

# A simple and general explanation for the evolution of altruism

Jeffrey A. Fletcher<sup>1,3,\*</sup> and Michael Doebeli<sup>1,2</sup>

<sup>1</sup>*Department of Zoology, University of British Columbia, No. 2370-6270 University Boulevard, Vancouver, British Columbia, Canada V6T 1Z4*

<sup>2</sup>*Department of Mathematics, University of British Columbia, Vancouver, British Columbia, Canada V6T 1Z2*

<sup>3</sup>*Systems Science Graduate Program, Portland State University, Portland, OR 97207-0751, USA*

We present a simple framework that highlights the most fundamental requirement for the evolution of altruism: assortment between individuals carrying the cooperative genotype and the helping behaviours of others with which these individuals interact. We partition the fitness effects on individuals into those due to self and those due to the ‘interaction environment’, and show that it is the latter that is most fundamental to understanding the evolution of altruism. We illustrate that while kinship or genetic similarity among those interacting may generate a favourable structure of interaction environments, it is not a fundamental requirement for the evolution of altruism, and even suicidal aid can theoretically evolve without help ever being exchanged among genetically similar individuals. Using our simple framework, we also clarify a common confusion made in the literature between alternative fitness accounting methods (which may equally apply to the same biological circumstances) and unique causal mechanisms for creating the assortment necessary for altruism to be favoured by natural selection.

**Keywords:** altruism; assortment; cooperation; fitness accounting; interaction environment; population structure

## 1. INTRODUCTION

The evolution of altruism poses a problem in evolutionary theory: how can natural selection favour individuals that carry helping traits, over those that carry selfish ones? Historically (as well as recently; Wilson 2005, 2008; Fletcher *et al.* 2006; Foster *et al.* 2006; Nowak 2006; West *et al.* 2008), competing theories have sparked much controversy and this debate has been couched in terms of which theories best explain the evolution of altruism and under what conditions one theory may be superior to another. Here, we present what we believe to be the most fundamental explanation for how altruistic traits evolve by returning to first principles—the conditions necessary for any genetic trait to increase (altruistic or not). This framework focuses on determining when carriers of altruistic genes on average receive more net fitness benefits than carriers of alternative genes. The framework is an alternative to the standard theories (e.g. kin selection, multi-level selection, reciprocal altruism) that underlies them all. It thus supports efforts to unify multi-level selection and inclusive fitness theories (Wade 1980; Queller 1985, 1992; Frank 1998; Sober & Wilson 1998), as well as inclusive fitness and reciprocal altruism theories (Queller 1985; Frank 1998; Fletcher & Zwick 2006).

For a focal genotype of interest to increase in frequency in a population, carriers must, on average, end up with more net direct fitness benefits than average population members. This net direct fitness must account not only for any costs and/or benefits to the focal carrier due to its own

behaviour associated with the trait, but also for any fitness benefits received from other individuals (related or not; Queller 1985; Fletcher & Zwick 2006). In particular, any altruistic trait that causes carriers via their own behaviours to put themselves at a disadvantage compared with those they interact with, will only increase if the benefits received from others are sufficient to make up for this disadvantage. Here, we provide a general and transparent approach that resolves this problem by emphasizing the role of assortment, i.e. the association between carriers of the altruistic genotype and the helping behaviours they receive from others. The models we use to explain this approach are well known, but our interpretation of these models and the perspective it provides for the evolution of cooperation appear to be novel.

The generality of our approach has a trade-off. On the one hand, it leads to a very simple and straightforward solution to the problem of altruism on the basis of assortment; but on the other hand, it does not specify the mechanisms that can generate this assortment between focal genotype and help from others. In fact, in our view, the real problem of altruism lies exactly in identifying the biological mechanisms for such assortment. Each of the aforementioned theories has its strengths in emphasizing different (and in some cases overlapping) mechanisms for assortment, e.g. limited dispersal, kin recognition, group structure, conditional behaviour. Yet, this can become confusing when these theories keep track of fitness in different ways without identifying the underlying commonality, i.e. the requirement for assortment. Thus, what appear to be different causal explanations can in fact be just different ‘fitness accounting techniques’ applied to similar situations. In particular, several recent papers have claimed

\* Author and address for correspondence: Systems Science Graduate Program, Portland State University, Portland, OR 97207-0751, USA (jeff@pdx.edu).

Table 1. Total pay-offs in groups of  $N$  players with  $k$  cooperators.

phenotype	pay-off received from own behaviour	pay-off received from the behaviour of others in an interaction environment (excluding self)	total direct pay-off (within group)
cooperate (C)	$(b/N) - c$	$(k-1)b/N$ (there are $k-1$ cooperators and $N-k$ defectors)	$(kb/N) - c$
defect (D)	0	$kb/N$ (there are $k$ cooperators and $N-k-1$ defectors)	$kb/N$

that kin selection is a theoretically necessary mechanism for the evolution of altruism (Foster *et al.* 2006; Lehmann & Keller 2006; West *et al.* 2007a,b), while others have argued that multi-level selection is an equally valid explanation (Wilson 2005, 2008; Wilson & Hölldobler 2005). We address these claims in §4.

## 2. INTERACTION ENVIRONMENTS IN THE PUBLIC GOODS GAME

The public goods game (Olson 1971) is a simple, but fundamental and powerful metaphor for the problem of cooperation that has been used, in various forms (including the Prisoner's Dilemma) in a large part of the altruism and cooperation literature. Here, we use a simple evolutionary model based on this game to clarify various issues and assumptions, as well as the basic conceptual conditions that need to be satisfied for altruism to evolve.

The most basic setup of the public goods game is as follows. There are two possible strategies: cooperate (C) and defect (D). The game is played within interaction groups of  $N$  players, in which a C behaviour contributes an amount  $b$  to the public good at a cost  $c$  to the cooperator, where it is assumed that the enhanced value to the public good exceeds the cooperator's cost, i.e.  $b > c$ . D behaviours contribute nothing to the public good and impose no costs. We first consider what happens in a given interaction group of  $N$  players,  $k$  of which cooperate and  $N-k$  of which defect ( $1 \leq k \leq N-1$ , so that the interaction group contains both cooperators and defectors). In such a group, the total size of the public good is  $kb$ , which is distributed evenly among all  $N$  interacting players. So, all group members receive a pay-off of  $kb/N$ . However, cooperators pay a cost, whereas defectors do not. Therefore, the net pay-off to cooperators is  $kb/N - c$ , whereas the net pay-off to defectors is  $kb/N$ . It follows that within any given interaction group, defectors have a higher pay-off than cooperators, which is the basic dilemma of altruism.

### (a) *The within-group interaction environment*

It is useful to unveil the structure of this dilemma by partitioning the pay-offs received by individual cooperators and defectors according to the contribution to the pay-off made by the individual itself, and the contribution made by the other individuals it interacts with, as shown in table 1. From its own contribution to the public good, a cooperator receives the net direct pay-off  $b/N - c$ , whereas it receives the pay-off  $(k-1)b/N$  from the other  $k-1$  cooperators in its interaction group (it does not receive anything from the  $N-k$  defectors). Thus, the partitioning  $kb/N - c = (b/N - c) + (k-1)b/N$  of the total pay-off received by a single cooperator divides the total pay-off into one part that comes from the cooperator itself and one part that comes from its interaction environment, which consists of  $k-1$  cooperators and  $N-k$  defectors. This decomposition into pay-off due to self (one's own

behaviour) and pay-off due to the individual's environment (behaviour of others) is natural and useful given our individual-based fitness accounting method. For a defector, the same partitioning into pay-off received from self and pay-off received from the group environment can be made. In a given interaction group, a defector gains a pay-off of zero from its own action and receives a pay-off  $kb/N$  from the  $k$  cooperators.

This partitioning of the pay-off reveals a number of things. First, it shows whether the net pay-off to a cooperator from its own cooperative act is more or less than what defectors give themselves (i.e. positive or negative given that we set the fitness consequences of defectors to zero). Some researchers use this distinction to delineate weak from strong altruism (Wilson 1979, 1990), with the former referring to the case when the net pay-off from self is positive (i.e. when  $b/N > c$ ) and the latter referring to the case when the net pay-off from self is negative (i.e. when  $b/N < c$ ). However, in any given interaction group of  $N$  individuals, this distinction is not fundamental, because the cooperators do worse than the defectors regardless of whether the pay-off to self for cooperators is positive or negative, i.e. regardless of whether altruism is strong or weak (Wilson 2004).

Table 1 clearly shows the reason for this: within any given interaction group, defectors experience different interaction environments to an extent that makes the difference between strong and weak altruism irrelevant within each interaction group. The interaction environment of a defector contains  $k$  cooperators, but the interaction environment of a cooperator contains only  $k-1$  cooperators. As long as the direct cost for cooperating ( $c$ ) is itself positive, the net gain for a cooperator from its own action is never enough to make up for the deficit of one less cooperator in its environment. This is true regardless of whether  $b/N > c$  or  $b/N < c$ .

### (b) *The average interaction environment*

The above considerations based on partitioning of the total pay-off into pay-off due to self and pay-off due to the environment cannot only be applied at the level of a given interaction group, but they can be generalized to the level of the whole population, where they reveal the basic solution to the problem of altruism. At the level of the whole population, focal individuals of a particular genotype can experience different types of interaction environments, i.e. interaction groups with different compositions. To determine the average pay-offs of cooperators and defectors, we therefore have to know the composition of the *average* interaction group in which a cooperator or a defector finds itself. Let  $e_C$  be the number of cooperators among the  $N-1$  interaction partners in the average interaction group of a focal C individual. The average pay-off to a C individual from its interaction environment is then simply  $e_C b/N$ . Combining this with the pay-off received by a cooperator from its own

Table 2. Total average pay-offs at the population level.

phenotype	pay-off received from own behaviour	pay-off received from others' behaviours in an average interaction environment (excluding self)	average total pay-off
cooperate	$(b/N) - c$	$e_C b/N$	$(e_C b/N) + (b/N) - c$
defect	0	$e_D b/N$	$e_D b/N$

cooperative act,  $b/N - c$ , yields the total pay-off of the cooperator as  $e_C b/N + b/N - c$ . Similarly, if  $e_D$  is the number of cooperators among the  $N - 1$  interaction partners in the average interaction environment of a focal D individual, the average pay-off received by the focal D from its interaction environment is  $e_D b/N$ . Since D does not receive pay-offs due to self, this is also the total average pay-off received by D individuals. The partitioning of the total average pay-off of C and D into pay-off due to self and pay-off due to the average interaction environment is shown in table 2.

The average pay-offs to C and D determine their evolutionary fate. In particular, the C genotype increases in frequency if its average pay-off is bigger than that of D, i.e. if

$$\frac{e_C b}{N} + \frac{b}{N} - c > \frac{e_D b}{N}. \tag{2.1}$$

For example, assume that the population is well mixed, so that interaction groups are formed at random from a large population. In this case, the distribution of different interaction group compositions is binomial. For example, if  $p$  is the frequency of cooperators in the population, the frequency of interaction groups consisting entirely of cooperators is  $p^N$ . Averaging over all interaction groups containing a focal C or a focal D player, it immediately follows that  $e_C = e_D = p(N - 1)$ . In particular, the average interaction environment is exactly the same for C and D, which is of course a simple consequence of the assumption of random interactions. Therefore, at the population level, the average pay-off received by focal cooperators and defectors from their interaction environments are the same, and the only difference is the pay-off due to self. It follows that cooperators do better than defectors if and only if the net pay-off for a cooperator's own behaviour ( $b/N - c$ ) is positive, i.e. if and only if altruism is weak.

In this special case, the distinction between weak and strong altruism is indeed the dividing line between selection for or against cooperators. In the literature, this seems to be the basis for the contention that the distinction between weak and strong altruism is fundamental (i.e. that weak altruism is not true altruism; Nunney 1985, 2000; Maynard Smith 1998; Foster *et al.* 2006; Lehmann & Keller 2006) because whether it evolves depends on the net pay-off to self being positive. However, it is important to realize that this distinction only makes the correct prediction if interactions are random, i.e. if the groups in which the public goods game is played are formed at random. If interactions are not random, then it is possible that cooperation loses out even if it amounts to weak altruism. More importantly, non-random interactions can lead to the evolution of cooperation even if altruism is strong, i.e. even if  $b/N - c < 0$ .

In fact, this is the central insight that can be gained from the pay-off partitioning given in table 2. If  $b/N - c < 0$ , it is clear from table 2 that for C to increase in frequency,

$e_C$  must be bigger than  $e_D$ . In other words, for (strong) altruism to evolve, the average interaction environment of C individuals must contain more cooperation than the average interaction environment of D individuals. Thus, the evolution of altruism requires (positive) assortment between focal C players and cooperative acts in their interaction environment, and such assortment is the basic mechanism by which altruism can evolve. Note that in the simple public goods model described above, cooperative behaviours are hardwired to the cooperation allele, so that  $e_C$  and  $e_D$  are simply the average numbers of cooperative genotypes in the interaction neighbourhoods of focal C and D players. More generally, however,  $e_C$  and  $e_D$  refer to the average number of cooperative behaviours in interaction neighbourhoods irrespective of whether this behaviour is coded for by the focal C genotype. This becomes important, e.g. when there are multiple alleles coding for similar phenotypes, with conditional behaviour, and for interspecific mutualisms. As we will discuss below, the essential insight that altruism requires  $e_C > e_D$  also applies in these situations, showing that in principle, it is the fitness consequences of behaviours on the focal C genotype that matter, not the genetic similarity of donors and recipients (i.e. relatedness).

A rearrangement of equation (2.1) quantifies the amount of relative assortment necessary for the evolution of altruism in this simple model

$$e_C - e_D > \frac{cN}{b} - 1. \tag{2.2}$$

For example, consider a hypothetical situation of extreme assortment, in which, e.g. due to experimental design or conditional interactions, cooperators only interact with other cooperators and defectors only interact with other defectors. Then  $e_C = N - 1$  and  $e_D = 0$ , and hence inequality (2.2) is satisfied whenever  $b > c > 0$ . Thus, given the basic assumption that benefits produced outweigh costs, no matter how costly cooperation is, it always evolves under such extreme assortment. On the other hand, imagine a situation in which every interaction group contains exactly one cooperator, so that there is 'over-dispersion' of cooperators, i.e. negative assortment. Then  $e_C = 0$ , while  $e_D = 1$ , and inequality (2.2) is never satisfied. Thus, under this interaction structure, the C type is always selected against at the population level, regardless of whether altruism is weak or strong. In particular, the explanation that weak altruism is selected for because weak altruists give themselves a positive pay-off (e.g. Nunney 1985, 2000), whereas D types give themselves nothing, does not hold in general. Even if cooperators give themselves a large positive pay-off, i.e.  $b/N \gg c$ , defectors still win because they experience, on average, more cooperative interaction environments. Note that even if labelled 'weak', cooperators can still be considered fully altruistic in that they always give their interaction partners

more than they give themselves (table 1). Thus, even weak C types are selected against unless they receive enough fitness benefits from others, measured across all possible interaction environments, i.e. unless  $e_C$  is large enough. The pay-off due to self (i.e. whether altruism is strong or weak) does of course influence the evolutionary outcome, but the crux of understanding the evolution of altruism lies in understanding the mechanisms that lead to different degrees of assortment between carriers of the altruistic genotype and help from those they interact with.

### (c) *Hamilton's rule*

Inequality (2.2) above readily generates a Hamilton's (1964, 1975) rule for the public goods game when the inequality is written in the equivalent form

$$\left(\frac{e_C + 1}{N} - \frac{e_D}{N}\right)b > c. \quad (2.3)$$

The term  $r = ((e_C + 1)/N) - (e_D/N)$  plays the role of 'relatedness', and inequality (2.3) is most easily interpreted in terms of direct fitness (Taylor & Frank 1996; Taylor *et al.* 2007):  $(e_C + 1)/N$  is the probability that a focal C player receives help from any given individual (including the focal player) in an average interaction environment of the focal C player, whereas  $e_D/N$  is the probability that a focal D player receives help from any given individual in its average interaction environment. The difference between these two probabilities, multiplied by the amount of benefit, measures how much more help a focal C player receives on average compared with a focal D player; which in turn determines whether or not the cost of cooperation is on average outweighed by benefits received. In our examples above,  $r=1$  in the case of extreme assortment ( $e_C=N-1$ ,  $e_D=0$ ) and  $r=0$  in the case with negative assortment ( $e_C=0$ ,  $e_D=1$ ). Thus, the examples reflect the fact that according to Hamilton's rule, altruism always evolves when  $r=1$  (if  $b > c > 0$  as always) and never evolves when  $r=0$ .

Note that although we have an analogous expression to Hamilton's rule, we do not need to invoke inclusive fitness. While the  $r$  term in Hamilton's rule is usually interpreted as a measure of the relatedness between actors and recipients, here we have interpreted it strictly in terms of the relative amount of helping behaviours in the average interaction environments of carriers of the cooperator and defector genotypes. We note that this more general condition is always met when altruism is selected for in both kin selection and multi-level selection models. (A formal treatment of Hamilton's  $r$  term as a measure of the assortment between the focal genotype and the phenotypes of focal individuals and their interaction partners can be found in Queller (1985).)

### 3. A SIMPLE MODEL OF NON-KIN ALTRUISM

To further illustrate that altruism can in principle evolve even when carriers of an altruistic genotype do not provide help to other carriers of the same genotype (i.e. in the absence of 'kin selection'), consider the following thought experiment. We start with an (infinite) population of haploid individuals (e.g. bacteria) and assume that altruism consists of producing a common good, e.g. an enzyme made available to the other members of an interaction group, at a cost to self. Imagine that the metabolic pathways needed to produce the common good can be activated

through two different and independent genetic regulatory mechanisms, which are encoded by two different loci. At the first locus, allele  $A$  activates the first regulatory mechanism that induces the metabolic pathway generating the common good, while allele  $a$  does not activate this regulatory mechanism. At the second locus, allele  $B$  activates the other regulatory mechanism that induces the metabolic pathway generating the common good, while allele  $b$  does not activate that second regulatory mechanism. For simplicity, we can assume that the genotype  $AB$  is inviable (e.g. due to overproduction of some toxic metabolite common to both pathways).

We then have three genotypes in the population:  $ab$ ;  $Ab$ ; and  $aB$ . The first of these is a defector, while the other two are cooperators. Now assume that the costs and benefits in interactions among these phenotypes (i.e. cooperators and defectors) can be modelled as a two-player Prisoner's Dilemma game with pairwise interactions (which by definition implements strong altruism; Fletcher & Zwick 2007). As a check on whether genetic similarity is necessary for cooperation, an experimenter then imposes assortment between cooperators (and between defectors) in the following way:  $Ab$  always interacts pairwise with  $aB$ , and  $ab$  always interacts pairwise with other  $ab$  genotypes. In other words, the experimenter imposes the strongest possible assortment between cooperators, but in such a way that carriers of the altruistic allele  $A$  never interact with other carriers of  $A$  and carriers of the altruistic allele  $B$  never interact with other carriers of  $B$ . It is clear that in this situation, both alleles  $A$  and  $B$  will increase in frequency, and the genotype  $ab$  will go extinct, leaving a population consisting of cooperators only.

However, it is also clear that the help which carriers of  $A$  provide never goes to other carriers of  $A$  and the help that carriers of  $B$  provide never goes to other carriers of  $B$ . Thus, even though cooperation evolves, it does not evolve due to kin selection (genetic similarity among those that interact). Instead, it evolves owing to assortment between phenotypic cooperators. More precisely, cooperation evolves because carriers of cooperative alleles, whether  $A$  or  $B$ , receive help from cooperative phenotypes, not from other carriers of the same allele. Note that our thought experiment also works for initially rare altruistic alleles, i.e. altruistic alleles that are invading a population consisting of defectors. In this case, one would need to assume that both altruistic alleles are present at low frequencies. To ensure that there are enough individuals for the experimenter to be able to impose strong assortment, one could assume that all genotypes experience a selectively neutral growth phase before the action of selection (as, e.g. envisioned in the model of Ackermann *et al.* 2008). Our example is admittedly very artificial. Nevertheless, it is logically consistent, and it shows that it is not relatedness *per se* that is fundamental for the evolution of cooperation. Rather, the fundamental mechanism enabling the evolution of altruism is assortment between carriers of altruistic alleles and phenotypes exhibiting helping behaviours. In the example given, this mechanism would be cumbersome to capture in terms of inclusive fitness, but is very easy to describe using direct fitness based on average environments experienced by different types of individuals.

To emphasize the generality of the direct fitness approach, we can extend the above example by assuming that cooperative acts are suicidal. Of course, in this case,

an altruistic gene cannot unconditionally specify the altruistic phenotype, because for such genes to increase in frequency at least some carriers must have offspring. Specifically, in the above example, assume that individuals carrying a cooperative allele (*A* or *B*) have a probability  $q$  to actually perform the altruistic act (before reproducing) and die as a consequence (e.g. because lysis is necessary to release the helpful enzyme). Then, still assuming the same thought experimental setup as above, the expected pay-off for an individual carrying a cooperative allele is  $(1 - q)qb$ , where  $b$  measures the benefit due to enzyme production of the partner in a given interacting pair. This is because for any individual carrying a cooperative allele, the probability to survive is  $1 - q$ , and the probability that the individual receives help from the other individual in the pair, i.e. that the other individual dies, is  $q$ . Thus, cooperation will evolve whenever the quantity  $(1 - q)qb$  is larger than the pay-off that  $ab$  individuals receive in pairwise interactions with other  $ab$  individuals.

This example shows that in principle, suicidal altruism can evolve even when the suicidal act does not benefit related genotypes, i.e. in the absence of kin selection. The crucial requirement is that when compared with carriers of defector alleles, carriers of alleles encoding the altruistic phenotype interact more often with individuals expressing the altruistic act. With suicidal altruism, it is important to view the cooperation genotype as incorporating conditionality, i.e. as encoding the *total* phenotype 'commit altruistic suicide with a certain probability (or under certain conditions); do not cooperate otherwise'. Evolution of such altruism then requires assortment between carriers of altruistic *genotypes* and altruistic *phenotypes* of others. For a more elaborate mathematical model of stochastic suicide in public goods games, see [Ackermann \*et al.\* 2008](#).

#### 4. DISCUSSION

The models presented here are intentionally simple in order to emphasize the most basic requirement for the evolution of altruism: positive assortment between carriers of the altruistic genotype and altruistic behaviour of others. The model can easily be modified to accommodate other situations such as those where altruists give only to others ([Pepper 2000](#)) or where cooperative behaviour is not strictly specified by genotype ([Ackermann \*et al.\* 2008](#)). Our brief example of suicidal aid involves both of these points and illustrates the need to clearly distinguish between phenotype and genotype when analysing the conditions necessary for the evolution of altruism. It can be confusing when altruism is defined in terms of phenotypic behaviours, but the evolution of altruism is defined in terms of the frequency of a focal genotype, without clearly describing the relationship between the two (e.g. [Foster \*et al.\* 2006](#); [Lehmann & Keller 2006](#); [West \*et al.\* 2007b](#)). This is especially true if some carriers of the focal genotype must be phenotypically non-altruists in order for altruism to evolve, as is the case with suicidal aid. This is the reason we have emphasized the interaction environment of carriers of genes for cooperation. The interaction environment needs to contain enough phenotypic helping behaviours for altruism to evolve, independent of whether this helping behaviour is coded for by the same or different genes. This is very similar to the

argument presented in [Kerr & Godfrey-Smith \(2002, appendix 2\)](#), which is also based on the expected environments of altruists and defectors. Accordingly, what is necessary for the evolution of altruism is assortment between focal genotype and phenotypic help, rather than the assortment among genetic types often emphasized in kin selection theory. As a consequence, our framework can be applied not only to interactions among relatives or those sharing the focal genotype, but also to interactions among non-relatives and even to interactions across species in mutualistic interactions. In fact, the basic requirements for the evolution of altruism between two different species are conceptually exactly the same as those for intraspecific altruism: cooperative genotypes in species 1 must receive sufficiently more cooperation from species 2 individuals than non-cooperative genotypes in species 1, and similarly for cooperative genotypes in species 2. In other words, there must be assortment between cooperative genotypes of either species and cooperative behaviours in the other species. For example, [Doebeli & Knowlton \(1998\)](#) showed how the assortment necessary for the evolution of mutualism can be generated by spatial structure, and [Fletcher & Zwick \(2006\)](#) showed how it can be generated by conditional interspecific behaviours.

In models of social interactions, we assume benefits come from a donor and go to a recipient. We can think of this as a flow of fitness benefits, and we can keep track of the total benefit produced by donors in a population by counting benefits as they leave focal donors headed for other individuals (the indirect fitness method emphasizing the origin of this flow), or as they arrive at focal recipients (the direct fitness method emphasizing the destination of this flow). Note that if we count benefits in both places, this would result in a double counting error. So each method necessarily ignores one end of the flow of fitness benefits due to social interactions. Note that these are merely two different methods of keeping track of fitness (i.e. different fitness accounting methods), but not alternative mechanisms by which a focal genotype can increase in a population. If the average indirect fitness for carriers of a certain genotype is higher than for the alternative type, then the average net direct fitness for carriers will also be higher than for the alternative, and vice versa.

Confusing alternative fitness accounting methods with different causal explanations may contribute to the erroneous claim that altruism can only evolve via indirect fitness ([Foster \*et al.\* 2006](#); [Lehmann & Keller 2006](#); [West \*et al.\* 2007b](#)). For instance, [West \*et al.\* \(2007b, p. 417\)](#) state that 'Direct benefits explain mutually beneficial cooperation whereas indirect benefits explain altruistic cooperation', where in this quote a distinction is being made between true 'altruistic cooperation' and 'mutually beneficial cooperation', which the authors consider non-sacrificing (i.e. weak). By contrast, we have shown here that direct fitness benefits can, and indeed must, explain the evolution of ('strong') altruistic cooperation.

This confusion appears to have its seeds in Hamilton's original papers. For example, in summarizing the basic causal explanation in inclusive fitness theory, Hamilton states: '...a gene may receive positive selection even though disadvantageous to its bearers if it causes them to confer sufficiently large advantages on relatives' ([Hamilton 1964, p. 17](#)). This quote implies that there are two types of individuals who experience two very

different fitness effects: bearers, who suffer disadvantages, and relatives, who garner advantages. But of course, the only relatives that matter when viewing this situation from the inclusive fitness perspective are those that are themselves bearers (of the altruistic gene). Moreover, the word ‘causes’ in Hamilton’s sentence could be interpreted as implying that the altruistic genotype has control over the degree to which benefits fall to other cooperators (or relatives), that is, control over its interaction environment. Clearly, this is generally not the case, and it seems preferable to consider the interaction environment to be an emergent property of the population structure, and to formulate Hamilton’s sentence in terms of direct fitness: ‘Even though disadvantageous to its bearers, a gene may receive positive selection if bearers receive sufficiently large advantages from others’.

It is possible that advocates of particular theories about the evolution of altruism will see our model as fitting squarely into their framework. What we have tried to emphasize is that there is a basic and very general requirement that underlies all these theories: an increase in the frequency of an altruistic genotype requires that carriers of the genotype are overcompensated for their altruistic sacrifice by benefits received from others. Not all carriers must help or be helped, but on average, carriers must end up with higher direct fitness benefits than carriers of alternative genotypes. This is a basic principle of natural selection and true, regardless of whether one prefers to think in terms of kin selection, multi-level selection, reciprocal altruism or other frameworks. The basic understanding that cooperators and defectors must experience different average interaction environments for altruism to evolve provides a very simple and general perspective that focuses attention on what we believe to be the central question about cooperation in nature: what are the biological mechanisms that can generate the necessary assortment between carriers of altruistic genes and the altruistic behaviour of others?

Perhaps the best-known mechanism for such assortment is ‘population viscosity’, e.g. brought about by limited dispersal. This mechanism has already been envisaged by Hamilton as a potent driver of altruism, especially when cooperation is initially rare (Hamilton 1964; Axelrod & Hamilton 1981). Based on the seminal paper by Nowak & May (1992) on the spatial Prisoner’s Dilemma, many theoretical studies have examined the role of spatial population structure for the evolution of cooperation (e.g. Nowak & Sigmund 2000; Hauert & Doebeli 2004; Lieberman *et al.* 2005). We think that the evolution of eusociality may also be influenced by this mechanism, since eusociality tends to have evolved in colony-forming species, in which interactions are highly localized (Wilson & Hölldobler 2005). Another mechanism generating the necessary assortment between altruistic genotypes and the helping behaviour of others is mediated by conditional behaviours, even in the absence of spatial structure, i.e. even with random interactions (Fletcher & Zwick 2006). For example, the famous ‘tit-for-tat’ strategy in the iterated Prisoner’s Dilemma game (Axelrod & Hamilton 1981) can be viewed as a strategy that creates assortment between cooperative genotypes and cooperative behaviour of interaction partners, because it generates cooperation as a behavioural response to cooperation, and defection as a response to defection. This assortment is essentially the

reason for the success of tit-for-tat and other, similar strategies in the iterated Prisoner’s Dilemma (e.g. Nowak & Sigmund 1992, 1993). A more recent model where conditional behaviour creates this necessary assortment involves the coevolution of choosiness and cooperation, where individuals terminate interactions if their partners are not cooperative enough (McNamara *et al.* 2008).

It seems worth noting that traditionally, evolution of cooperation due to spatial structure is put into the category of kin selection, while evolution of cooperation due to conditional behaviour such as tit-for-tat falls under the seemingly different category of reciprocal altruism. However, the concept of assortment provides a unifying perspective in which these two mechanisms can be seen as special cases of a general principle (Fletcher & Zwick 2006). A similar remark applies to interspecific cooperation, i.e. to mutualism. Traditionally, mutualism is seen as a problem that is separate from intraspecific altruism (partly owing to obvious limitations of kin selection approaches to mutualism), but again the concept of assortment between cooperators and cooperation immediately provides a unifying perspective.

In general, each of the major theories explaining the evolution of cooperation can be seen as growing out of efforts to understand the role of specific biological mechanisms that lead to assortment. For example, kin selection relies on the mechanisms of limited dispersal and kin recognition, multi-level selection on the mechanism of competition among (versus within) groups and reciprocal altruism on the mechanism of conditional behaviour. Each theory also emphasizes its own fitness accounting technique and decomposition (e.g. direct+indirect or within-group+between-group). Unnecessary disagreements arise partly because over time particular fitness accounting methods are generalized until they are seen as synonymous with the idea of assortment itself—as illustrated in the claim that altruism only evolves via indirect fitness benefits.

Of course, the method of analysing the public goods game presented here also simply presents a particular fitness accounting method, which (like the others) relies on the fundamental mechanism of assortment. However, our accounting, which is based on a fitness decomposition into ‘fitness due to self’ and ‘fitness due to the (interaction) environment’, places assortment, i.e. fitness due to the interaction environment, squarely into the centre of attention. We certainly do not claim that our approach exclusively explains the evolution of altruism in any particular instance. However, the existence of the fitness accounting method we advocate, which does not use the concepts of indirect fitness or competition among groups, makes clear that while inclusive fitness and multi-level selection theories may provide sufficient explanations for how altruism evolves, they are not necessary for understanding the evolution of altruism. Put bluntly, based on the concept of assortment, we would be able to fully understand the evolution of cooperation in a world in which the concepts of kin and group selection are absent. Focusing on the underlying role of assortment points to the biologically most relevant problem for understanding cooperation: identifying and understanding the mechanisms by which biological systems assort carriers of genotypes that induce them to help others with help from those they interact with.

We are grateful to four anonymous reviewers for their very helpful comments. J.A.F. was supported by the NSF International Fellowship Program (USA) and M.D. was supported by NSERC (Canada).

## REFERENCES

- Ackermann, M., Stecher, B., Freed, N. E., Songhet, P., Hardt, W.-D. & Doebeli, M. 2008 Self-destructive cooperation mediated by phenotypic noise. *Nature* **454**, 987–990. (doi:10.1038/nature07067)
- Axelrod, R. & Hamilton, W. D. 1981 The evolution of cooperation. *Science* **211**, 1390–1396. (doi:10.1126/science.7466396)
- Doebeli, M. & Knowlton, N. 1998 The evolution of interspecific mutualisms. *Proc. Natl Acad. Sci. USA* **95**, 8676–8680. (doi:10.1073/pnas.95.15.8676)
- Fletcher, J. A. & Zwick, M. 2006 Unifying the theories of inclusive fitness and reciprocal altruism. *Am. Nat.* **168**, 252–262. (doi:10.1086/506529)
- Fletcher, J. A. & Zwick, M. 2007 The evolution of altruism: game theory in multilevel selection and inclusive fitness. *J. Theor. Biol.* **245**, 26–36. (doi:10.1016/j.jtbi.2006.09.030)
- Fletcher, J. A., Zwick, M., Doebeli, M. & Wilson, D. S. 2006 What's wrong with inclusive fitness? *Trends Ecol. Evol.* **21**, 597–598. (doi:10.1016/j.tree.2006.08.008)
- Foster, K. R., Wenseleers, T. & Ratnieks, F. L. W. 2006 Kin selection is the key to altruism. *Trends Ecol. Evol.* **21**, 57–60. (doi:10.1016/j.tree.2005.11.020)
- Frank, S. A. 1998 *Foundations of social evolution*. Princeton, NJ: Princeton University Press.
- Hamilton, W. D. 1964 The genetical evolution of social behavior I and II. *J. Theor. Biol.* **7**, 1–52. (doi:10.1016/0022-5193(64)90038-4)
- Hamilton, W. D. 1975 Innate social aptitudes of man: an approach from evolutionary genetics. In *Biosocial anthropology* (ed. R. Fox), pp. 133–155. New York, NY: Wiley.
- Hauert, C. & Doebeli, M. 2004 Spatial structure often inhibits the evolution of cooperation in the snowdrift game. *Nature* **428**, 643–646. (doi:10.1038/nature02360)
- Kerr, B. & Godfrey-Smith, P. 2002 On Price's equation and average fitness. *Biol. Philos.* **17**, 551–565. (doi:10.1023/A:1020572704223)
- Lehmann, L. & Keller, L. 2006 The evolution of cooperation. A general framework and a classification of models. *J. Evol. Biol.* **19**, 1365–1376. (doi:10.1111/j.1420-9101.2006.01119.x)
- Lieberman, E., Hauert, C. & Nowak, M. A. 2005 Evolutionary dynamics on graphs. *Nature* **433**, 312–316. (doi:10.1038/nature03204)
- Maynard Smith, J. 1998 The origin of altruism. *Nature* **393**, 639–640. (doi:10.1038/31383)
- McNamara, J. M., Barta, Z., Fromhage, L. & Houston, A. I. 2008 The coevolution of choosiness and cooperation. *Nature* **451**, 189–192. (doi:10.1038/nature06455)
- Nowak, M. A. 2006 Five rules for the evolution of cooperation. *Science* **314**, 1560–1563. (doi:10.1126/science.1133755)
- Nowak, M. A. & May, R. M. 1992 Evolutionary games and spatial chaos. *Nature* **359**, 826–829. (doi:10.1038/359826a0)
- Nowak, M. A. & Sigmund, K. 1992 Tit for tat in heterogeneous populations. *Nature* **355**, 250–253. (doi:10.1038/355250a0)
- Nowak, M. A. & Sigmund, K. 1993 A strategy of win-stay lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game. *Nature* **364**, 56–58. (doi:10.1038/364056a0)
- Nowak, M. A. & Sigmund, K. 2000 Games on grids. In *The geometry of ecological interactions: simplifying spatial complexity* (eds U. Dieckmann, R. Law & J. A. J. Metz), pp. 135–150. Cambridge, UK: Cambridge University Press.
- Nunney, L. 1985 Group selection, altruism, and structured-deme models. *Am. Nat.* **126**, 212–230. (doi:10.1086/284410)
- Nunney, L. 2000 Altruism, benevolence and culture: commentary discussion of Sober and Wilson's 'Unto others'. *J. Conscious. Stud.* **7**, 231–236.
- Olson, M. 1971 *The logic of collective action: public goods and the theory of groups*. Cambridge, MA: Harvard University Press.
- Pepper, J. W. 2000 Relatedness in trait group models of social evolution. *J. Theor. Biol.* **206**, 355–368. (doi:10.1006/jtbi.2000.2132)
- Queller, D. C. 1985 Kinship, reciprocity and synergism in the evolution of social behaviour. *Nature* **318**, 366–367. (doi:10.1038/318366a0)
- Queller, D. C. 1992 Quantitative genetics, inclusive fitness, and group selection. *Am. Nat.* **139**, 540–558. (doi:10.1086/285343)
- Sober, E. & Wilson, D. S. 1998 *Unto others, the evolution and psychology of unselfish behavior*. Cambridge, MA: Harvard University Press.
- Taylor, P. D. & Frank, S. A. 1996 How to make a kin selection model. *J. Theor. Biol.* **180**, 27–37. (doi:10.1006/jtbi.1996.0075)
- Taylor, P. D., Wild, G. & Gardner, A. 2007 Direct fitness or inclusive fitness: how shall we model kin selection? *J. Evol. Biol.* **20**, 301–309. (doi:10.1111/j.1420-9101.2006.01196.x)
- Wade, M. J. 1980 Kin selection: its components. *Science* **210**, 665–667. (doi:10.1126/science.210.4470.665)
- West, S. A., Griffin, A. S. & Gardner, A. 2007a Evolutionary explanations for cooperation. *Curr. Biol.* **17**, R661–R672. (doi:10.1016/j.cub.2007.06.004)
- West, S. A., Griffin, A. S. & Gardner, A. 2007b Social semantics: altruism, cooperation, mutualism, strong reciprocity and group selection. *J. Evol. Biol.* **20**, 415–432. (doi:10.1111/j.1420-9101.2006.01258.x)
- West, S. A., Griffin, A. S. & Gardner, A. 2008 Social semantics: how useful has group selection been? *J. Evol. Biol.* **21**, 374–385. (doi:10.1111/j.1420-9101.2007.01458.x)
- Wilson, D. S. 1979 Structured demes and trait-group variation. *Am. Nat.* **113**, 606–610. (doi:10.1086/283417)
- Wilson, D. S. 1990 Weak altruism, strong group selection. *Oikos* **59**, 135–140. (doi:10.2307/3545133)
- Wilson, D. S. 2004 What is wrong with absolute individual fitness? *Trends Ecol. Evol.* **19**, 245–248. (doi:10.1016/j.tree.2004.02.008)
- Wilson, D. S. 2008 Social semantics: toward a genuine pluralism in the study of social behaviour. *J. Evol. Biol.* **21**, 368–373. (doi:10.1111/j.1420-9101.2008.01500.x)
- Wilson, E. O. 2005 Kin selection as the key to altruism: its rise and fall. *Soc. Res.* **72**, 159–168.
- Wilson, E. O. & Hölldobler, B. 2005 Eusociality: origin and consequences. *Proc. Natl Acad. Sci. USA* **102**, 13 367–13 371. (doi:10.1073/pnas.0505858102)