So, CONTRITE TIT FOR TAT demands, under all circumstances, only one single act of cooperation for complete compensation.

It is difficult to associate CONTRITE TIT FOR TAT with a familiar

[6] The concept of subgame perfectness may be understood as a transferral of the idea of 'perfect equilibrium' to dynamic games. Perfect equilibrii are those that remain stable when faced with minor deviations, i.e. trembles. Cf. Selten 1975.

[7] Wu and Axelrod 1995 provide a survey of the existing basic approaches to coping with the problem; see this source for additional references.

[8] If no two mistakes occur in two immediately succeeding games, both strategies will generate the very same course of the game no matter what strategy the partner chooses.

emotional reaction or with a commonly known rule of conduct. But, it has a remarkable structural advantage: for suitable parameters, (CONTRITE TIT FOR TAT, CONTRITE TIT FOR TAT) is a subgame perfect equilibrium. To be precise, it is a subgame perfect equilibrium if and only if[9]

$$\omega \leq max\left\{\frac{T-R}{T-P}, \frac{T-R}{R-S}, \frac{P-S}{R-S}\right\}.$$

This condition is only slightly stronger than the corresponding condition for Nash equilibrium. So, in most relevant cases (that is if P and R differ sufficiently and the sucker's payoff S is not exceedingly low) if CONTRITE TIT FOR TAT is collectively stable in the sense of Axelrod, (CONTRITE TIT FOR TAT, CONTRITE TIT FOR TAT) is a subgame perfect equilibrium. This may explain its success in the computer tournament with noise.

The main concern here is with the strategic consequences of the emotion of envy. As the example above shows MODERATE ENVY is often better able to deal with deviant decisions than TIT FOR TAT. However, in contrast to CONTRITE TIT FOR TAT (MODERATE ENVY, MODERATE ENVY) is never a subgame-perfect equilibrium. This is because MODERATE ENVY demands an excessive degree of fairness and compensation. There are always n_i such that condition (1) is not fulfilled. After a player deviating from MODERATE ENVY defects sufficiently often, a return to MODERATE ENVY may force him to provide compensation that cannot be offset by the prospect of future cooperative profit. In such situations—and only in such situations—MODERATE ENVY prescribes rationally unjustifiable behavior.

In most cases, this will constitute no more than a minor weakness as it affects behavior in subgames that are quite far removed from the equilibrium path. Apart from this, the weakness can easily be removed by a slight amendment of MODERATE ENVY. MODERATE ENVY is to be modified in as far as a player who finds himself in such a hopeless situation generally is to defect. To be precise: Let n be the largest possible integer such that condition (1) holds for all $n_i \leq n$.[10] A player i who follows the modified strategy will act in accordance with the following rule:

SOPHISTICATED ENVY: Cooperate at time t if and only if $0 \leq n_i \leq n$ applies to the parameter n_i as determined by the history of the game so far.

As long as deviations are excluded, following SOPHISTICATED ENVY amounts to the same as following MODERATE ENVY (and TIT FOR TAT).

[9] Two different types of states of the world may occur during the course of the game: Both players are in good standing or just one player is. If both players are in good standing, CONTRITE TIT FOR TAT is the best answer to itself iff $\omega \leq max\left\{\frac{T-R}{T-P}, \frac{T-R}{R-S}\right\}$. If only one player is in good standing, his opponent's optimum answer to CONTRITE TIT FOR TAT is CONTRITE TIT FOR TAT iff $\omega \geq \frac{P-S}{R-S}$

[10] Using the INTEGER function n may be defined as:
$n := INTEGER\left(\frac{log(P-S)-log(R-S)}{logw}\right)$

SOPHISTICATED ENVY is nice, provocable, maximally discriminating and forgiving. Whenever TIT FOR TAT is a collectively stable strategy, SOPHISTICATED ENVY is also collectively stable and additionally forms a subgame-perfect equilibrium with itself:

> The strategy profile (SOPHISTICATED ENVY, SOPHISTICATED ENVY) is a Nash equilibrium in a prisoner's dilemma supergame with discount factor ω if and only if:
>
> $$\omega \leq max\left\{\frac{T-R}{T-P}, \frac{T-R}{R-S}\right\}.$$
>
> If it is a Nash equilibrium, it is also a subgame perfect equilibrium.

A player following SOPHISTICATED ENVY behaves according to the following rules:

- Be prepared to cooperate and do not seek to gain an advantage over your opponent!
- Try to balance out any advantage of your opponent!
- Fulfil your opponent's right to compensation for disadvantages if this good-will is suitably rewarded by the prospect of reciprocal cooperation.

These seem to be the rules according to which a moderately selfish, sensible person will behave; rules which are by no means foreign to us.

6. Clarity and Complexity

Axelrod praises another characteristic of TIT FOR TAT: "it has great clarity: it is eminently comprehensible to the other player" (122). In the case at hand, being clear and comprehensible evidently does not mean being actually identifiable. As long as mistakes are excluded, TIT FOR TAT is indistinguishable from MODERATE ENVY and SOPHISTICATED ENVY. If mistakes occur more often it is difficult to see how strategies can at all be clearly identifiable since deviant moves cannot be immediately identified as such. For a strategy, being comprehensible can only mean that it results in simple, recognizable behavior patterns. A strategy that is clear in this sense enables the opponent to develop reliable expectations about the future run of the game. Such a characteristic is always favorable when some common interests exist and cooperation in the pursuit of those would be mutually beneficial.

TIT FOR TAT is indeed comprehensible in this sense. But so are MODERATE ENVY and SOPHISTICATED ENVY each in their own way. It should be noted that regularity is always easily identifiable if it corresponds to known patterns that conform to the respective context. To my mind, this especially applies to the two strategies of envy presented above, or rather to the behavior patterns induced by them. A player who acts in accordance with them will be 'understood' because his behavior will be interpretable on the basis of our everyday theories about the usual motives governing our behavior and our emotional reactions. TIT FOR TAT, by contrast, appears comparatively mechanical and inhuman (and so does CONTRITE TIT FOR TAT).

TIT FOR TAT is actually an extremely simple strategy that makes few demands on the ability of an organism to react to its environment in an appropriate and differentiated manner. Like with CONTRITE TIT FOR TAT, an extremely simple finite automaton can select decisions in accordance with TIT FOR TAT. The decisive point here is that TIT FOR TAT does not demand any complex memory ability. The only information needed to make a decision concerns the opponent's last move.[11] Contrary to that, in order to use MODERATE ENVY or SOPHISTICATED ENVY, one always has to keep in mind a characteristic of the entire game history. The required information is simple and can be expressed by an integer subsuming the experience gained during numerous interactions. Still, if we take the assumptions of the supergame seriously, namely that we are dealing with a randomly large and unforeseeable number of similar interactions, it becomes clear that no finite automaton can replace one of these strategies.[12] Our lives, however, are not infinite and people are not automatons. To the extent that it is necessary to deal with real situations, humans have the necessary skills to summarize and store information. We actually use these skills every day. When measured against our usual way of dealing with a complex and constantly changing environment, these are indeed minor requirements. MODERATE ENVY and SOPHISTICATED ENVY are simple because they conform to known human behavior patterns—and they are even simpler than TIT FOR TAT in this respect.

So this is my advice with regard to how to behave in iterated prisoner's dilemma situations: be cooperative, but also be moderately envious! Take great care to let no one gain a one-sided advantage. If this should still happen, insist on compensation. But also respect the right of your partner to compensation in the reverse case.

Naturally this is not meant as serious advice. With regard to the original computer tournament, it would be far too late anyway, and, moreover, it does not constitute a real alternative for TIT FOR TAT in this context. With regard to real interactions, people do not seem to need my advice. The advice

[11] Strategies with such a memory restriction are called *reactive* (Nowak/Sigmund 1992).

[12] The strategies can at least be generated with Turing machines.

much rather mirrors actual human behavior. To a great extend it was Robert Axelrod, who taught us to view and understand human behavior in this way.

Bibliography

Axelrod, R. (1984), *The Evolution of Cooperation*, New York

Behr, R. L. (1981), Nice Guys Finish Last—Sometimes, in: *Journal of Conflict Resolution 25 (2)*, 289–300

Boyd, R./J. P. Lorberbaum (1987), No Pure Strategy is Evolutionary Stable in the Repeated prisoner's Dilemma Game, in: *Nature 327*, 58–59

Cambell, D. T. (1986): The Agenda Beyond Axelrod's "The Evolution of Cooperation", in: *Political Psychology 7 (4)*, 793–796

de Sousa, R. (1987), *The Rationality of Emotion*, Cambridge-London

Elster, J. (1999), *Alchemies of the Mind. Rationality and the Emotions*, Cambridge-New York

Gowa, J. (1986), Anarchy, Egoism, and Third Images: The Evolution of Cooperation and International Relations, in: *International Organization 40*, 167–186

Homans, G. C. (1985), "The Evolution of Cooperation" by Robert Axelrod, in: *Theory and Society 14*, 893–897

Lorberbaum, J. (1994), No Strategy is Evolutionary Stable in the Repeated Prisoner's Dilemma, in: *Journal of Theoretical Biology 168*, 117–130

Mackie, J. L. (1985), Morality and the Retributive Emotions, in: J. L. Mackie, *Persons and Values*, Oxford, 206–219

Nowak, M./K. Sigmund (1992), TIT FOR TAT in Heterogeneous Populations, in: *Nature 355*, 250–253

Selten, R. (1975), Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games, in: *International Journal of Game Theory 4*, 25–55

Sugden, R. (1986), *The Economics of Rights, Co-operation and Welfare*, Oxford

Taylor, M. (1987), *The Possibility of Cooperation*, Cambridge

Westermarck, E. (1932), *Ethical Relativity*, London

Wu, J./R. Axelrod (1995), How to Cope with Noise in the Iterated Prisoner's Dilemma, in: *Journal of Conflict Resolution 39*, 183–189