



Okasha, S., & Martens, J. (2016). Hamilton's rule, inclusive fitness maximization, and the goal of individual behaviour in symmetric two-player games. *Journal of Evolutionary Biology*, 29(3), 473-482.  
<https://doi.org/10.1111/jeb.12808>

Peer reviewed version

Link to published version (if available):  
[10.1111/jeb.12808](https://doi.org/10.1111/jeb.12808)

[Link to publication record on the Bristol Research Portal](#)  
PDF-document

This is the peer reviewed version of the following article: Okasha, S. and Martens, J. (2016), Hamilton's rule, inclusive fitness maximization, and the goal of individual behaviour in symmetric two-player games. *Journal of Evolutionary Biology*, 29: 473–482, which has been published in final form at doi: 10.1111/jeb.12808. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Self-Archiving.

## University of Bristol – Bristol Research Portal

### General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available:  
<http://www.bristol.ac.uk/red/research-policy/pure/user-guides/brp-terms/>

**Title:** Hamilton's Rule, Inclusive Fitness Maximization, and the Goal of Individual Behaviour in Symmetric Two-Player Games

**Authors:** Samir Okasha and Johannes Martens

**Affiliation:** Department of Philosophy, University of Bristol, U.K.

**Running title:** Hamilton's Rule and Goal of Behaviour

**Corresponding author:** Dr. Samir Okasha, Department of Philosophy,  
Cotham House, University of Bristol, Bristol BS6 6JL, U.K.

Tel: 0117 3310697; Fax: 0117 9546654; Email: Samir.Okasha@bristol.ac.uk

## Abstract

2 Hamilton's original work on inclusive fitness theory assumed additivity of  
costs and benefits. Recently it has been argued that an exact version of  
4 Hamilton's rule for the spread of a pro-social allele ( $rb > c$ ) holds under non-  
additive payoffs, so long as the cost and benefit terms are defined as partial  
6 regression coefficients rather than payoff parameters. This paper examines  
whether one of the key components of Hamilton's original theory can be  
8 preserved when the rule is generalized to the non-additive case in this way,  
namely that evolved organisms will behave as if trying to maximize their  
10 inclusive fitness in social encounters.

12 **Keywords:** inclusive fitness, altruism, Hamilton's rule, game theory

# 1 Introduction

14 Inclusive fitness theory is a widely-used framework for studying the evolution  
of social behaviour. Hamilton’s original formulation of the theory contains  
16 two distinct though related ideas (Hamilton 1964, 1971). The first is Hamil-  
ton’s rule, the famous criterion ( $rb > c$ ) for when an allele coding for a  
18 social behaviour will be favoured by selection. This aspect of the theory  
fits with the “gene’s eye” view of evolution. The second is maximization of  
20 inclusive fitness, rather than classical fitness, as the “goal” towards which  
an individual’s social behaviour will appear designed. This aspect fits with  
22 the traditional individualist view of evolution, and is frequently employed by  
behavioural ecologists.

24 The relation between these two aspects of inclusive fitness theory is not  
fully settled. Much theoretical work has focused solely on the first aspect;  
26 indeed the notion of individuals “trying” to maximize their inclusive fitness  
is often omitted from expositions of kin selection theory. However, recently  
28 Grafen (2006, 2009), Queller (2011) and Gardner *et al.* (2011) have argued  
for the central importance of inclusive fitness maximization as the “goal” of  
30 individual behaviour; Grafen (2006) provides a population-genetic foundation  
for the idea. This goes some way to reconciling the two aspects of Hamilton’s  
32 theory.

The work of Grafen, Queller and Gardner *et al.* suggests an intriguing  
34 link between social evolution and rational choice theory. For in effect, these

authors are arguing that inclusive fitness plays the role of a utility function in  
36 rational choice, i.e. it is the quantity that an evolved organism will behave as  
if it is trying to maximize. Thus Gardner *et al.* (2011) write: “we can imagine  
38 the individual adjusting her inclusive fitness...by altering her behaviour”,  
before choosing an action which brings maximal inclusive fitness (p.1039-  
40 40). This way of thinking about evolution is an instance of what Sober  
(1988) called “the heuristic of personification”, which says that a trait will  
42 be favoured by natural selection if and only if a rational individual, seeking to  
maximize its fitness, would choose that trait over the alternatives. In effect,  
44 Gardner *et al.* are suggesting that this heuristic is valid in social settings,  
where the trait in question is a social action, so long as “fitness” is defined  
46 as inclusive fitness.

Our aim here is to propose a particular way of formalizing this “rational  
48 actor heuristic” in the context of social evolution, and to ask how generally  
it applies. This is a pressing question because Grafen’s (2006) argument  
50 that evolution will lead to inclusive fitness maximizing behaviour assumes  
additivity of costs and benefits. This assumption is quite restrictive since in  
52 many social situations, the benefit that a given action confers on a recipient  
may depend on the recipient’s own type (Frank 1998, Lehmann and Rousset  
54 2014a). In our simple model below we find that if the additivity assumption  
is made, then the rational actor heuristic, with inclusive fitness as the indi-  
56 vidual’s utility function, applies neatly. However matters are more complex  
if there is non-additivity.

58        Asking whether the rational actor heuristic applies is different from asking  
whether Hamilton’s rule itself applies in non-additive scenarios. This latter  
60 question has been extensively discussed in the literature. The upshot is that  
an exact version of Hamilton’s rule does apply under non-additivity, so long  
62 as the cost and benefit terms are suitably defined (Queller 1992; Frank 1998,  
2013; Gardner *et al.* 2011); though the biological significance of the resulting  
64 rule has been questioned (Allen *et al.* 2013, Birch and Okasha 2015, Birch  
2015). However this does not settle the issue about individual maximization  
66 that is our focus here.

The structure of this paper is as follows. Section 2 studies social evolution  
68 using a simple additive Prisoner’s dilemma, and shows how the rational actor  
heuristic applies to it. Section 3 considers a non-additive variant of the same  
70 model and asks whether a similar conclusion holds. Section 4 discusses the  
results obtained.

## 72    **2    The case of additive payoffs**

### **2.1   Additive Prisoner’s dilemma**

74 Consider a simple model of the evolution of social behaviour of the sort used  
in evolutionary game theory. An infinite population of haploid asexual organ-  
76 isms engage in pairwise social interactions in every generation. Organisms  
are of two types, altruists ( $A$ ) and selfish ( $S$ ).  $A$  types perform an action that  
78 is costly for themselves but benefits their partner;  $S$  types do not perform

the action. Type is hard-wired genetically and perfectly inherited.

80 An organism’s payoff from the social interaction depends on its own type  
 and its partner’s type. Payoffs are interpreted as increases in lifetime repro-  
 82 ductive fitness over a unit baseline. The social action is assumed to affect  
 only the actor and their partner, thus local interaction is assumed absent.  
 84 An  $A$  type incurs a cost of  $-c$  as a result of its action and confers a benefit  
 of  $b$  on its partner, where  $c > 0$  and  $b > 0$ ; thus the game is a Prisoner’s  
 86 dilemma.

Payoffs to the actor, referred to as “personal payoffs”, are shown in Table  
 88 1. We let  $V(i, j)$  denote the payoff to an actor from playing  $i$  when her oppo-  
 nent plays  $j$ , where  $i, j \in \{A, S\}$ . Note that payoffs are additive: an altruist  
 90 alters their own payoff by  $-c$  and their partners’ payoff by  $b$ , irrespective of  
 the type of their partner.

		<b>Partner</b>	
		<b>A</b>	<b>S</b>
<b>Actor</b>	<b>A</b>	$b - c$	$-c$
	<b>S</b>	$b$	$0$

**Table 1: Additive Prisoner’s Dilemma**

92 There are three pair-types in the population,  $AA$ ,  $AS$  and  $SS$ , whose  
 relative frequencies in the initial generation are  $f_{AA}$ ,  $f_{AS}$  and  $f_{SS}$  respectively,  
 94 where  $f_{AA} + f_{AS} + f_{SS} = 1$ . The overall frequency of the  $A$  type in the initial  
 generation is denoted  $p$ , where  $p = f_{AA} + \frac{1}{2}f_{AS}$ . The change in  $p$  over one  
 96 generation is denoted  $\Delta p$ .

The sign and magnitude of  $\Delta p$  depend on the rules by which the pairs are

98 formed. If pairing is random, then the  $S$  type must be fitter overall, so  $\Delta p$   
will be negative. However if pairing is assortative then the  $A$  type may be  
100 fitter overall; for the benefits of altruistic actions then fall disproportionately  
on other altruists. Random pairing means that the frequency distribution  
102 of the pair-types will be binomial, i.e.  $f_{AA} = p^2$ ,  $f_{AS} = 2p(1 - p)$  and  
 $f_{SS} = (1 - p)^2$ .

104 Where pairing is non-random, a simple regression analysis yields a mea-  
sure of the statistical correlation between social partners. We use the variable  
106  $p_i$  to indicate an organism's own type and  $p'_i$  to indicate its partner's type;  
thus  $p_i = 1$  if the  $i^{th}$  organism is an  $A$ ,  $p_i = 0$  otherwise; and  $p'_i = 1$  if the  $i^{th}$   
108 organism is paired with an  $A$ ,  $p'_i = 0$  otherwise. We then compute the linear  
regression of  $p'_i$  on  $p_i$ , given by  $b_{p'p} = Cov(p', p)/Var(p)$ , which is a standard  
110 way of defining the  $r$  term of Hamilton's rule. Henceforth we refer to  $b_{p'p}$  as  
 $r$ .

112 In the early kin selection literature,  $r$  was often defined in genealogical  
terms, e.g. as the probability that actor and partner share an allele that  
114 is identical by descent, yielding the familiar values of  $\frac{1}{2}$  for full sibs,  $\frac{1}{2}$  for  
offspring and  $\frac{1}{4}$  for grandoffspring (see Michod and Hamilton 1980). In some  
116 ways this definition of  $r$  is the more natural one for expressing the idea that  
organisms value their relatives' reproduction in proportion to how closely  
118 related they are. However the statistical definition of  $r$ , above, yields a  
version of Hamilton's rule that is more generally applicable.

120 In the context of pairwise interactions,  $r$  can be conveniently expressed

as a difference in conditional probabilities:

$$r = Pr(\text{partner is } A \mid \text{actor is } A) - Pr(\text{partner is } A \mid \text{actor is } S)$$

It follows that  $r$  ranges from  $-1$  (perfect disassortment) to  $1$  (perfect assortment); when pairing is random,  $r = 0$ .

## 2.2 Evolutionary analysis

In the Appendix, we show that the change in  $p$  over one generation is given by:

$$\Delta p = (rb - c) \cdot Var(p) / \bar{w} \tag{1}$$

where  $\bar{w}$  is average population fitness. Since  $Var(p)$  is non-negative, this tells us that so long as  $0 < p < 1$ , the  $A$  type will increase in frequency in the population whenever  $rb > c$ , which is of course Hamilton's rule. Since  $b$  and  $c$  are fixed parameters of the payoff matrix, this condition for the spread of  $A$  is frequency-independent so long as  $r$  itself does not change as the population evolves. Constancy of  $r$  across generations will sometimes be a reasonable assumption, for the pattern of assortment in the population, which is what  $r$  measures, may be determined by biological factors, e.g. dispersal, which are independent of the social trait that is evolving. This assumption is considered further in the Discussion section.

With the constant  $r$  assumption, the outcome of the evolutionary process is easily determined. If  $rb > c$  the  $A$  type will spread to fixation; if  $rb < c$  the

140  $S$  type will spread to fixation; if  $rb = c$  there will be no evolutionary change.

### 2.3 Rational actor analysis: preliminaries

142 To apply a rational actor analysis, we transpose our evolutionary model to a  
rational choice context. We consider two players playing a symmetric game.  
144 Each player has two pure strategies,  $A$  and  $S$ . If a player plays a *mixed*  
*strategy* this means that they randomize over their pure strategies; thus  $\pi_A$   
146 denotes the mixed strategy in which  $A$  is played with probability  $\pi_A$  and  $S$   
with probability  $1 - \pi_A$ . The payoff to a mixed strategy is then simply its  
148 expected payoff.

Each player has a utility function which measures how desirable they  
150 find the possible outcomes of the game; we assume that both players have  
the same utility function. Each player's goal is to maximize their utility  
152 function. One possibility is that the utility function is given by the personal  
payoffs in Table 1 above, in which case we write  $U(i, j) = V(i, j)$ , where  
154  $U(i, j)$  is the utility a player gets from playing  $i$  when their partner plays  $j$ ;  
 $i, j \in \{A, S\}$ . There are other possibilities too; see below.

156 Once the players' utility function has been specified, the next step is to  
seek the Nash equilibrium (or equilibria) of the game. (A Nash equilibrium  
158 is a pair of strategies, possibly mixed, one for each player, each of which is  
a best response to the other.) Game theory predicts that if the players are  
160 rational, they will end up at a Nash equilibrium of the game (see for example  
Binmore 2007). We can then ask whether the Nash equilibria of the game

162 correspond to the outcomes of the evolutionary process described above. If  
 163 so, we can conclude that evolution will lead organisms to behave as if trying  
 164 to maximize the utility function in question.

This is a natural way of formalizing the rational actor heuristic in a game-  
 166 theoretic context. It differs somewhat from Grafen’s (2006) formalization of  
 the same idea, which posits “links” between gene-frequency change and indi-  
 168 vidual optimization. Our approach allows recovery of Grafen’s main result;  
 and by taking optimization to include best-response, i.e. optimal choice con-  
 170 ditional on the other player’s choice, extends easily to the non-additive case.  
 A similar approach is found in Alger and Weibull (2012) and Lehmann *et al.*  
 172 (2015).

### 2.3.1 Utility as inclusive fitness

174 One possibility to explore is that a player’s utility function depends on their  
 partner’s payoff as well as their own. For example, suppose that a player’s  
 176 utility for any outcome is given by the quantity: personal payoff plus  $r$  times  
 partner’s payoff, i.e.  $U(i, j) = V(i, j) + rV(j, i)$ . Applying this transforma-  
 178 tion to the personal payoffs yields Table 2 below, which we refer to as the  
 “inclusive fitness payoff matrix”.

		<b>Partner</b>	
		<b>A</b>	<b>S</b>
<b>Actor</b>	<b>A</b>	$(b - c)(r + 1)$	$-c + rb$
	<b>S</b>	$b - rc$	$0$

**Table 2: Additive PD with inclusive fitness payoffs**

180 This transformation was first suggested by Hamilton (1971), and has been  
discussed by Grafen (1979), Bergstrom (1995), Wade and Breden (1980), Day  
182 and Taylor (1998), Taylor and Nowak (2007) and Martens (2015). It is a  
natural formalization of the idea that an actor, in their social behaviour, will  
184 care about their partner’s payoff, discounted by relatedness, as well as their  
own payoff. Clearly, other transformations of the personal payoff matrix are  
186 also conceivable.

The payoffs in Table 2 do not correspond exactly to the verbal definition  
188 of inclusive fitness in Hamilton (1964), which was: “the personal fitness  
which an individual actually expresses...once it is stripped of all components  
190 which can be considered as due to the individual’s social environment...then  
augmented by certain fractions of the quantities of harm and benefit which  
192 the individual himself causes to the fitnesses of his neighbours...The fractions  
in question are simply the coefficients of relationship” (p. 8). This definition  
194 is sometimes but not always adhered in the literature.

The discrepancy between Table 2 and Hamilton’s definition arises because  
196 in the left column, the actor’s payoff has not been stripped of the component  
that is due to the partner’s altruistic action ( $b$ ), and has been augmented  
198 by  $r$  times the partner’s entire payoff, rather than the portion of that payoff  
that is caused by the actor ( $b$ .) Applying Hamilton’s definition exactly would  
200 lead to the payoff matrix in Table 3 below.

Note that Table 3 derives from Table 2 by subtraction of the quantity  
202  $b - rc$  from the left column. In game-theoretical terms, Table 3 is thus

		Partner	
		A	S
Actor	A	$-c + rb$	$-c + rb$
	S	0	0

**Table 3: Additive PD with Hamilton (1964) payoffs**

a "local shift" of Table 2 (and vice-versa), which means that their Nash  
 204 equilibria are necessarily identical (Weibull 1995). Therefore, if the players'  
 utility function is given by Table 3, game theory predicts exactly the same  
 206 outcome(s) as if it were given by Table 2. So although taking Table 2 as  
 the definition of inclusive fitness involves an element of "double counting" –  
 208 which Hamilton's definition was designed to avoid – it is harmless.

In fact there is a positive reason to prefer Table 2 as the definition of  
 210 inclusive fitness, in a game-theoretic context. For Hamilton's definition does  
 not generalize easily to non-additive payoffs. With non-additivity it is un-  
 212 clear how to decide which component of the actor's payoff is "caused" by  
 its partner's action and vice-versa (cf. Allen *et al.* 2013). By contrast, the  
 214 definition used in Table 2 – actor payoff plus  $r$  times partner payoff – ap-  
 plies just as well to the non-additive case. In order not to prejudge the issue  
 216 of whether inclusive fitness maximization, or a similar result, obtains under  
 non-additivity, this is the definition preferred here.

## 218 **2.4 Rational actor analysis: results**

Suppose firstly that the utility function is personal payoff (Table 1). It is  
220 easy to see that  $(S, S)$  is the only Nash equilibrium of the game, since  $S$   
strongly dominates  $A$ , i.e. each player does strictly better by playing  $S$   
222 irrespective of their partner's choice. This familiar result shows that the  
rational actor heuristic fails for this choice of utility function, since it would  
224 have us conclude that altruism can never evolve, which we know to be false.

What if the utility function is inclusive fitness payoff (Table 2)? In that  
226 case, we can show the following. If  $rb > c$  then  $(A, A)$  is the unique Nash  
equilibrium; if  $rb < c$  then  $(S, S)$  is the unique Nash equilibrium; if  $rb = c$   
228 then  $(A, A)$  and  $(S, S)$  are both Nash equilibria, as is every pair of mixed  
strategies, so game theory makes no prediction about the players' choices  
230 (See Appendix for proof).

It follows that with additive payoffs, defining utility as inclusive fitness  
232 makes the rational actor heuristic valid. The condition for the  $A$  type to  
evolve,  $rb > c$ , is identical to the condition for  $(A, A)$  to be the unique Nash  
234 equilibrium of the rational game; and similarly for  $S$  (Table 4). This supports  
the idea that evolution will lead organisms to appear as if trying to maximize  
236 their inclusive fitness, just as Hamilton originally argued.

An equivalent perspective on the situation is this. The quantity  $(rb - c)$   
238 equals the difference in a player's inclusive fitness payoff between playing  
 $A$  and  $S$ , irrespective of what its partner does (see Table 2). Thus we can

$rb > c$	$\iff$	$A$ evolves	$\iff$	$(A, A)$ is unique Nash equilibrium
$rb < c$	$\iff$	$S$ evolves	$\iff$	$(S, S)$ is unique Nash equilibrium
$rb = c$	$\iff$	no evolution	$\iff$	all pairs of strategies, pure and mixed, are Nash equilibria

Note:  $\iff$  means “if and only if”

**Table 4: Rational actor heuristic with utility = inclusive fitness**

240 determine whether the  $A$  type will evolve by asking whether a rational agent,  
 who wants to maximize their inclusive fitness, would choose  $A$  over  $S$ . In  
 242 short, equating utility with inclusive fitness ensures that the rational agent’s  
 choice coincides with the “choice” made by natural selection.

## 244 2.5 A caveat: uniqueness

One important caveat is needed. In the above model, the inclusive fitness  
 246 payoff matrix (Table 2) is not the unique utility function that yields the  
 $rb > c$  condition for action  $A$  to be chosen over  $S$ . In game theory, the utility  
 248 function is only ever unique up to choice of origin and unit; so any affine  
 transformation (of the form  $U' = aU + b$ , where  $a, b \in \mathbb{R}$ ,  $a > 0$ ) will leave  
 250 all Nash equilibria of the game unchanged. Furthermore, a “local shift” of  
 the utility function, which involves adding a constant to any column of the  
 252 utility matrix, will also leave unchanged the Nash equilibria, as noted above.

One local shift of the inclusive fitness payoff matrix (Table 2) is of par-  
 254 ticular interest. If we add the quantity  $(rc - rb)$  to the left-hand column of  
 Table 2, we get the matrix in Table 5 below.

		<b>Partner</b>	
		<b>A</b>	<b>S</b>
<b>Actor</b>	<b>A</b>	$(b - c)$	$-c + rb$
	<b>S</b>	$(1 - r)b$	$0$

**Table 5: Additive PD with Grafen 1979 payoffs**

256 The payoffs in Table 5 are related to the personal payoffs (Table 1) by  
the transformation  $U(i, j) = rV(i, i) + (1 - r)V(i, j)$ . This transformation  
258 was first suggested by Grafen (1979), hence the label ‘Grafen 1979 payoff’;  
see Bergstrom (1995), Day and Taylor (1998), Alger and Weibull (2012) for  
260 discussion. By contrast with the inclusive fitness payoffs (Table 2), which  
involve adding  $r$  times partner’s payoff to the actor’s personal payoff, the  
262 Grafen 1979 payoffs involve taking an  $(r, 1 - r)$  weighted average of the  
personal payoff that would accrue to the actor if their partner had chosen  
264 the same as the actor and if their partner made the choice that they actually  
did.

266 Since the Grafen 1979 payoff matrix (Table 5) is a local shift of the in-  
clusive fitness payoff matrix (Table 2), the Nash equilibria of the resulting  
268 games are identical; thus the rational actor heuristic works equally well with  
either. (This is because in both cases,  $(rb - c)$  is the payoff difference be-  
270 tween playing  $A$  and  $S$ .) Therefore while our simple model has vindicated  
Hamilton’s claim that evolution will lead organisms to behave as if trying to  
272 maximize their inclusive fitness, it is important to see that inclusive fitness  
(whether defined our way (Table 2) or in Hamilton’s original way (Table 3)),  
274 is not the unique quantity of which this maximization claim is true.

### 3 Non-additive payoffs

276 To determine whether the above results generalize to the non-additive case,  
we consider a modified Prisoner’s dilemma in which the payoff to an  $A$  type  
278 paired with another  $A$  type is  $(b - c + d)$  rather than  $(b - c)$ . So the parameter  
 $d$  quantifies the deviation from payoff additivity, or synergistic effect, when  
280 two  $A$  types are paired together;  $d$  can be either positive or negative. The  
resulting payoff structure (Table 6) is sometimes referred to as a ‘synergy  
282 game’ (van Veelen 2009).

		Partner	
		A	S
Actor	A	$b - c + d$	$-c$
	S	$b$	$0$

**Table 6: Non-additive Prisoner’s dilemma (‘synergy game’)**

Again, we assume that pairs of organisms are drawn from an infinite  
284 population to play the game; type is genetically hard-wired and mutation is  
absent.

#### 286 3.1 Evolutionary analysis

As before,  $\Delta p$  denotes the change in frequency of the  $A$  type over a genera-  
288 tion. Unsurprisingly,  $rb > c$  is no longer the condition for  $\Delta p$  to be positive.  
However an exact version of Hamilton’s rule can be recovered by suitably  
290 defining the cost and benefit terms, as emphasized by Gardner *et al.* (2011),  
whose approach we follow here. (A different approach, not discussed here,

292 incorporates non-additive payoffs into Hamilton’s rule by a weak selection  
approximation; see for example Lehmann and Rousset 2014b).

294 For each individual  $i$ , we let  $w_i$  denote its actual reproductive fitness  
(number of offspring). We then write  $w_i$  as a linear regression on  $p_i$  and  $p'_i$ :

$$w_i = \alpha + b_{wp.p'}p_i + b_{wp'.p}p'_i + e_i \quad (2)$$

296 where  $\alpha$  is baseline fitness;  $b_{wp.p'}$  is the partial regression of an individual’s  
fitness on their own type, controlling for their partner’s type;  $b_{wp'.p}$  is the  
298 partial regression of an individual’s fitness on their partner’s type, controlling  
for their own type; and  $e_i$  is the residual. These partial regression coefficients  
300 quantify the average effect (*sensu* Fisher 1930) of the actor’s action, and their  
partner’s action’s, on the actor’s fitness.

302 Following Hamilton (1964), instead of considering the effect on the actor’s  
fitness of their partner’s action  $b_{wp'.p}$ , we can consider the effect on their  
304 partner’s fitness of the actor’s action, denoted  $b_{wp.p'}$ . These two partial  
regression coefficients are numerically identical (Taylor *et al.* 2007). (This  
306 is the well-known switch from ‘neighbour-modulated’ to ‘inclusive’ fitness.)  
Following Gardner *et al.* (2011), we denote the  $b_{wp.p'}$  and  $b_{wp'.p}$  coefficients  
308 as  $-C$  and  $B$  respectively.

Importantly, equation (2) can be fitted whether or not the true relation  
310 between  $w$ ,  $p_i$  and  $p'_i$  is linear. In the non-additive case under considera-  
tion that relation is non-linear (since  $d > 0$ ), which implies that the partial  
312 regression coefficients  $-C$  and  $B$  will be functions of population-wide gene

frequencies, and liable to change as the population evolves. Therefore un-  
 314 like  $c$  and  $b$ , which are fixed payoff parameters,  $-C$  and  $B$  are population  
 variables.

316 Following Gardner *et al.* (2007, p. 219), we can write explicit expressions  
 for  $-C$  and  $B$  in terms of  $r$ ,  $p$ , and the parameters of the payoff matrix  $b$ ,  $c$   
 318 and  $d$ . This yields:

$$-C = (-c) + (d) \cdot [r + p(1 - r)]/[1 + r] \quad (3)$$

$$B = (b) + (d) \cdot [r + p(1 - r)]/[1 + r] \quad (4)$$

320 We can then derive the following expression for evolutionary change:

$$\Delta p = (rB - C) \cdot \text{Var}(p)/\bar{w} \quad (5)$$

where  $\bar{w}$  is average population fitness (see Appendix). Equation (5) tells us  
 322 that when  $0 < p < 1$ , the  $A$  type will increase in frequency if and only if  
 $rB > C$ . This is a generalized version of Hamilton's rule, applicable whether  
 324 payoffs are additive or not.

The quantity  $(rB - C)$ , whose sign determines whether altruism spreads,  
 326 can be computed by adding equation (3) to  $r$  times equation (4). After  
 simplifying this yields:

$$rB - C = (rb - c) + d[r + p(1 - r)] \quad (6)$$

328 Note that  $(rB - C)$  is a function of  $p$ , so satisfaction of  $rB > C$  in gen-

eration  $t$  does not imply its satisfaction in generation  $t + 1$ . Selection is thus  
 330 frequency-dependent, and neither type will necessarily spread to fixation. A  
 polymorphic equilibrium will obtain when  $p = [c - r(b + d)]/d[1 - r]$ ; the  
 332 stability of this equilibrium depends on the sign of  $d$ . The full evolutionary  
 dynamics are summarized in Table 7 below; see Appendix for proof.

**Case 1:  $r < 1, d > 0$**

(i) $rb - c + rd \geq 0$	$A$ evolves to fixation
(ii) $rb - c + d \leq 0$	$S$ evolves to fixation
(iii) $rb - c + d > 0 > rb - c + rd$	unstable polymorphism at $p = [c - r(b + d)]/d[1 - r]$

**Case 2:  $r < 1, d < 0$**

(i) $rb - c + d \geq 0$	$A$ evolves to fixation
(ii) $rb - c + rd \leq 0$	$S$ evolves to fixation
(iii) $rb - c + d < 0 < rb - c + rd$	stable polymorphism at $p = [c - r(b + d)]/d[1 - r]$

**Case 3:  $r = 1$**

(i) $b - c + d > 0$	$A$ evolves to fixation
(ii) $b - c + d < 0$	$S$ evolves to fixation
(iii) $b - c + d = 0$	no evolutionary change

**Table 7: Evolutionary dynamics of non-additive PD**

334 The general version of Hamilton’s rule embodied in equation (5) raises  
 interesting interpretive questions. Some have argued that the rule in this  
 336 form has little explanatory value (Nowak *et al.* 2011, Allen *et al.* 2013);  
 while others have seen the generality of the rule as an advantage, a proof that  
 338 inclusive fitness theory does not rely on restrictive assumptions (Gardner *et al.*  
*et al.* 2011). This debate has been analyzed elsewhere (Birch 2015, Birch and  
 340 Okasha 2015) and is not the focus here.

Instead our question is this. Given that equation (5) is true, and given the  
342 resulting evolutionary dynamics, can the rational actor heuristic be applied?  
Will evolution lead organisms to behave as if maximizing a utility function,  
344 and if so what is it?

Importantly, the answer to this question cannot simply be read off equa-  
346 tion (5). In the additive case there was a simple link between Hamilton’s rule  
and a utility function with the desired property:  $rb - c > 0$  was the condition  
348 for the  $A$  type to spread, and  $(rb - c)$  the utility difference between playing  
 $A$  and  $S$ . One might hope to extrapolate this to the non-additive case by  
350 simply replacing  $(rb - c)$  with  $(rB - C)$  in Table 2. However since  $B$  and  $C$   
are functions of  $p$ , they cannot meaningfully feature in the utility function.

352 The reason is as follows. The point of the rational actor heuristic is to  
find a link between gene-frequency dynamics and a “goal” that organisms  
354 behave as if they are trying to achieve. Such a link would be trivial if the  
“goal” were allowed to change as gene frequencies change. For the heuristic  
356 to have any value, the goal must remain fixed. So our task is to find a utility  
function *whose arguments are restricted to the payoff parameters ( $b$ ,  $c$  and*  
358  *$d$ ), and the relatedness coefficient  $r$* , which makes the rational actor heuristic  
work.

### 360 **3.2 Rational actor analysis**

To address this question, we again transpose the evolutionary model to a  
362 rational choice context and study the Nash equilibria of the resulting game.

Suppose firstly that the utility function is given by the inclusive fitness payoff transformation, i.e. personal payoff plus  $r$  times partner payoff. This yields the payoffs in Table 8 below.

		Partner	
		A	S
Actor	A	$(b - c + d)(r + 1)$	$-c + rb$
	S	$b - rc$	0

**Table 8: Non-additive PD with inclusive fitness payoffs**

The Nash equilibria are then as follows:

$(A, A)$  is a Nash equilibrium if and only if  $rb - c + d(r + 1) \geq 0$

$(S, S)$  is a Nash equilibrium if and only if  $rb - c \leq 0$

$(\pi_A, \pi_A)$  is a mixed strategy Nash equilibrium where  $\pi_A = (c - rb)/d(1 + r)$ , so long as  $0 < \pi_A < 1$ .

It follows that, unlike in the additive case, the rational actor heuristic does not work when utility is defined as inclusive fitness. The condition for  $(A, A)$  to be a Nash equilibrium is not identical to the condition for  $A$  to evolve to fixation; similarly for  $S$ . Furthermore, the condition for there to be a mixed-strategy Nash equilibrium is not the same as the condition for there to be a polymorphism. So it is not true that at evolutionary equilibrium, organisms will behave as if trying to maximize their inclusive fitness.

Can we find a utility function *modulo* which the rational actor heuristic works? The answer is yes. The Grafen 1979 payoff matrix, which to recall

376 is derived from the personal payoff matrix by the transformation  $U(i, j) =$   
 $rV(i, i) + (1 - r)V(i, j)$ , does the trick. This yields the payoffs in Table 9  
 378 below.

		<b>Partner</b>	
		<b>A</b>	<b>S</b>
<b>Actor</b>	<b>A</b>	$(b - c + d)$	$-c + rb + rd$
	<b>S</b>	$(1 - r)b$	$0$

**Table 9: Non-additive PD with Grafen 1979 payoffs**

The Nash equilibria are then as follows:

$(A, A)$  is a Nash equilibrium if and only if  $rb - c + d \geq 0$

$(S, S)$  is a Nash equilibrium if and only if  $rb - c + rd \leq 0$

$(\pi_A, \pi_A)$  is a mixed strategy Nash equilibrium, where  $\pi_A$   
 $= (c - r(b + d))/d(1 - r)$ , so long as  $0 < \pi_A < 1$ .

380 This restores the rational actor heuristic. In particular, if  $(A, A)$  is the  
 only pure-strategy Nash equilibrium, then  $A$  evolves to fixation; if  $(S, S)$   
 382 is the only pure-strategy equilibrium, then  $S$  evolves to fixation. If there  
 is a mixed-strategy Nash equilibrium but no pure strategy equilibria, the  
 384 population evolves to a stable polymorphism; if there is a mixed-strategy  
 Nash equilibrium and both  $(A, A)$  and  $(S, S)$  are pure-strategy equilibria,  
 386 then there is an unstable polymorphism; in both cases, the weights on  $A$  and  
 $S$  in the mixed-strategy Nash equilibrium equal the proportions of  $A$  and  
 388  $S$  in the polymorphism. Thus there is a tight correspondence between the

Nash equilibria and the evolutionary dynamics, summarized in Table 10 (see  
 390 Appendix for proof)

$(A, A)$ is only pure N.E.	$\implies$	$A$ evolves to fixation
$(S, S)$ is only pure N.E.	$\implies$	$S$ evolves to fixation
$(\pi_A, \pi_A)$ is only N.E.	$\iff$	stable polymorphism at $p = \pi_A$
$(\pi_A, \pi_A), (A, A), (S, S)$ all N.E.	$\iff$	unstable polymorphism at $p = \pi_A$

Note:  $\pi_A = (c - r(b + d)/d(1 - r))$

**Table 10: Rational actor heuristic, utility = Grafen 1979 payoff**

The upshot is that with non-additive payoffs, the rational actor heuristic  
 392 will work so long as the utility function is defined as Grafen 1979 payoff,  
 rather than inclusive fitness payoff. Again any affine transformation of the  
 394 Grafen 1979 payoff matrix, or any local shift, will also preserve the correspon-  
 dences above. Note that, unlike in the additive case, the Grafen 1979 payoff  
 396 matrix (Table 9) is *not* a local shift of the inclusive fitness payoff matrix  
 (Table 8). This is why the rational actor heuristic fails if utility is defined as  
 398 inclusive fitness in the non-additive case.

## 4 Discussion

400 Hamilton's original formulation of inclusive fitness theory assumed additivity  
 of costs and benefits. A number of authors have emphasized that an exact  
 402 version of Hamilton's rule holds with non-additive payoffs, so long as the  $-C$   
 and  $B$  terms are suitably defined. Here we have focused on the relevance of  
 404 payoff additivity not for Hamilton's rule itself, but for Hamilton's (logically

distinct) claim that evolution will lead organisms to behave as if trying to  
406 maximize their inclusive fitness, understood here to mean personal payoff  
plus  $r$  times partner payoff.

408 In a recent critique, Allen *et al.* (2013) observe that arguments for inclu-  
sive fitness maximization all rely on payoff additivity, and that where selec-  
410 tion is frequency-dependent, fitness maximization need not generally occur.  
They write: “evolution does not, in general, lead to the maximization of  
412 inclusive fitness or any other quantity” (p. 20138).

Our analysis partly supports this conclusion. Here we have understood  
414 maximization to include best-response, so that the presence of frequency-  
dependence does not automatically preclude a maximization principle from  
416 holding; and we have allowed the utility function to be any function of the  
payoff parameters  $b$ ,  $c$  and  $d$  and the relatedness coefficient  $r$ . At the evo-  
418 lutionary equilibrium of our simple non-additive model, it is not true that  
organisms behave as if trying to maximize their inclusive fitness payoff. How-  
420 ever there is a somewhat similar quantity – Grafen 1979 payoff – that organ-  
isms do behave as if they are trying to maximize.

422 It is an open question whether our positive result – maximization of  
Grafen 1979 payoff – extends to more complicated models of social evolu-  
424 tion, e.g. that incorporate local interaction, multiple social partners, or class  
structure, or to more realistic genetic architectures than haploid inheritance.  
426 There is no guarantee that it does, as such models typically lead to more  
complicated evolutionary dynamics than those assumed here. As has been

428 emphasized before, a valid maximization argument must always deduce the  
quantity being maximized, if any, from the underlying evolutionary dynamics  
430 (Mylius and Diekmann 1995).

Also, we have assumed that the coefficient of relatedness,  $r$ , remains con-  
432 stant as the population evolves. Without this assumption, it makes little  
sense to allow the utility function to depend on  $r$ , as this would be tanta-  
434 mount to positing a changing “goal” so would again trivialize the rational  
actor heuristic. In some inclusive fitness models,  $r$  is in fact a dynamic vari-  
436 able rather than a constant (e.g. van Baalen and Rand 1998), so it cannot be  
assumed that our results, or ones like them, can be derived for these models.

438 Our negative result, that maximization of inclusive fitness only holds  
with additive payoffs, is in line with previous results by Bergstrom (1995)  
440 and Lehmann and Rousset (2014a); it supports some of the claims made  
by opponents of inclusive fitness theory such as Allen and Nowak (2015).  
442 The key logical point to note is that although a version of Hamilton’s rule is  
indeed a fully general evolutionary principle, as Gardner *et al.* (2011) stress,  
444 no principle about individual maximization can be deduced directly from this  
form of the rule. Whether such a principle holds, and if so what the quantity  
446 being maximized is, needs to be shown on a case-by-case basis.

Finally, what are the implications for biological practice? Behavioural  
448 ecologists have often used inclusive fitness maximization as a way to interpret  
observed behaviour in the field, in line with Hamilton’s original suggestion.  
450 Our analysis suggests that this will not always be possible. If an observed

social behaviour fails to maximize an individual's inclusive fitness, defined as  
452 personal payoff plus  $r$  times partner's payoff, the behaviour may nonetheless  
be adaptive and the population at an evolutionary equilibrium. Moreover  
454 the quantity we have called "Grafen 1979 payoff" will serve the needs of the  
behavioural ecologist seeking to identify the "goal" of evolved behaviour in  
456 a broader range of cases than will inclusive fitness itself.

## Acknowledgements

458 Thanks to Ken Binmore, Bengt Autzen, Jonathan Birch, Steve Frank, Herb  
Gintis, Alan Grafen, Andy Gardner and two anonymous referees for their  
460 comments and discussion. This work was supported by the European Re-  
search Council Seventh Framework Program (FP7/20072013), ERC Grant  
462 agreement no. 295449.

## References

- 464 Alger, I. & Weibull, J. W. 2012. A generalization of Hamilton's rule—love others  
how much? *J. Theor. Biol.* **299**: 42-54.
- 466 Allen B., & Nowak M.A. 2015. Games among relatives revisited. *J. Theor. Biol.*  
**194**: 391-407.
- 468 Allen B., Nowak M.A., & Wilson, E.O. 2013. Limitations of inclusive fitness.  
*Proc. Natl. Acad. Sci. USA* **110**: 21035-20139.
- 470 Bergstrom, T. 1995. On the evolution of altruistic ethical rules for siblings. *Am.*  
*Econ. Rev.* **85**: 58-81.
- 472 Binmore, K. 2007. *Playing for Real*. Oxford University Press, Oxford.
- Birch J. 2015. Hamilton's rule and its discontents *Brit. J. Philos. Sci.*, in press.
- 474 Birch J. & Okasha, S. 2015. Kin selection and its critics. *Bioscience* **65**: 22-32.
- Day, T. & Taylor, P.D. 1998. Unifying genetic and game theoretic models of kin  
476 selection for continuous traits. *J. Theor. Biol.* **194**: 391-407.
- Fisher R.A. 1930. *The Genetical Theory of Natural Selection*. Clarendon Press,  
478 Oxford.
- Frank S.A. 1998. *Foundations of Social Evolution*. Princeton University Press,  
480 Princeton, New Jersey.
- Frank S.A. 2013. Natural selection VII. History and interpretation of kin  
482 selection theory. *J. Evol. Biol.* **26**: 1151-1184.
- Gardner A., West S.A., & Barton, N. 2007. The relation between multilocus

- 484 population genetics and social evolution theory. *Am. Nat.* **169**: 207-226.
- Gardner A., West S.A., & Wild, G. 2011. The genetical theory of kin selection.  
486 *J. Evol. Biol.* **24**: 1020-1043.
- Grafen A. 1979. The hawk-dove game played between relatives. *Anim. Behav.*  
488 **27**: 905-907.
- Grafen A. 2006. Optimization of inclusive fitness. *J. Theor. Biol.* **238**: 541-563.
- 490 Grafen A. 2009. Formalizing Darwinism and inclusive fitness theory. *Phil. Trans.  
R. Soc. B*: **364**: 3135-3141.
- 492 Hamilton W.D. 1964. The genetical evolution of social behaviour. *J. Theor.  
Biol.* **7**: 1-52.
- 494 Hamilton W.D. 1971. Selection of selfish and altruistic behaviour in some  
extreme models. In: *Narrow Roads of Gene Land Volume 1*, pp. 198-228. W. H.  
496 Freeman, New York.
- Lehmann L., Alger, I. and Weibull, J. 2015. Does evolution lead to maximizing  
498 behaviour. *Evolution* **69**: 1858-1873.
- Lehmann L., & Rousset, F. 2014a. Fitness, inclusive fitness and optimization.  
500 *Biol. Philos.* **29**: 181-195.
- Lehmann L., & Rousset, F. 2014b. The genetical theory of social behaviour.  
502 *Phil. Trans. R. Soc. B* **369**: 20130357.
- Martens, J. 2015. Hamilton meets causal decision theory. *Brit. J. Philos. Sci.*, in  
504 press.
- Michod, R. & Hamilton, W.D. 1980. Coefficients of relationship in sociobiology.

- 506 *Nature* **288**: 694-697.
- Mylius, S.D., & Diekmann, O. 1995. On evolutionarily stable life histories,  
508 optimization and the need to be specific about density dependence. *Oikos* **74**:  
218-224.
- 510 Nowak, M.A. & Tarnita, C.E. & Wilson, E.O. 2011. Nowak et. al. reply. *Nature*  
**471**: E9-E10.
- 512 Price, G.R. 1970. Selection and covariance. *Nature* **227**: 520-521.
- Queller D.C. 1992. A general model for kin selection. *Evolution* **46**: 376-380.
- 514 Queller D.C. 2011. Expanded social fitness and Hamilton's rule for kin, kith and  
kind. *Proc. Natl. Acad. Sci. USA* **108**: 10792-10799.
- 516 Sober, E. 1988. Three differences between evolution and deliberation. In:  
*Modeling Rationality, Morality and Evolution* (P. Danielson, ed.), pp. 408-422.  
518 Oxford University Press, Oxford.
- Taylor C., & Nowak, M.A. 2007. Transforming the dilemma. *Evolution* **61**:  
520 2281-2292.
- Taylor P.D., Wild, G. & Gardner, A. 2007. Direct fitness or inclusive fitness:  
522 how shall we model kin selection? *J. Evol. Biol.* **20**: 301-309.
- van Baalen, M. & Rand, D.A. 2007. The unit of selection in viscous populations  
524 and the evolution of altruism. *J. Theor. Biol.* **193**: 631-648.
- van Veelen M. 2009. Group selection, kin selection, altruism and cooperation:  
526 when inclusive fitness is right and when it can be wrong. *J. Theor. Biol.* **259**:  
589-600.

- 528 Wade, M., & Breden, F. 1980. The evolution of cheating and selfish behaviour.  
*Behav. Ecol. Sociobiol.* **7**: 167-72.
- 530 Weibull J. 1995. *Evolutionary Game Theory*. MIT Press, Massachusetts.