Robert Axelrod, "The Emergence of Cooperation among Egoists" APSR 75 (1994): 306-318

SUMMARY/ABSTRACT

Puzzle: What are the conditions under which cooperation will emerge in a world of egoists without central authority. Methods: The tournament approach, the ecological approach and the evolutionary approach (most general). Conclusions: 1) No better strategy than TFT exists under certain conditions 2) necessary and sufficient conditions for a strategy to be collectively stable are identified 3) cooperation can emerge from a small cluster of individuals even if everyone else is unconditionally defecting.

SET UP AND THE BEST STRATEGY

Axelrod claims that his approach is superior, because *all* possible strategies are taken into account and that the mechanism of move from non-cooperation to cooperation is also specified in addition to the equilibrium conditions. He then reviews the set-up and the logic of the Prisoners' Dilemma, which I will not reiterate here. One interesting comment by Axelrod is the R>(T+S)/2 rule. The rule is to insure that an even chance of exploitation (T payoff) or being exploited (S payoff) is not as good as an outcome of mutual cooperation (R payoff). Problems with resolving PD: 1) No mechanism available to the players to make enforceable threats or commitments 2) No way to be sure what the other player will do on a given move 3) No way to change the other player's utilities. Two things to be specified: 1) constant discount rate (w) per move: the smaller w is, the less important later moves are relative to earlier ones 2) a *strategy* is a function from the history of the game so far into a probability of cooperation on the next move (n.b. this is different from the usual game theoretical definition of a strategy). Axelrod argues that identification of the "best strategy" is dependent on the environment and not possible a priori, but by Theorem 1, if w is sufficiently high, there is no best strategy independent of the strategy used by the other player. (e.g. Congress, institutionalization (low turn-over rate) leads to greater probability of future interaction among members, which leads to greater cooperation based on norms of reciprocity.)

THE THREE APPROACHES

The Tournament approach: TIT FOR TAT (TFT) submitted by Anatol Rapoport won the two rounds of round robin strategy tournament organized by Axelrod. TFT scored the most number of points over 200 rounds. The Ecological approach: Axelrod set the submitted strategies against one another over time. TFT displaced other strategies and emerged as "fixation." The Evolutionary approach: What are the characteristics of the strategy that is stable in the long run? "Invasion": equivalent to the single mutant individual being able to do better than the population average. So, a strategy is *collectively stable* if no strategy can invade it. (i.e. this strategy is in Nash equilibrium.) Because specifying all strategies and evaluating their collective stability is difficult, Axelrod approaches the problem through the "backdoor"—by arriving at the three conclusions summarized above.

THE THREE CONCLUSIONS

TFT as a collectively stable strategy: TFT can only avoid being invadable by such a rule if the game is likely to last long enough for the retaliation to counteract the temptation to defect. No rule can invade TFT if the discount parameter, w, is sufficiently large (Theorem 2). Accordingly, if the other player is unlikely to be around for long, the perceived value of w falls and TFT is no longer stable (e.g. failing business, congressman likely to be defeated, etc.). On the other hand, to prevent cooperation, one should keep the same individuals from interacting too regularly with each other.

The characterization of collectively stable strategies: B has a *secure position* over A on move n if no matter what A does from move n onwards, $V(A|B) < V(B|B)$, assuming that B defects from move n onwards. So B can prevent invasion by A. ➔ If you want to employ a collectively stable strategy, you should only cooperate when you can afford an exploitation by the other side and still retain your secure position (Theorem 3). Two consequences: 1) a strategy has the flexibility to either cooperate or defect and still be collectively stable, as long as the other player has not accumulated too great a score. 2) a nice rule (one which never defects first) has the most flexibility, since it has the highest possible score when playing against an identical strategy. But a nice strategy must be provoked (exploited) by the very first defection of the other player (Theorem 4). Moreover, any rule, which may be the first to cooperate, is collectively stable only when w is sufficiently large (Theorem 5). Lastly, the strategy of ALL D(efect) is always collectively stable (Theorem 6).

The implications of clustering: A strategy, ALL D, may be invaded by a cluster (of TFT). If p is the proportion of an individual invader interacts with another invader to another strategy, a p-cluster of A invades B if $pV(A|A) + (1-p)V(A|B) > V(B|B)$. Axelrod shows that for a p any greater than .05, TFT can invade ALL D. More generally, the strategies that can invade ALL D in a cluster with the smallest value of p are those which are maximally discriminating, such as TFT (Theorem 7). Finally, if a nice strategy cannot be invaded by a single individual, it cannot be invaded by any cluster of individuals either (Theorem 8). Once cooperation is established, it remains stable!!!

CONCLUSION

Even when everyone is playing ALL D, a small group of discriminating individuals who have some probability of interacting with one another can eventually cooperation for the whole group, which can then be sustained. The invaders must have a nice strategy (never defecting first) that is provokable.