Brit. J. Phil. Sci. **47** (1996), 533–551

Modelling Reciprocal Altruism

Christopher Stephens

ABSTRACT

Biologists rely extensively on the iterated Prisoner's Dilemma game to model reciprocal altruism. After examining the informal conditions necessary for reciprocal altruism, I argue that formal games besides the standard iterated Prisoner's Dilemma meet these conditions. One alternate representation, the *modified* Prisoner's Dilemma game, removes a standard but unnecessary condition; the other game is what I call a *Cook's Dilemma*. We should explore these new models of reciprocal altruism because they predict different stability characteristics for various strategies; for instance, I show that strategies such as *Tit-for-Tat* have different stability dynamics in these alternate models.

1 The altruism puzzle and the standard model

Scientists use mathematical models to make vague ideas and claims more precise. Besides having the virtue of making such assertions easily testable, the precision of mathematical models often yields surprising theoretical results that could not be predicted a priori by pre-mathematical intuition. But care must also be taken in formalizing our attempts to describe the world—scientists face many choices in deciding how to make their models and theories rigorous; one hopes to avoid the relatively unimportant details and preserve what is needed for a model to have predictive and explanatory success. Theoretical biologists are no exception, and face these sorts of decisions.

One of the puzzles facing biologists is the pervasiveness of cooperative behaviour among organisms. Prima facie, the theory of evolution by

natural selection implies that helping behaviour should not exist because organisms that do not help should do better than helpers by reaping the rewards of the help without incurring the costs. But, of course, organisms frequently do help one another. Why is this?

William Hamilton [1964] developed *inclusive fitness theory* (kin selection theory) to provide a partial explanation for the prevalence of cooperation. Helping behaviour among close relatives is selected for because close relatives share a large percentage of genes. But kin selection theory cannot provide a complete explanation of cooperative behaviour because helping often occurs between organisms that are not close relatives (Trivers [1971]; Wilkinson [1988]). Biologists appeal to *reciprocal altruism* to explain cooperation that cannot be accounted for by kin selection.[1] Roughly, reciprocal altruism evolves because organisms do better by accepting the immediate costs of helping another organism in order to reap the comparatively greater benefits of receiving help at a later time.

In order to develop this idea precisely, biologists rely extensively on the two-person iterated Prisoner's Dilemma game (Trivers [1971]; Axelrod and Hamilton [1981]; Maynard Smith [1982]; Axelrod [1984]; Dugatkin [1988]; Nowak and Sigmund [1993, 1994]). In a Prisoner's Dilemma game, each player has two possible behaviours: defect or cooperate. The game may be represented in the following way:

		Player 2	
		Cooperate	Defect
Player 1	Cooperate	W, W	X, Y
	Defect	Y, X	Z, Z

In biological models, W, X, Y and Z represent fitness payoffs to organisms.[2] The first letter in each outcome pair represents the fitness payoff to player 1, the second, the payoff to player 2.[3] Nearly all of the biological literature assumes, either explicitly or implicitly, that the Prisoner's

[1] Throughout this paper, I follow the standard biological practice and use 'reciprocal altruism' and 'reciprocal helping' synonymously to refer to behaviour in which fitness costs and benefits are exchanged between two organisms. Wilson and Sober [1994] argue that inclusive fitness theory (kin selection) and reciprocal altruism should both be considered special forms of group selection. Although I find their arguments persuasive, none of my claims about modelling reciprocal altruism depends on accepting their framework. A full discussion of the units of selection problem is obviously beyond the scope of this paper. See Wilson and Sober [1994] and the commentaries that follow for a recent critical exchange on this issue.

[2] There is no reason to assume that the organisms in question have a conscious mind or are able to deliberate rationally; as a result of natural selection, organisms which cooperate incur a fitness disadvantage, while organisms that defect incur a fitness advantage. However, it is necessary for organisms to have the cognitive capacity to recognize other individual organisms. Dugatkin [1988], for example, argues that guppies do have such a capacity.

[3] Axelrod [1984] uses the following names for the payoffs: $W = R$ (reward), $Y = T$ (temptation), $Z = P$ (punishment), $X = S$ (sucker). In my alternate games, the connotations of Axelrod's names can be misleading, so I use more neutral letters to represent these outcomes.

Dilemma game requires two main conditions:[4]

(C1) $Y > W > Z > X$ [*the ordering condition*] and
(C2) $(Y + X) < 2W$ [*the anti-exploitation condition*]

The situation is a dilemma because although 'defect' is the dominant choice for each player in a one-shot game, mutual cooperation is better than mutual defection.[5]

The Prisoner's Dilemma is useful for modelling reciprocal altruism because the game requires that the players be self-interested and that the interests of one player partially conflict with the interests of another. Just as individuals in a one-shot Prisoner's Dilemma maximize self-interest by defecting, organisms that act self-interestedly in nature should do better than altruists (see Williams [1966] and Dawkins [1989]). In his book, *The Evolution of Cooperation*, Axelrod expands the use of these game theory models and uses computer tournaments to make his case that in indefinite games cooperation is not, after all, very difficult to achieve, and once achieved proves surprisingly stable.[6,7]

Axelrod's work has been influential, and many have concluded that his computer tournaments provide evidence that strategies like *Tit-for-Tat* do well in a wide variety of environments. *Tit-for-Tat* (*TFT*) is the strategy which instructs the organism to cooperate on the first move of any given pairing, and then do whatever its opponent did on the previous move. The success of *TFT* is particularly interesting because it does well by eliciting cooperation from its opponents. One general lesson of these results is that it is beneficial to help others when such help is *conditionally reciprocated* by one's opponent.[8] So the success of strategies like *TFT* helps explain why

[4] Nearly all of the biological literature (at least since Axelrod and Hamilton [1981] and Maynard Smith [1982], pp. 202) require both conditions. Nowak and Sigmund [1994] is a notable exception. In non-iterated accounts of the Prisoner's Dilemma game, such as Luce and Raiffa [1957], only the first condition is mentioned; however, they do not consider evolutionary cases. The ordering condition defines a Prisoner's Dilemma game; the second condition is introduced so that an alternating exploitation will not evolve. See Sections **2** and **3** for critical discussion of these conditions.

[5] One strategy *dominates* another if an individual is better off playing that strategy over any alternative, regardless of what the other player does. See Luce and Raiffa [1957].

[6] In indefinite games, the termination of the paired interaction cannot be determined ahead of time by the organisms. Axelrod argues that it is only in indefinite games where cooperation is easily achieved and surprisingly stable. The worry is that in definite games, both organisms will defect on every move as a result of the so-called backward induction argument (Luce and Raiffa [1957]). But see Sober [1992] for critical discussion.

[7] Technically, the length of the game is determined by the probabilistic parameter w. It gives the probability that the game will continue until the next move. For example, games where $w = 0.5$ have an average length of 2 moves. In general, w has a fixed value between 0 and 1, and the expected length is $1/(1-w)$. See Axelrod [1984], pp. 13.

[8] Strategies like *Pavlov*, which have a high probability of cooperating after receiving W or Z and a high probability of defecting after receiving X or Y (see Nowak and Sigmund [1993]), share Tit-for-Tat's property of helping others when such help is conditionally reciprocated.

cooperation can become stable in cases where kin and group selection do not occur.

In this paper I argue that the iterated Prisoner's Dilemma game is but one way formally to model claims about reciprocal altruism. The standard iterated Prisoner's Dilemma game analysis makes an unnecessary assumption, and furthermore the Prisoner's Dilemma game is itself unnecessary to formally represent reciprocal altruism.[9] I will proceed roughly as follows: first, I develop carefully the intuitive, *informal* conditions necessary for reciprocal altruism; second, I argue that both Axelrod's ordering condition and the anti-exploitation condition are unnecessary. Next I show that other kinds of models, which I call the *modified* Prisoner's Dilemma and *Cook's Dilemma* games meet the informal conditions on reciprocal altruism and therefore can be used to model certain kinds of reciprocal altruism. Then I apply these alternate games to examples and briefly explore how the stability dynamics in the alternate models differ from the standard model. Finally, I conclude with reference to some general issues in adaptationism.

2 Informal conditions for reciprocal altruism

Before leaping immediately into the formal apparatus, it is useful to understand the *informal* characteristics that are necessary in modelling reciprocal altruism. After all, the hasty regimentation of reciprocal altruism may be precisely what leads others to ignore the possibility of alternate models. Once we have a handle on these crucial features, we will formalize them in game theoretic terms, thus enabling us to generate precise predictions about reciprocal altruism.

The following four conditions are individually necessary and jointly sufficient for an instance of reciprocal altruism:[10]

(i) the behaviour must reduce a donor's fitness relative to a selfish alternative;

(ii) the fitness of the recipient must be elevated relative to non-recipients;

(iii) the performance of the behaviour must not depend on the receipt of an immediate benefit;

[9] Boyd [1988] defends the Prisoner's Dilemma game as an appropriate model for reciprocal altruism; however, he is merely concerned to show that it is possible to model at least some cases of reciprocal altruism with the Prisoner's Dilemma game. I am not disputing that; rather, I am simply arguing that some cases of cooperation may best be modelled using other games.

[10] See Wilkinson [1984, 1988, 1990] for a similar effort to describe these informal conditions. As far as I can tell, there is widespread agreement that these conditions are necessary. However, most authors do not make them explicit.

(iv) conditions (i), (ii), and (iii) must apply to both individuals engaging in reciprocal helping.

Conditions (i) and (ii) are what make the behavior *altruistic*. Condition (iii) distinguishes reciprocal altruism from *mutualism*, in which the donor acts altruistically only if the recipient *simultaneously* provides a return benefit. Condition (iv) makes the altruism *reciprocal*.

At least two further conditions are necessary for reciprocal altruism to *evolve*:[11]

(v) a mechanism for detecting 'cheaters' must exist;
(vi) a large (indefinite) number of opportunities to exchange aid must exist.

Condition (v) is required so that altruists have a way of punishing organisms which do not cooperate; without this condition, non-altruists would always take advantage of altruists, so reciprocal altruism would not evolve. The mechanism does not have to be 'conscious'; a simple conditioning device is enough. Condition (vi) is necessary to avoid the consequences of the one-shot Prisoner's Dilemma game or backwards induction problem characteristic of games with a known finite number of interactions (see footnote 6). With these informal conditions in mind, let us now turn to the formal models.

3 Criticism of Axelrod's justification of the anti-exploitation condition

Axelrod [1984] argues that the anti-exploitation condition (C2) is necessary for models of reciprocal altruism. Most of the biological literature cites Axelrod and Hamilton [1981] and Axelrod [1984] or Maynard Smith [1982] as the reason for requiring the condition that $(Y + X) < 2W$. Axelrod ([1984], p. 10) states that the *definition* of the Prisoner's Dilemma requires both conditions (C1) and (C2). But the second condition is not necessary for the dilemma to exist. In the standard version of the game, where $(Y + X) < 2W$, the dilemma exists because although the rational choice is defection, both players would have been better off had they both cooperated. If the opponent cooperates, it is better to defect because $Y > W$, and if the opponent defects, it is also better to defect because $Z > X$. In either case, the dominant (and therefore rational) choice is to defect. Both players go through the same reasoning and hence both defect.

[11] These conditions are not sufficient for the evolution of reciprocal altruism; whether reciprocal altruism evolves depends on many factors, such as the frequency of the cooperation, the particular mix of strategies in a population (see Section 6), as well as random genetic drift, mutation, recombination and migration.

If the anti-exploitation condition (C2) is removed, and $(Y + X)$ is greater than or equal to $2W$, the dilemma remains. I call this model of reciprocal altruism the *modified* Prisoner's Dilemma game. $Y > W$ and $Z > X$ so therefore the rational choice in the noniterated case of the modified Prisoner's Dilemma is still defection. The condition that $(Y + X) < 2W$ is not necessary for the dilemma to exist in the noniterated case. But perhaps it is necessary in iterated (repeated) games. Axelrod explains that the second condition is necessary so that 'players cannot get out of their dilemma by taking turns exploiting one another' ([1984], p. 10). And since we are interested in situations where cooperation can develop even when all the players are self-interested, one requires the condition. Perhaps in the iterated game, when $(Y + X)$ is greater than $2W$, cooperation cannot develop and hence the modified Prisoner's Dilemma game *cannot be the proper model for reciprocal altruism.*

In support of this worry, one might point out that when (C2) holds, strategies like *TFT* do well in a wide variety of interactions with other strategies because *TFT* offers conditionally reciprocated help and as such tends to elicit cooperation from those it interacts with (Axelrod [1984]). Surely, if $(Y + X)$ is greater than $2W$, then the best strategy for both players is no longer mutual cooperation; rather, it is a strategy of alternate exploitation. Pairs of organisms which take turns receiving the Y (temptation) and X (sucker) payoffs will do better than pairs of organisms which receive W (reward) each turn. Hence, the argument continues, cooperation would not evolve. And consequently, the modified Prisoner's Dilemma game must be regarded as an inadequate model of reciprocal altruism because it does not even predict the basic empirical fact that reciprocal altruism exists.

Prima facie, the preceding argument seems correct: alternate exploitation is better than concurrent cooperation from both players when $(Y + X) > 2W$; therefore cooperation could be undermined by strategies which employ alternate exploitation. My response to this problem is simple: consider the alternate 'exploitation' a form of *cooperation.* After all, the benefits are mutual.[12] In this sense the game can also model cooperative behaviour in nature—instead of an immediately reciprocated cooperation, it is a *delayed cooperation.*[13] Each player allows itself to be

[12] This is not meant to suggest that the players are no longer acting only in their own interest, any more than the move 'C' entails that the players are not acting in their own interest in the standard Prisoner's Dilemma game analysis. The idea is simply that organisms which tend to get involved in reciprocal 'exploitation' of this type will do better (in terms of fitness) than organisms which tend to engage in simultaneous mutual cooperation or mutual defection.

[13] Nowak and Sigmund [1994] independently also propose using an alternating C-D as a kind of cooperation. Their paper complements and reinforces some of the conclusions reached here. They do not, however, consider the possibility that Cook's Dilemma games (see Section 4) could model reciprocal altruism.

exploited in exchange for the opportunity to exploit the other player. Axelrod denies that this is cooperation; a more judicious assessment would be that in the modified Prisoner's Dilemma game the cooperation is not necessarily represented as immediate; rather, it can occur over a two-turn period.

Removing the anti-exploitation condition may make modelling reciprocal altruism more complicated and fail to give any increased predictive accuracy. One might therefore require the second condition because it is a more accurate model, or because there is nothing theoretically interesting about the other models. These would be good reasons to ignore these other games, and I will respond to these sorts of concerns in Section **6**. My point here is simply that these alternatives should not be dismissed a priori.

Since the dilemma exists even when the condition $(Y + X) < 2W$ does not hold, it is wrong to view this condition as a requirement on Prisoner's Dilemma modelling. Furthermore, it is possible for cooperation to develop without this second condition. The second condition should merely be thought of as a convenient starting point of Prisoner's Dilemma analysis. Finally, and perhaps most importantly, the possibility of delayed cooperation will be crucial to examining an alternate game in which Axelrod's ordering condition $[Y > W > Z > X]$ does not hold.

4 A menu of formal models of reciprocal altruism

Only conditions (i)–(iv) (from Section **2**) apply to the game theory payoff structure; formally, they require that:

		Player 2		
		Cooperate		Defect
Player 1	Cooperate	W	$>$	X
		$\wedge$		$\wedge$
	Defect	Y	$>$	Z

For simplicity, these values represent the fitness payoffs only to player 1. Condition (i) requires that the cooperative choice reduce a donor's fitness (player 1) relative to some selfish alternative. This condition is met because $Y > W$ and $Z > X$. Condition (ii) may be read *strongly* or *weakly*. In the strong sense, (ii) requires that the fitness of the recipient (player 2) be raised relative to being a non-recipient of cooperation, *regardless of whether player 2 cooperates or defects*. This holds because $W > X$ and $Y > Z$. Condition (ii) may also be understood weakly, so that only $Y > Z$. In this case, the recipient of the helping benefits *only if he or she defects*. Note

that *none* of these informal conditions, either alone or collectively, requires the *ordering* or *anti-exploitation* conditions.

Sometimes too much helping is bad, just as too many cooks can spoil the soup. I call such a situation a *Cook's Dilemma* game. In a Cook's Dilemma game, simultaneous cooperation (W) is worse than simultaneous defection (Z). Nevertheless, reciprocal altruism can still exist in such situations, because nonsimultaneous cooperation is mutually beneficial (see below). In some cases, simultaneous cooperation is so disastrous that it is the worst payoff; in other cases, cooperating when the other player defects is the worst.

A careful specification of the informal conditions yields the following exhaustive list of formal models of reciprocal altruism:

Model 1: the standard Prisoner's Dilemma game:	$Y > W > Z > X$ and $(Y + X) < 2W$
Model 2: the modified Prisoner's Dilemma game:	$Y > W > Z > X$ and $(Y + X) \geqslant 2W$
Model 3: the unstable Cook's Dilemma game:	$Y > Z > W > X$ and $(Y + X) < 2Z$
Model 4: the strong Cook's Dilemma game:	$Y > Z \geqslant W > X$ and $(Y + X) \geqslant 2Z$
Model 5: the weak Cook's Dilemma game:	$Y > Z > X > W$ and $(Y + X) \geqslant 2Z$

In models of the first kind (and some instances of model 2), altruism is represented in the traditional way by *simultaneous cooperation* ('C-C'); in all other cases, reciprocal altruism is represented by *delayed cooperation* (an alternating 'C-D'/'D-C' for each player). Although model 3 meets all of the informal conditions (i) through (iv), it is not a stable kind of reciprocal altruism because both players would do better with simultaneous defection. Models of type 3 meet the conditions for reciprocal altruism, but not for a form of reciprocal altruism which will be selected for.[14]

In contrast, models 2, 4, and 5 represent kinds of reciprocal altruism in which delayed cooperation could be stable. Which kind of model one uses depends on the nature of the problem at hand. I will now provide a few examples of different kinds of models.

5 Modelling reciprocal altruism in guppies, baboons, and bats

The following examples illustrate the various models; in some cases, more than one model can represent the basic features of the situation. Since the models make different predictions (see Section **6**), which model is most appropriate is a matter that should be decided *empirically*. My point in the subsequent sections is simply to show how the new models could be used.

[14] The idea is that sooner or later some organisms will start defecting and these defecting pairs will do much better than those engaged in reciprocal altruism.

5.1 Modelling simultaneous cooperation in guppies

Example 1. One to a few guppies will break off from their school and go to inspect a predator to gain information about the predator's likelihood of attacking. Consider a pair of fish; each guppy has a choice—approach the predator (cooperate) or play safe and hang back (defect). If a guppy approaches the predator, it increases its chance of being eaten, but gains valuable information about the predator. If two guppies inspect together, the risk of being eaten for each inspector is reduced (see Dugatkin [1988]).

This example may be represented as a Prisoner's Dilemma game:

Standard Prisoner's Dilemma game (Model #1):

		Guppy #2	
		Approach	Hang Back
Guppy #1	Approach	3, 3	0, 5
	Hang Back	5, 0	1, 1

The cooperation between the two guppies is *simultaneous*, and the model represents it as such. Several biologists have run experiments which suggest that fish involved in repeated inspection visits play *Tit-for-Tat*, or some similar strategy (Dugatkin [1988]; Milinski *et al.* [1990]; and Dugatkin and Alfieri [1991a, b]). In one experiment (Dugatkin [1988]), the test fish moved closer to the predator when the control fish was made to appear closer to the predator. If the control fish 'defected' by failing to move close to the predator, the test fish did not move as close to the predator. The test fish's behaviour is compatible with a *Tit-for-Tat* strategy, as well as *Tit-for-Two-Tats* and similar strategies.

In all of these studies, the models presuppose that Axelrod's anti-exploitation condition on the Prisoner's Dilemma game [that $(Y + X) < 2W$] is true. In this case, it amounts to assuming that the risk of being eaten or injured cannot be too high relative to the gain in information (but reduced risk) of simultaneous cooperation. But this is an empirical claim which should not be assumed a priori. It is entirely plausible that the risk of being eaten/injured is very high, such that: $(Y + X)$ is *greater than* $2W$. The relative risk of being eaten can increase without raising the values of X or W. The standard Prisoner's Dilemma game cannot represent such possibilities.

Notice, however, that the second kind of game, the *modified* Prisoner's Dilemma, can model such possibilities. Even in cases of simultaneous helping, there are crucial empirical assumptions which should be tested to determine which model is more appropriate. By adjusting the payoff

values, one can generate an alternate model which yields different predictions. These other models ought to be tested to see if they provide better fit with the data.

5.2 Nonsimultaneous cooperation in baboons

At first glance, it might appear that reciprocal altruism should be represented on one of the nonstandard models because they represent reciprocal altruism as occurring over a two-turn period. After all, cooperation often does not occur simultaneously. It is difficult for two organisms to scratch each other's backs at the same time. On models 2–5, one can represent delayed reciprocal altruism by having one player receive benefits while the other player receives no benefits until some later time. And in the standard game, since all play is simultaneous and cooperation can exist only if both players choose 'cooperate' on the same turn of a game, it is puzzling to see how to model nonsimultaneous reciprocal altruism.

It is possible, however, to represent delayed cooperation on the standard model. A single iteration of the game is thought of as taking place over an interval which starts with an action by the first player and ends with an action by the second. Thus if organism *A* helps organism *B* at time t_1 and organism *B* helps *A* at time t_2, one can represent their interaction by treating the time interval t_1–t_2 as just *one* iteration in a Prisoner's Dilemma game where both *A* and *B* choose 'cooperate'. It is therefore important to see that the alternate models make different predictions about reciprocal altruism. I defend this claim in Section **6**.

Example 2. Packer [1977] reports that one male olive baboon (*Papio anubis*) will often solicit help from another male in fighting a rival baboon in order to mate with a female baboon. The situation can be characterized by a two-by-two game between the two males:the solicited male has a choice of whether or not to help—he gains no benefit at the time (and puts himself at risk) while the other (soliciting) male stands to gain access to the female baboon.

In this example, it is clear that the reciprocal helping is nonsimultaneous. Only one baboon receives the rewards; the other baboon simply incurs a cost of helping. It is natural to represent this sort of helping by a nonstandard model. The situation can be represented either on model 1 or on model 2. As with the guppies, which model is more appropriate depends on the empirical details.

5.3 Modelling reciprocal altruism with the Cook's Dilemma

In some cases, non-simultaneous cooperation may best be represented by a

Cook's Dilemma game. Cook's Dilemma's occur when too much helping is bad—'too many cook's spoil the broth'. Consider the following illustration:

Example 3. Wilkinson ([1984, 1988]) has observed Wild vampire bats (*Desmodus rotundus*) engaged in reciprocal altruism by regurgitating blood to other bats who are starving. The regurgitation is a form of food sharing in which the bat is sacrificing its own fitness (in the short run) for the fitness benefit of another bat. The regurgitation and donation of blood occurs even among non-kin, and as controlled experiments reveal, vampire bats will not continue to donate blood to individuals who do not reciprocate. Furthermore, the gain in health is not related in a linear way to gain in blood, so that a given quantity of blood is worth more to a starving bat than to a bat with plenty of blood.

Strong Cook's Dilemma game (Model #4)

		Bat #2	
		Cooperate	Defect
Bat #1	Cooperate	3,3	0,7
	Defect	7,0	3,3

How are we to understand 'cooperate' and 'defect' in this situation? Apparently, the bats know whether or not another bat has blood to give, and only solicit blood from those who have it; hence, they do not simultaneously regurgitate blood to one another. This is a clear case where too much helping (simultaneous regurgitation) would not be a good outcome for either bat.[15] Consequently, simultaneous cooperation does not occur. How then can one represent reciprocal food sharing among vampire bats? Since giving and receiving blood are conditional on having and needing blood, the following possibility suggests itself:

Cooperate: If asked for, donate blood; otherwise, wait.
Defect: If need blood, ask; if have blood, refuse to give.

Since cooperate and defect are now conditional, dispositional behaviours, there is no immediate behavioural difference between two bats that are

[15] One interesting feature of this example is that the donor–recipient relationship depends on the situation; which member is the recipient depends on which individual finds food. Consequently, an individual may be a donor several times in succession by chance. I am assuming that the donor can distinguish this chance element from defection. Of course, one bat may attempt to deceive another about whether it has blood; however, taking this into account would make the model more complicated than I can explore here. Sober [1994] develops a model of when such deception would evolve.

simultaneously cooperating and two that are simultaneously defecting. Consequently, I have assigned the same values to these two outcomes. The game is therefore a kind of Cook's Dilemma. The bats can only engage in nonsimultaneous reciprocal altruism. This game still shares with the standard Prisoner's Dilemma game the idea that there can be both reciprocators and non-reciprocators. Reciprocators, if asked for help, do so. Non-reciprocators, if asked for help, do not.

Of course, one could try to model the vampire bats' behaviour on a standard Prisoner's Dilemma. To do so, one would have to consider one move of simultaneous cooperation in the game as representing a delayed cooperation between two bats in which one bat regurgitates blood for another at one time and the second bat reciprocates at some later time. As before, this model yields different predictions, and neither should be eliminated a priori.

In order to illustrate model 5, consider the following hypothetical situation:

Example 4. Each member of a mated pair of sexually reproducing organisms can either forage for food (cooperate) or stay home and guard the young from predation (defect). We can represent their situation as a Cook's Dilemma game:

Weak Cook's Dilemma game (Model 5)

		Organism # 2	
		Forage (C)	Stay Home (D)
Organism # 1	Forage (C)	0, 0	1, 5
	Stay Home (D)	5, 1	2, 2

This is an example of a model 5 type of game: $Y(5) > Z(2) > X(1) > W(0)$

Staying at home is the non-cooperative behaviour, because you don't have to expend energy to gather food or risk predation. On the other hand, foraging is the cooperative choice—you go out and gather food for the offspring. However, if both parents forage at the same time, the offspring are in danger of predation from a predator, let us suppose, that is easily scared off by the adults but for whom the young are no match. As a result, simultaneous cooperation is actually worse than if you forage but your partner stays at home. The best outcome is staying at home and having your mate go forage; the next best outcome is for both you and your mate to remain at home; the third best outcome is for you to forage and your partner to remain at home, and the worst outcome for you is if both you

and your partner go out and forage, leaving the nest unprotected from predators.[16]

6 Characteristics of the alternate models

The alternate models would be uninteresting if the stability results for various strategies were the same as results of the standard model. In this section I briefly illustrate differences between the stability dynamics of the standard and new models of reciprocal altruism. A couple of examples will illustrate that the stability results are significantly different.

Consider the following strategies:

S1: *Tit-for-Tat* (*TFT*): cooperates on the first move, then does whatever its opponent did on its previous move.
S2: *Alternator* (*ALT*): alternates cooperate and defect, regardless of what the individual its paired up with does (50% of the time, it starts by defecting, and 50% of the time, it starts by cooperating).
S3: *All Defect* (*All D*): always defects, regardless of what the other player does.
S4: *All Cooperate* (*All C*): always cooperates, regardless of what the other player does.

It is important to distinguish different sorts of questions that one can ask about the stability of a strategy. One can enquire about its initial evolution, how it develops, or how resistant it is to invasion by other strategies. Here I simply give a few examples to show the difference between the standard and nonstandard models regarding the resistance to invasion of a particular strategy (e.g. *TFT*) by another strategy.[17]

In the standard game (model 1), *TFT* is a *Collectively Stable Strategy* (*CSS*) because for all S_j: Val(*TFT*/*TFT*) is greater than or equal to Val(S_j/*TFT*); (where Val(S_j/S_k) is read 'the payoff value to strategy *Sj* when it plays strategy S_k'). *TFT* always does at least as well playing itself as any other strategy does when it plays *TFT*.

[16] Of course, in some situations, foraging when your partner stays home is the best outcome, because you get both the advantage of a meal and having your young protected from predators. Whether a situation is best represented in one way rather than another depends on the empirical details: what is the risk of predation—both for you and your offspring—and how important is getting food (both for you and your offspring) relative to the predation risk? How much more helpful is it if both parents forage or stay at home rather than just one? Different answers to these questions may generate different models of reciprocal altruism. My project here is simply to illustrate the kind of empirical possibilities that would suggest that model 5 (Cook's Dilemma game) is appropriate to represent reciprocal altruism.

[17] For example, one could use computer tournaments analogous to those in Axelrod [1984] to explore both the initial evolution of a strategy and how it develops. Nowak and Sigmund [1994] run a computer tournament for model 2; the results suggest that this model favors *Generous Tit-for-Tat*, instead of the 'win-stay, lose-shift' (*Pavlov*) strategy more characteristic of the standard Prisoner's Dilemma game models.

Once we consider the alternate models, we can see that there are some significant differences in the strategies' resistance to invasion. For example, in model 2, if we let $Y = 7$, $W = 3$, $Z = 1$ and $X = 0$, then *TFT* is no longer a *CSS*; this is because Val(*ALT*/*TFT*) > Val(*TFT*/*TFT*). In model 3, *All D* is a *CSS*, and cannot be invaded. Hence, as remarked in Section **4**, model 3 is not an appropriate representation for the evolution of reciprocal altruism; however, it is useful to see why: *All D* is as good or better than any alternatives. In such situations, cooperation does not pay. In models 4 and 5, *TFT* is also no longer a *CSS*. To explore the consequences of assigning different values to W, X, Y, and Z, one can consult the following table. I have also provided a few graphs to illustrate more easily the resistance to invasion by various strategies in various kinds of games. (See pp. 550–1.)

Let n = number of rounds; W, X, Y, Z are defined as above:

	TFT	*ALT*	*ALLD*	*ALLC*
TFT	nW	$0.5nX + 0.5nY$	$X + (n-1)Z$	nW
ALT	$0.5nX + 0.5nY$	$0.25n[X + Y + W + Z]$	$0.5nX + 0.5nZ$	$0.5nW + 0.5nY$
ALLD	$Y + (n-1)Z$	$Y + (n-1)Z$	nZ	nY
ALLC	nW	$0.5nW + 0.5nX$	nX	nW

These graphs illustrate the resistance to invasion results for models 1, 2, 4, and 5.

Graph 1 (Model 1)
(Standard Prisoner's Dilemma Game)

	C	D
C	3,3	0,5
D	5,0	1,1

Here *TFT* is a *CSS*; *ALT* can never do better than *TFT*.

Graph 2 (Model 2)
(Modified Prisoner's Dilemma Game)

	C	D
C	3,3	0,7
D	7,0	1,1

The evolutionarily stable state occurs at 40% *ALT* and 60% *TFT*.

Graph 3 (Model 4)
(Strong Cook's Dilemma Game)

	C	D
C	3,3	0,7
D	7,0	3,3

The evolutionarily stable state occurs at 66.6% *ALT* and 33.3% *TFT*.

Graph 4 (Model 5)
(Weak Cook's Dilemma Game)

	C	D
C	0,0	1,5
D	5,1	2,2

The evolutionarily stable state occurs at 75% *ALT* and 25% *TFT*.

In models 2, 4, and 5, there is a stable polymorphism with these two strategies. In general, as simultaneous cooperation gets more costly, *TFT* tends to do worse against *ALT*.

7 Conclusions

The standard iterated Prisoner's Dilemma game has dominated biological models of reciprocal altruism. I have argued that there are alternate games

which meet the informal conditions on reciprocal altruism, and which yield both quantitatively and qualitatively different predictions. For instance, *TFT* is no longer a *CSS* in models 2, 4, and 5. We should explore these alternatives to see if they fit the data better. Of course, one might object that for all we know, the standard models will succeed in modelling all kinds of reciprocal altruism. After all, they succeed in predicting the basic fact that conditional reciprocators like *TFT* do well. The new models also predict that conditional reciprocators will prosper, but they differ over the details. Why, one might ask, should we care about these details?

One of the most important and heated debates in recent evolutionary biology concerns adaptationism and the relative strength of natural selection as compared to other forces such as random genetic drift and migration. Critics often charge that using optimality models often assumes that a particular trait is optimal, and that it is too easy to make *post hoc* adjustments concerning the fit between theory and data. In so far as these optimality models resist quantitative or even rough qualitative formulations, it is impossible to reveal whether natural selection has had merely some influence or has had a major influence. If we wish to answer these questions, we need to examine carefully the predictions of specific models.[18]

Of course, there is no guarantee that these game theory models will do the trick. Obviously, all of these models are simplified in various ways, and the only way to find out if they need further modifications (or a whole different approach) is by detailed analysis. At the present time, we do not yet have enough data to make this judgement. My point here is simply that these alternate models should not be ruled out a priori.

Acknowledgements

I hope to reciprocate all the help that Robin Andreasen, André Ariew, David Lorvick, Steve Orzack, Denis Walsh, David Sloan Wilson, and especially Elliott Sober provided me. Thanks also to an anonymous referee of this journal for helpful comments on an earlier draft.

Department of Philosophy
University of Wisconsin
Madison, WI 53706
USA

[18] This conclusion echoes that of Orzack and Sober [1994], who argue that clear quantitative statistical standards of the goodness of fit between theory and evidence must be used to test adaptationism.

REFERENCES

Axelrod, R. [1984]: *The Evolution of Cooperation*, NY, Basic Books, Inc.

Axelrod, R. and Hamilton, W. D. [1981]: 'The Evolution of Cooperation', *Science*, **211**, pp. 1390–6.

Boyd, R. [1988]: 'Is the Repeated Prisoner's Dilemma a Good Model of Reciprocal Altruism?' *Ethology and Sociobiology*, **9**, pp. 211–22.

Dawkins, R. [1976 &1989]: *The Selfish Gene*, Oxford, Oxford University Press.

Dugatkin, L. [1988] : 'Do Guppies Play Tit for Tat during Predator Inspection Visits?' *Behavioral Ecology and Sociobiology*, **23**, pp. 395–9.

Dugatkin, L. and Alfieri, M. [1991a]: 'Tit-for-Tat in Guppies', *Evolutionary Ecology*, **5**, 300–9.

Dugatkin, L. and Alfieri, M. [1991b]: 'Guppies and the Tit for Tat Strategy Preference Based on Past Interaction', *Behavioral Ecology and Sociobiology* , **28**, pp. 243–6.

Hamilton, W. D. [1964]: 'The Genetical Evolution of Social Behavior', *Journal of Theoretical Biology*, **7**, pp. 1–16, 17–32.

Luce, D. and Raiffa, H. [1957]: *Games and Decisions*, NY, Dover.

Maynard Smith, J. [1982]: *Evolution and the Theory of Games*, Cambridge, Cambridge University Press.

Milinski, Kulling, and Kettler, [1990]: 'Tit for Tat: Sticklebacks "Trusting" a Cooperative Partner', *International Society for Behavioral Ecology*, **1**, pp. 7–10.

Nowak, M. [1994]: 'The Alternating Prisoner's Dilemma', *Journal of Theoretical Biology*, **168**, pp. 219–26.

Nowak, M. and Sigmund, K. [1993]: 'A Strategy of Win–Stay, Lose–Shift that Outperforms Tit-for-Tat in the Prisoner's Dilemma Game', *Nature*, **364**, pp. 56–8.

Orzack, S. H. and Sober, E. [1994]: 'Optimality Models and the Test of Adaptationism', *American Naturalist*, **143**, 3, pp. 361–79.

Packer, C. [1977]: 'Reciprocal Altruism in *Papio Anubis*', *Nature*, **265**, pp. 441–3.

Sober, E. [1992]: 'Sable Cooperation in Iterated Prisoner's Dilemmas', *Economics and Philosophy*, 8, p. 127–39.

Sober, E. [1994]: 'The Primacy of Truth-telling and the Evolution of Lying', in *From a Biological Point of View*, Cambridge, Cambridge University Press.

Trivers, R. L. [1971]: 'The Evolution of Reciprocal Altruism', *Quarterly Review of Biology*, **46**, pp. 35–57.

Wilkinson, G. [1984]: 'Reciprocal Food Sharing in the Vampire Bat', *Nature*, **308**, pp. 181–4.

Wilkinson, G. [1988]: 'Reciprocal Altruism in Bats and Other Mammals', *Ethology and Sociobiology*, 8, pp. 85–100.

Wilkinson, G. [1990]: 'Food Sharing in Vampire Bats', *Scientific American*, February, pp. 76–82.

Williams, G. C. [1966]: *Adaptation and Natural Selection*, Princeton, Princeton University Press.

Wilson, D. S. and Sober, E. [1994]: 'Re-introducing Group Selection to the Human Behavioral Sciences', *Behavioral and Brain Sciences*, **17**, pp. 585–654.

Appendix: Resistance to invasion results for TFT and ALT

General definitions:

Strategy A is said to invade a population consisting of players using strategy B if Val(A/B) > Val(B/B). If no strategy exists which can invade B, B is said to be *collectively stable* (Axelrod [1984]). A strategy B is an *evolutionarily stable strategy* (*ESS*) relative to some strategy A if, either: (i) Val(A/B) < Val(B/B), or (ii) Val(A/B) = Val(B/B) and B receives a higher payoff than A in populations composed almost entirely of B's. A strategy is absolutely evolutionarily stable if it meets the above conditions for all A. All absolutely evolutionarily stable strategies are collectively stable, but not vice versa. For more information, see Maynard Smith (1982). An *evolutionarily stable state* is an equilibrium mix of strategies such that anytime the population deviates from such a state, evolution will favor organisms which lead the population back to the state.

Proofs for the resistance to invasion results for *TFT* and *ALT*:

In all cases, assume p = the frequency of Alternators in the population.

Graph 1 (*Model 1*): $W = 3$, $X = 0$, $Y = 5$, $Z = 1$

ALT population payoff (on average, in the long run) = Val(*ALT*/*ALT*)p + Val(*ALT*/*TFT*)(1 – p) = 2.25p + 2.5(1 – p) = 2.5 –0.25p. *TFT* payoff = Val(*TFT*/*TFT*)(1–p) + Val(*TFT*/*ALT*) p = 3(1 – p) + 2.5p = 3 – 0.5p. Since 3 – 0.5p > 2.5 – 0.25p for all frequencies $1 \geq p \geq 0$, *TFT* always does better than *ALT*, and hence is an *ESS* relative to *ALT*, and a *CSS* generally (proof in Axelrod [1984]).

Graph 2 (*Model 2*): $W = 3$, $X = 0$, $Y = 7$, $Z = 1$

ALT population payoff = 2.75p + 3.5(1 – p) = 3.5 – 0.75p. *TFT* population payoff = 3(1 – p) + 3.5p = 3 + 0.5p. Setting the two equations equal to each other and solving for p, we get: p = 0.4. So an evolutionarily stable state occurs at 40% *ALT*, 60% *TFT*.

Graph 3 (*Model 4*): $W = 3$, $X = 0$, $Y = 7$, $Z = 3$

ALT population payoff = 3.25p + 3.5(1 – p) = 3.5 – 0.25p *TFT* payoff = 3.5p + 3(1 – p) = 3 + 0.5p

Here p = 2/3, so an evolutionarily stable state occurs at 66 and 2/3% *ALT*, 33 and 1/3% *TFT*.

Graph 4 (*Model 5*): $W = 0$, $X = 1$, $Y = 5$, $Z = 2$

ALT population payoff = 2p + 3(1 – p) = 3 – p; *TFT* payoff = 3p + 0 (1 – p)

Here p = 0.75; an evolutionarily stable state occurs at 75% *ALT* and 25% *TFT*.

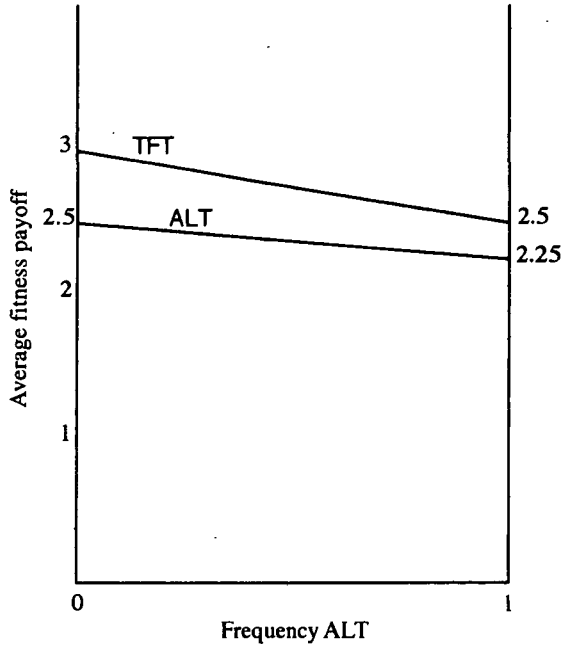

Fig 1

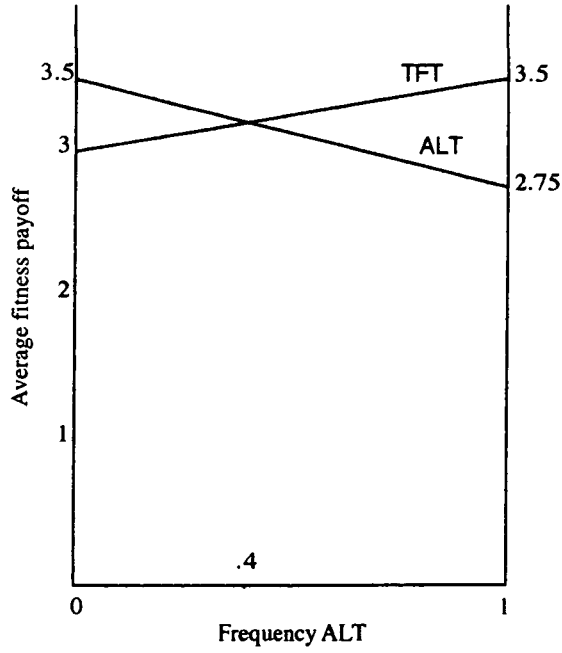

Fig 2

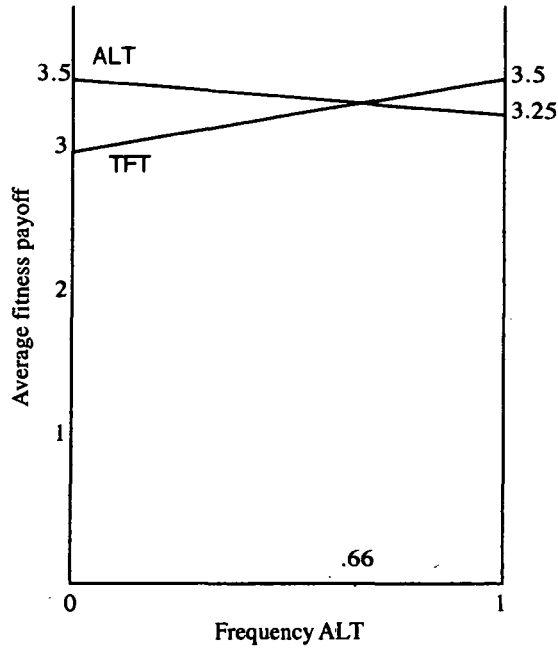

Fig 3

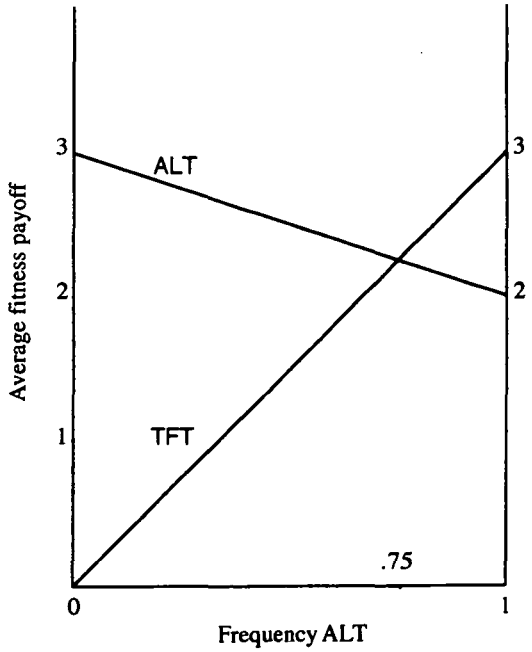

Fig 4