Momentum

Volume 1 | Issue 1 Article 17

4-18-2012

The Evolution of Altruism

Eithan Graber
egraber@sas.upenn.edu

The Evolution of Altruism

Abstract
In *The Evolution of Cooperation*, Axelrod and Hamilton (A&H) provide a game theoretic approach to the evolution of reciprocal altruism (cooperation). They argue that because Tit-For-Tat (TFT) is robust, stable, and initially viable, it can be used to explain how reciprocal altruism evolved in nature. There are three important issues regarding A&H's approach to which the author responds. First, Nowak and Sigmund challenge the robustness of TFT by arguing that a strategy of Win-Stay, Lose-Shift (WSLS) can outperform TFT. Second, although A&H use clustering to account for the initial viability of TFT, they owe an explanation of how a cluster of TFTs could arise in the first place. Finally, A&H account for the stability of TFT by assuming, without empirical or theoretical support, that the individuals who interact have a sufficiently large probability of meeting again. The author will respond to these three issues in order to improve and extend A&H's approach to the evolution of cooperation.

Eithan Graber

The Evolution of Altruism

In *The Evolution of Cooperation,* Axelrod and Hamilton (A&H) provide a game theoretic approach to the evolution of reciprocal altruism (cooperation). They argue that because Tit-For-Tat (TFT) is robust, stable, and initially viable, it can be used to explain how reciprocal altruism evolved in nature. There are three important issues regarding A&H's approach to which I will be responding. First, Nowak and Sigmund challenge the robustness of TFT by arguing that a strategy of Win-Stay, Lose-Shift (WSLS) can outperform TFT. Second, although A&H use clustering to account for the initial viability of TFT, they owe an explanation of how a cluster of TFTs could arise in the first place. Finally, A&H account for the stability of TFT by assuming, *without* empirical or theoretical support, that the individuals who interact have a sufficiently large probability of meeting again. I will respond to these three issues in order to improve and extend A&H's approach to the evolution of cooperation.

Axelrod and Hamilton begin their paper by claiming that many of the benefits which living organisms seek are disproportionally available to cooperating groups.[1] The problem, they claim, is that although one organism can benefit from cooperating with another organism, each individual can do even better by not cooperating and exploiting the cooperation of others. A&H use the Prisoner's Dilemma as a way of formalizing the "strategic possibilities inherent in [these] situations."[2] In the Prisoner's Dilemma game, two individuals play each other and can either cooperate or defect. The payoff to the players is measured in terms of fitness gain. If both players defect, they both receive 1 unit of fitness (the Punishment (P) payoff). If Player 1 cooperates and

1" Axelrod, Robert, and William Hamilton. "The Evolution of Cooperation." Science. 211.4484 (1981): 1390-96. Print. Pg. 1391.

2" Ibid.

Player 2 defects, Player 1 gets 0 units (the Sucker (S) payoff) and Player 2 gets 5 units of fitness (the Temptation (T) payoff). Since the game is symmetric, if Player 1 defects and Player 2 cooperates, Player 1 will get the T payoff and Player 2 will get the S payoff. Lastly, if both players cooperate, they will both receive 3 units of fitness (the Reward (R) payoff).

One of the authors, Robert Axelrod, conducted a computer tournament of the Prisoner's Dilemma game. 15 strategies were submitted by professors, professionals, and hobbyists from around the world. The game lasted 200 moves and the 15 strategies were paired in a round-robin tournament.[3] Although some complex strategies were submitted, TFT attained the highest average score. TFT is a strategy of reciprocal cooperation; it cooperates on the first move and then does what the other player did the previous round for all remaining rounds. According to A&H, TFT is capable of beating the other strategies because it is never the first to defect, it can retaliate immediately, and it is forgiving (it cooperates after just one act of cooperation by the other player).[i] From these results, A&H conclude that TFT is robust; that is, TFT thrives in an environment with numerous kinds of strategies.[4]

In their 1993 paper, Nowak and Sigmund challenge the robustness of TFT. They suggest that "a strategy of win-stay, lose-shift outperforms tit for tat in the Prisoner's Dilemma."[5] Win-Stay, Lose-Shift (WSLS) is a relatively simple strategy: it repeats its move on the previous round if it obtained the R or T payoff, but switches moves if it only received the P or S payoff.[ii] Nowak and Sigmund demonstrate that WSLS is more effective than TFT against an All-Out-Cooperator (AC). If TFT plays an AC, both players will receive the R payoff every round. This will occur

3" Axelrod and Hamilton, 1393.

4" Ibid.

5 Nowak, Martin, and Karl Sigmund. "A strategy of win-stay, lose-shift that outperforms tit-for-that in the Prisoner's Dilemma game." *Nature*. 364. (1993): 56-8. Print. Pg. 56.

because TFT will cooperate the first round and follow AC into cooperating all subsequent rounds. On the other hand, if WSLS plays AC, once WSLS defects and AC cooperates, WSLS will defect in all subsequent rounds to obtain the T payoff. In this way, WSLS is able to exploit the cooperation of AC. Hence, against an AC, WSLS has an advantage over TFT.

Notwithstanding, TFT has two significant advantages over WSLS. First, if we consider these strategies from the baseline condition of All-Out Defection (AD), which is what existed in nature before any cooperative strategies evolved, TFT is more viable than WSLS. If TFT plays AD, TFT will get the S payoff the first round and the P payoff in all subsequent rounds. Assuming that that TFT plays AD six times, TFT will get a payoff of 5 points.[iii] If WSLS plays AD, however, WSLS will oscillate between the S payoff and the P payoff. If WSLS plays AD six times, WSLS will only obtain a payoff of 3 points.[iv] Hence, against an AD, TFT will get a higher payoff than WSLS after a minimum of 4 rounds.[v] Since defection was the strategy that existed in nature before cooperative strategies evolved, TFT's advantage against AD makes it more initially viable than WSLS. Additionally, WSLS would only have an advantage over TFT if, in addition to TFT and WSLS evolving, an AC also evolved to play the other strategies. TFT's success does not require the emergence of an AC, whereas WSLS' success does. Therefore, although WSLS has a theoretical advantage over TFT, the constraints of natural conditions make it such that the winning situation for WSLS would be extremely unlikely to arise.

In the last two paragraphs, I have argued that even if WSLS does better than TFT against an AC, TFT has two significant advantages over WSLS. Hence, TFT remains the more robust strategy. I will now address an issue regarding the initial viability of TFT. A&H state the problem of initial viability for cooperation as follows: "Even if a strategy is robust and stable, how can it ever get a foothold in an environment that is predominantly noncooperative?"[6] A&H argue that

6" Axelrod and Hamilton, 1393.

clustering and genetic kinship theory are two ways in which cooperation could have evolved in a population of defectors. Although genetic kinship theory is interesting and compelling, I have chosen to focus on clustering for the sake of simplicity and clarity.[vi] A cluster may be informally defined as a number of individuals who employ the same strategy and constantly play each other. More formally, there are two necessary conditions for a cluster to exist. First, a high proportion (*p*) of the interactions of individuals using a strategy must be with others using the same strategy.[vii] Second, two individuals who play each other must have a large probability (*w*) of interacting again.

A&H claim that when *p* and *w* are sufficiently large, a cluster of TFTs can become initially viable in an environment composed overwhelmingly of defectors. To understand this claim, it is helpful to analyze the payoffs of the Prisoner's Dilemma under these conditions. In an environment entirely composed of defectors, all individuals will receive a payoff of 1**n*, where *n* stands for the number of interactions. Assuming *n* = 100, each defector will obtain a payoff of 100. Now, consider how a cluster of TFTs will fare in such an environment. Assuming that *n* = 100, a TFT will theoretically play other TFTs 70 times (*p* is large) and two defectors fifteen times each (*w* is large). A TFT will thus obtain a payoff of (70*3) + (14*2) = 238. As these simple but illuminating calculations show, a cluster of TFTs will do much better than the defectors in the population (238 > 100).

Axelrod and Hamilton convincingly show that a cluster of TFTs can thrive in an environment composed overwhelmingly of defectors. However, they do not explain how a cluster of TFTs could arise in the first place. This is an important issue regarding the initial viability of TFT. I will defend A&H's approach by providing two ways in which a cluster of TFTs could arise in a population of defectors. First, a cluster of TFTs could simply come about

by chance. In particular, this could occur if defectors who lived near each other had offspring that mutated into TFTs at the same time. The problem with this response is that it relies heavily on chance. Even considering the length of evolutionary history, the likelihood of a number of organisms mutating into a specific strategy in the same place and at the same time seems fairly low. Hence, claiming that a cluster of TFTs evolved purely by chance is not entirely convincing.

There is another explanation for the evolution of TFT clusters which is more plausible. In particular, a cluster of TFTs could evolve if, in a small group of defectors, a few defectors mutate into TFTs over time. Unlike the first explanation, this account does not require that several defectors mutate into TFTs at the same time. Instead, all it requires is that defectors mutate into TFTs in a nearby area over an extended period of time. It is important, however, that the mutants belong to a small group of defectors. In a large group of defectors, individuals are not likely to play each other repeatedly. Hence, in a large group, a defector who mutates into a TFT would not be very successful; it would constantly play distinct defectors and hence receive mostly S-payoffs (0). In a small group, on the other hand, individuals are more likely to play each other repeatedly. Therefore, a defector who mutates into a TFT would do much better in a small group. Although it would receive the S-payoff (0) in the first encounter, it would receive the P-payoff (1) in all subsequent encounters with the same individual. Hence, if a TFT is playing a small number of defectors repeatedly, it would obtain a payoff that is close to the defector's payoff.[viii] By doing so, TFTs could survive, reproduce, and form clusters over time.

So far, I have defended the initial viability of TFT by arguing that there are two ways in which clusters of TFTs could arise. Moreover, I have defended the robustness of TFT by arguing that TFT has two significant advantages over a strategy of Win-Stay, Lose-Shift. Still, the stability of TFT remains unsupported. A&H state the problem of stability for cooperation as

follows: "Under what conditions can such a strategy, once fully established, resist invasion by mutant strategies?"[7] TFT may be initially viable and robust. However, if mutant strategies are capable of invading it, TFT will not be stable. A&H make the strong claim that TFT can resist the invasion of any strategy. More specifically, they argue that once TFT is established, if the individuals who interact have a sufficiently high probability (w) of meeting again, TFT can resist invasion by any possible mutant strategy.[8] Just like with the initial viability of TFT, A&H use mathematical expressions to support the stability of TFT. In fact, they demonstrate that no strategy at all can invade TFT if and only if $w \geq (T - R)(T - P)$ and $w \geq (T - R)(R - S)$.

Although the mathematical details are interesting, they are not pressingly important. Instead, I will present A&H's argument for the stability of TFT in a more intuitive way. The general structure of the argument is the following: Neither Always-Cooperate nor Always-Defect can invade TFT. Moreover, a strategy that alternates between Cooperation and Defection cannot invade TFT. Hence, no strategy can invade TFT. I will go through each of the parts in greater detail. First, a strategy of Always-Cooperate (AC) cannot invade TFT. When TFT and AC play, they will both receive the R payoff (3 each) ever time. Because TFT and AC do equally well, AC cannot invade TFT. A strategy of Always-Defect (AD) is also incapable of invading TFT. If TFT plays AD numerous times, TFT will receive the S-payoff (0) the first encounter and the P-payoff (1) in all subsequent encounters. AD will receive the T-payoff (5) the first round and the P-payoff (1) for all remaining rounds. Hence, if they play n times, TFT will get a payoff of n-1 and AD will get a payoff of n+4. Although a difference of 5 units is a slight advantage, it is not significant enough for AD to invade TFT. There are two reasons why this difference is not particularly significant. First, since the value of w is high, the number of interactions (n) between

7" Axelrod and Hamilton, 1393.

8" Ibid.

two individuals will be large. Hence, for example, if $n = 50$, TFT will obtain a payoff of 59, while AD will receive a payoff of 64. Both individuals are obtaining large enough payoffs to potentially survive in the population. Secondly, these calculations do not take into account TFT's interactions with other TFTs. Every time TFT plays another TFT, it reaps the R-payoff (3). On the other hand, every time AD plays another AD, it only receives the P-payoff (1). Thus, as long as TFTs play each other a few times, they will obtain a larger payoff than ADs.[ix] Finally, a strategy that alternates between cooperation and defection will not be able to invade TFT. If TFT plays the Alternating Strategy (AS), the best AS can do is get a payoff that is 5 units greater than TFT.[x] Again, although this is a slight advantage, given a large *n* value and the interaction among TFTs, 5 units is not enough for AS to invade TFT. By showing that AC, AD, and AS are incapable of invading TFT, A&H conclude that no strategy is capable of invading TFT.[xi]

A&H convincingly argue that if individuals who interact have a sufficiently high probability (*w*) of meeting again, TFT can resist invasion by any possible mutant strategy. Nevertheless, A&H provide no theoretical or empirical support for the large value of *w.* The artificiality of this assumption is an important issue regarding the stability of TFT. I will defend the stability of TFT by arguing that even if the value of *w* is low, a group of TFTs can employ norms and metanorms to prevent other strategies from invading. Before developing this argument, however, it is important to clarify why a low value of *w* would threaten the stability of TFT. The point can be illustrated with the following example: Suppose that a defector begins playing a group of 4 TFTs. Since *w* is low, the defector will only play each TFT once. While the defector will reap a payoff of 12 (4 interactions * T-Payoff (3)), TFTs will receive a payoff of 0, since they will obtain the Sucker (0) payoff every time they play the defector. By obtaining such

large payoffs, defectors could survive, reproduce, and invade a group of TFTs. In this way, the stability of TFT is undermined by a sufficiently low *w.*

Even though the stability of TFT is undermined by a low value of *w*, TFTs can employ other mechanisms to strengthen their stability. In his paper *An Evolutionary Approach to Norms*, Robert Axelrod suggests a way in which a group could detect and punish defectors, thereby reinforcing its own stability. Although he does not directly argue that the use of norms and meta-norms strengthens the stability of TFT, his analysis is relevant and important. Axelrod defines a norm as existing "to the extent that individuals usually act in a certain way and are often punished when seen not to be acting in this way."[9] In order to investigate the growth and decay of norms, Axelrod developed the Norm Game. In the game, an individual has an opportunity to defect, accompanied by a chance of being observed by others. If the individual chooses to defect, he/she receives the Temptation payoff (3). Furthermore, if an individual sees another individual defecting, he/she may choose to punish the defector. In this case, the defector will receive the Painful Payoff (-9) and the enforcer will also pay an Enforcement Cost (-2). Axelrod tracks the players' strategies using two dimensions: boldness, which refers to how much a player defects, and vengefulness, which refers to how much a player punishes someone who is defecting.

Axelrod ran the Norm Game five rounds and found the following pattern. First, the boldness level of players decreased significantly. Since the vengefulness level in the population was initially quite high, it became very costly to be bold. Hence, individuals stopped defecting. Once the boldness level fell, vengefulness started decreasing as well. The reason this happened is that once boldness was sufficiently low, the incentive to punish a few defectors was not worth the Enforcement Cost (-2). Finally, once the vengefulness level fell to nearly zero, there was a

9" Axelrod, Robert. "An Evolutionary Approach to Norms."*American Political Science Review*. 80.4 (1986): 1095-1111. Print. Pg. 1097.

significant increase in boldness. Although one might expect vengefulness to rise with the increase in boldness, this did not occur. Instead, a state of high boldness and low vengefulness remained stable. From running the Norm Game several times, Axelrod concludes that the problem with establishing a norm is that no individual has a direct incentive to punish a defection.[10] As a result, Axelrod explores different mechanisms that can serve to support a norm.

One mechanism he suggests is the use of metanorms. A metanorm may be defined (in the context of Axelrod's game) as the expectation that individuals will punish those who do not punish a defection. Axelrod formulates a Metanorm Game to explore the effectiveness of this mechanism. The game is essentially the same as the Norm Game, but with the important addition that if an individual is caught seeing and not punishing a defection, the other players have a chance to punish that individual. Five runs of the Metanorm Game were conducted, yielding unambiguous results. In all runs of the game, the norm against defection was established; vengefulness quickly increased to a high level and boldness subsequently decreased. Axelrod claims that the results are not surprising. Since players had a personal incentive to be vengeful (that is, to escape punishment for not punishing an observed defection), individuals made sure to punish all instances of defection. As a result, the system became self-policing and the norm became well-established.[11]

The use of norms and metanorms is a mechanism which clusters of TFTs could employ to prevent other strategies from invading. There are two ways in which this mechanism strengthens the stability of a TFT cluster. First, by punishing an act of defection, the defector's fitness is directly undermined. This is represented in the game by the Painful Payoff (-9). Even if the defector obtains the Temptation Payoff (+5), the gain will be outweighed by the Painful Payoff (-

10" Axelrod, 1100.

11" Axelrod, 1102.

9) received through punishment. There is also a more subtle way in which the use of norms and metanorms can strengthen the stability of a TFT cluster. If a defector is punished by a cluster of TFTs, this sends a clear signal that if the defector decides to play another member of the cluster, it will suffer another strong punishment. As a result of this signal, the defector will be less likely to play another member of the cluster in the future. By preventing defectors from coming back, a cluster of TFTs can minimize the number of times that they interact with defectors and hence minimize the number of times that they receive the Sucker Payoff (0). This subtle advantage may not apply to all groups. In particular, it probably will only apply to organisms that have recognition capacities and memory systems. Still, it is an interesting way in which norms and metanorms can strengthen the stability of some groups of cooperators.

In the last couple of pages, I have argued that even if *w* is low, a group of TFTs can remain stable by employing norms and metanorms as mechanisms for punishing defectors. Although this has theoretical appeal, a salient charge is that it is empirically ungrounded. I will address this issue by making a few remarks about the empirical support for the use of norms and metanorms in nature. First, human beings employ a wide variety of norms and metanorms.[12] Sometimes they are enforced openly and aggressively, as in the case of international conflict when one country defects on another. Other times, norms are enforced subtly through the manipulation of feelings such as guilt and shame. There is also evidence that chimpanzees employ norms and metanorms. In his book "Chimpanzee Politics", Frans de Waal discusses the formation of large coalitions of chimpanzees. De Waal argues that chimpanzees act selectively when intervening in a conflict between other members of the group, taking into account the other chimpanzees' history of collaboration.[13] Thus, there is evidence that chimpanzees abide by norms

12 Fehr, Ernst, and Urs Fischbacher. "Social norms and human cooperation." *Trends in Cognitive Science*. 8.4 (2004): 185-90. Print. Pg. 185.

13 de Waal, Frans. *Chimpanzee Politics*. Baltimor: The Johns Hopkins University Press, 2007. 31. Print.

pertaining to conflict resolution. It is important to mention that species that have less complex social systems probably do not employ norms and metanorms. Hence, the stability of these groups must depend on other factors. However, for species that have intricate social systems, there is theoretical and empirical support that norms and metanorms can help strengthen the stability of cooperating groups.

I have now responded to the third and final issue regarding A&H's approach to the evolution of cooperation. In particular, I have argued that even if the value of *w* is low, socially intricate groups can remain stable by employing norms and metanorms. I provided both theoretical and empirical support for this claim. Additionally, I have responded to two other issues regarding A&H's model. First, I responded to Nowak and Sigmund's challenge by arguing that although WSLS does better than TFT against a Cooperator, TFT does better than WSLS against a Defector. Since Defection is the strategy that existed in nature before any cooperative strategies evolved, TFT is a more robust strategy than WSLS. Secondly, regarding the initial viability of TFT, I pressed A&H's account by arguing that they owe an explanation of how a cluster of TFTs could arise in a population of defectors. I responded to this issue by arguing that a cluster of TFTs could arise in two ways: First, it could arise simply by chance, and second, it could arise by the occurrence of several mutations in a small group of defectors. With these three arguments, I defended the robustness, stability, and initial viability of TFT.

To end, I will suggest one way of extending my analysis which I chose not to explore in this paper for the sake of focus and clarity. In defending A&H's model, I only discussed one-to-one player interactions. Extending my analysis to multiple-player interactions would further strengthen A&H's approach to the evolution of cooperation. For example, I argued that clusters of TFTs can arise if defectors mutate into TFTs in a small group of defectors. Exploring whether

this would still hold in groups with multiple-player interactions would extend A&H's model. Together with my defense of the robustness, stability, and initial viability of TFT, this analysis would significantly enhance A&H's general perspective on the evolution of cooperation.

i

ii

iii

iv

v

vi

vii

viii

ix

x

xi