# Inclusive Fitness and the Problem of Honest Communication

Bruner, Justin P.; Rubin, Hannah

*Published in:*
British Journal for the Philosophy of Science

*DOI:*
10.1093/bjps/axy028



*Publication date:*
2020

*Brit. J. Phil. Sci.* **0** (2018), 1–23



# Inclusive Fitness and the Problem of Honest Communication

Justin P. Bruner and Hannah Rubin

ABSTRACT

Inclusive fitness has been under intense scrutiny in recent years, with many critics claiming the framework leads to incorrect predictions. We consider one particularly influential heuristic for estimating inclusive fitness in the context of the very case that motivated reliance on it to begin with: the Sir Philip Sidney signalling game played with relatives. Using a neighbour-modulated fitness model, we show when and why this heuristic is problematic. We argue that reliance on the heuristic rests on a misunderstanding of what it means for two organisms to be related and perpetuates a mischaracterization of the role of the 'relatedness' parameter in inclusive fitness.



## 1 Introduction

Inclusive fitness has been under intense debate recently, following an article by (Nowak *et al.* [2010]). Among the many criticisms levied against the inclusive fitness framework, some have claimed the framework yields incorrect predictions (van Veelen [2009]; Nowak *et al.* [2010], [2011]). One response to criticisms of the inclusive fitness framework is that the authors confuse the framework as a whole with models of particular cases (Abbot *et al.* [2011]). However, it is more than just particular models that should give cause for concern. As we will show, one particularly influential method for estimating inclusive fitness is unreliable in that it yields predictions that are drastically different from explicit fitness calculations.


doi:10.1093/bjps/axy028

Maynard Smith ([1991]) famously used a heuristic method of calculating inclusive fitness to better explain the existence of honest communication among relatives with conflicting interests. Conventional wisdom holds that honest communication is possible when there are significant costs associated with signalling.[1] Unfortunately, this formulation of the so-called handicap principle has a rather peculiar status, for while it plays a central role in animal communication theory, there is surprisingly little empirical evidence that the high costs needed to sustain communication exist in nature (McCarty [1996]; Bachman and Chappell [1998]; Zollman [2013]). Maynard Smith's alternative way of resolving this puzzle is to notice that when senders and receivers are related, the conflict of interest preventing honest communication can be minimized. We show that the way Maynard Smith ([1991]) incorporates relatedness into inclusive fitness calculations, which has become the commonly used method in the literature (Johnstone and Grafen [1992]; Godfray [1995]; Bergstrom and Lachmann [1997]; Johnstone [1998]; Brilot and Johnstone [2003]; Huttegger and Zollman [2010]), misrepresents the biological reality of many of these interactions among kin.

We argue that reliance on this heuristic both rests upon and reinforces a misunderstanding about what it means for two organisms to be related. This is particularly significant since inclusive fitness as a framework for studying evolution has generated a variety of insights and is seen as indispensable by evolutionary theorists and field biologists (Queller [1992]; Abbot *et al.* [2011]; West and Gardner [2013]). Given its central status in evolutionary theory and the recent attention paid to debating its utility, it is of the utmost importance that techniques used within the inclusive fitness framework are methodologically and conceptually sound. We develop a model that correctly represents relatedness and show how the heuristic leads to different conclusions about the possibility of communication in the very case that motivated reliance on the heuristic to begin with.

Our article will proceed as follows: We first discuss Maynard Smith's heuristic approach to measuring inclusive fitness.[2] In Section 3, we introduce the Sir Philip Sidney game. We then develop our own model that uses so-called neighbour-modulated fitness in Section 4. The task of Section 5 will be discussing our predictions under this new model as well as comparing our results to previous work on the Sir Philip Sidney game. Section 6 concludes.

[1] For an overview of animal communications literature, see (Maynard Smith and Harper [2004]; Searcy and Nowicki [2005]).

[2] This heuristic approach is not limited to Maynard Smith and has also been referred to as a 'simple-weighted-sum' (Grafen [1982]). Since we are focusing on the animal communications literature, we will refer to it as Maynard Smith's heuristic.

## 2 Heuristic Inclusive Fitness

Maynard Smith ([1991]) presents a heuristic for exploring interactions between relatives. In short, he calculates the inclusive fitness of an organism by adding its own payoff and its relative's payoff, weighted by a relatedness parameter, $k$. Critics and supporters of inclusive fitness alike have argued that this is an incorrect definition (Grafen [1982], [1984]; Skyrms [2002]; Nowak *et al.* [2010]; Okasha and Martens [2016]; Birch [2013], [2016]). Despite recognition that this calculation is incorrect, it is often viewed as a useful heuristic for estimating the inclusive fitness of traits. One intuitive argument for why this heuristic should give adequate predictions is this: if we are interested in tracking gene frequencies, adding the relatedness-weighted payoff of a relative to the focal organism's payoff means that the focal organism's genes will be passed on more often. In other words, it captures the fact that an organism in some sense cares about the payoff, or reproductive success, of its relatives and this is exactly the phenomenon that the relatedness parameter in inclusive fitness is supposed to capture.

However, despite its intuitive appeal, Maynard Smith's way of accounting for the alignment of interests between relatives is not accurate. By just adding the (relatedness-weighted) fitness of an organism's social partner to its own, fitness is (at least partially) double counted. For example, say we have two relatives, organism $A$ and organism $B$, which interact and both have trait $j$. Under Maynard Smith's heuristic, when we calculate the fitness of trait $j$ we count $A$'s fitness twice: once when we consider $A$'s contribution to the fitness of the trait and again (at least partially, depending on the value of $k$) when we take into account $B$'s contribution to the fitness of the trait. We similarly double count $B$'s fitness.

The correct way to calculate inclusive fitness tends to be more mathematically and conceptually complex. It can be thought of as first stripping an organism's fitness of all the fitness effects of others, in order to avoid double counting these effects, and then adding the fitness effects the organism confers on its relatives (Hamilton [1964]). Thus we can think of inclusive fitness as measuring the offspring caused by a particular organism (weighted by a relatedness parameter), in contrast to the more standard way of calculating fitness, neighbour-modulated fitness, which measures the offspring the focal organism actually has. In other words, neighbour-modulated and inclusive fitness calculations provide alternative ways of partitioning the causal structure of social interactions (Frank [2013]). In order to understand this contrast, it is helpful to compare how both types of fitness are calculated.

Roughly, the neighbour-modulated fitness approach considers the number of offspring an organism is expected to have due to some social interaction. When $i$ interacts with another organism $j$, its fitness is affected by its own

action—call this fitness effect $s_{ii}$—and by the action of the other organism—call this fitness effect $s_{ji}$. The neighbour-modulated fitness of organism $i$ for an interaction of interest is then calculated as follows:

$$f_i = s_{ii} + s_{ji}.$$

Inclusive fitness provides an alternative way of accounting for fitness effects from social interactions (Hamilton [1970]). When $i$ interacts with another organism, it affects its own fitness by some amount ($s_{ii}$), but it also affects the fitness the other organism by some amount ($s_{ij}$). Further, $i$ and $j$ may be more or less related and we can describe this relatedness with a parameter, $k$. We can then calculate the inclusive fitness of an organism from an interaction of interest as follows:

$$f_i = s_{ii} + k \cdot s_{ij}.$$

Importantly, inclusive fitness is not calculated by counting the number of offspring an organism has then adding all the offspring its relatives have, weighted by relatedness, as it is in Maynard Smith's heuristic.

The inclusive fitness framework might initially seem counter-intuitive, so it is helpful to start with a basic observation: in general, a trait will increase in frequency when organisms with the trait have more offspring than the average organism in the population. To determine whether a trait of interest will increase in frequency, we want to determine the number of offspring organisms with that trait will have. Inclusive fitness gives us this information by telling us how many offspring are caused by an organism and how likely it is that these offspring are had by an organism with the trait of interest. In this general sense, $k$ can be thought of as a measure of how likely it is that $i$ and its social partner $j$ share the trait of interest, relative to the rest of the population. The relatedness of a focal organism to its social partner is the probability the social partner has a trait given the focal organism does, minus the probability the social partner has the trait given the focal organism does not:

$$k = P(T_j|T_i) - P(T_j|N_i).$$

Why this is the correct measure of relatedness is discussed by Skyrms ([2002]) and van Veelen ([2009]).[3]

This definition of relatedness captures the fact that in interactions among kin, there is correlation between traits: related organisms tend to have the same trait, or employ the same strategy in the game.

As noted, despite the fact that Maynard Smith's method for estimating inclusive fitness is clearly incorrect, this heuristic way of accounting for

[3] The equivalence between this measure of relatedness and more common measures derived from the Price equation is discussed in (Rubin [2018]), and their equivalence with the measure of of relatedness employed by Grafen ([1979]) is discussed by Marshall ([2015]).

relatedness has been influential and has been employed in a variety of different strategic contexts (Johnstone and Grafen [1992]; Johnstone [1998]; Nowak [2006]; Taylor and Nowak [2007]; Archetti [2009a], [2009b]). In particular, this heuristic is routinely used in discussions of one of the central games in the literature on the evolution of communication, the Sir Philip Sidney game (see, for instance, the massive literature surrounding the Sir Philip Sidney game summarized in Maynard Smith and Harper [2004] and Searcy and Nowicki [2005]). In fact, the heuristic is often seen as preferable to explicit calculation of inclusive fitness. When payoffs are additive—that is, when the causal effects of an organism on its social partner's fitness are the same irrespective of the type of its social partner (allowing us to just sum all these fitness effects up to determine an organism's fitness)—the heuristic correctly identifies the Nash equilibrium of a game. Further, the heuristic is easy to generalize to games where payoffs are not additive, like the Sir Philip Sidney game. It is difficult to use the correct calculation of inclusive fitness this type of game because it is often unclear what fitness effects an organism is causally responsible for (Okasha and Martens [2016]). Thus, the heuristic is thought to give us an idea of the evolutionary outcomes we should expect in these more complicated models, despite the fact that it is known to have a problem of double counting. That is, it is commonly used in more complex evolutionary models both because it is easier to generalize and because it captures the important feature of relatedness as generating a degree of common interest between interacting organisms.

However, this common use of the heuristic rests upon and perpetuates the incorrect notion that relatedness represents the degree of common interest. For instance, in discussing the evolution of honest communication, Zollman ([2013]) considers the literature on the evolution of biological altruism, where it is well-known that correlations between traits can allow altruism to evolve, and presents correlated interactions as an alternative to inclusive fitness theory. In discussing situations where honest communication can be seen as a type of altruistic action (because it is costly for the honest organism, but beneficial for their relative), Zollman claims that 'the most popular solution to the biological altruism problem, inclusive fitness theory, cannot help in this context, since parent–offspring conflicts arise despite the high relatedness between parents and offspring' (Zollman [2013], p. 130). He instead proposes that we look to solutions using correlation between types and notes that 'Relatedness might, beyond inclusive fitness, introduce additional correlation' (Zollman [2013], p. 131).

This loses sight of the fact that generating common interest is merely a feature of high relatedness.[4] This can be a useful way of thinking about

[4] Although if one is interested in modelling the evolution of honest communication when the interacting organisms are not relatives but there is some degree of common interest for some other reason, models based on the heuristic could be useful.

relatedness, but not when it obscures the fact that relatedness is fundamentally a measure of correlation between types. When using the relatedness calculation above, inclusive fitness models neatly incorporate correlations between types into evolutionary predictions; the contrast between inclusive fitness and correlation as solutions to the problem of altruism only makes sense when considering the heuristic rather than explicit inclusive fitness calculations. Reliance on the heuristic both rests upon and perpetuates this misunderstanding of the nature of relatedness and its role in inclusive fitness.

We will show why this heuristic cannot capture correlations between types. We deviate from the literature and develop a model that uses neighbour-modulated fitness in order to examine the evolutionary predictions when the correlation inherent in parent–offspring relationships is incorporated. This article will argue against conclusions drawn using Maynard Smith's heuristic in the very case that motivated reliance on the heuristic within the animal communications literature in the first place: the Sir Philip Sidney game. To be clear, this article will argue against a particularly influential method for estimating inclusive fitness, not against the inclusive fitness framework as a whole.

This article, in part, echoes a debate that occurred the late 1970s over the use of this heuristic in what's called the Hawk–Dove game, which is used as a model of animals fighting over resources.[5] The prediction for this game, when not played with relatives, is that the population will be composed of a mixture of Hawks and Doves at equilibrium (what mixture this will be depends on the particular payoffs of the game). Maynard Smith ([1978]) argued that we could use the heuristic method of calculating the inclusive fitness of organisms to predict the evolutionary outcome in the Hawk–Dove game played among relatives. Grafen ([1979]) showed that whether or not this works depends on how the mixture of strategies in the population arises: when each organism is either hawkish or dovish, and there is some mixture of these two pure strategies in the population, the heuristic gives the wrong answer, but when all the organisms in the population are playing mixed strategies (that is, alternating between acting hawkish or dovish with some probability) the heuristic 'amazingly' gives the right answer. The response by Hines and Maynard Smith ([1979]) was to grant that Grafen is right, but then show that for games with these sort of mixed strategy equilibria, the heuristic lets you calculate necessary, but not sufficient, conditions for something to be an equilibrium.

[5] In this game 'hawk' refers to an aggressive animal that fights for a resource and 'dove' refers to an animal that is unwilling to fight. A hawk then always gets the contested resource when encountering a dove, but against another hawk will split the resource and pay a cost from fighting. A dove will always surrender the resource to a hawk, but will split the resource peacefully with another dove.

Throughout the article, we make claims about the heuristic giving incorrect predictions and our model giving correct predictions, and it is important to clarify what exactly these claims should be taken to mean. We do not mean that our model is the correct description of any real population, or that the outcomes of models using the heuristic cannot be present in a real biological population. Our model, like the models we compare it to, is highly idealized and so is not meant to describe any real population. Instead, we mean that our model gives the correct predictions for a set of modelling assumptions. (We include the assumption that interactions are among kin as one of the modelling assumptions.) That is, if we think of idealized models as describing non-actual, fictional populations in which the idealizing assumptions are true (but that are still similar to real world populations in important respects) (Godfrey-Smith [2009]), we claim that our model gives the correct predictions for evolution in the fictional population, while models with the heuristic do not.[6,7]

In order to make our case against the heuristic calculation of inclusive fitness, we present a model with as many of the same assumptions as possible (idealizing assumptions and otherwise) as a model using the heuristic (from Huttegger and Zollman [2010]), but with a correct calculation of fitness. This is the same strategy employed by Grafen ([1979]) where the conclusion he reached (and to which Hines and Maynard Smith [1979] agreed) was that when the predictions of the heuristic do not match those given by an explicit (non-heuristic, if still idealized) calculations of fitness, the heuristic's predictions should be considered incorrect.

Grafen ([1979]) observed that Maynard Smith's heuristic does not capture important features of the population structure when organisms are interacting with relatives. Our discussion supports this point and goes beyond the debate over the Hawk–Dove game in discussing the evolutionary significance of the equilibria predicted by Maynard Smith's heuristic. We show a simple evolutionary model involving neighbour-modulated fitness leads to equilibria with completely different strategies than those predicted by the heuristic. We further show that the heuristic will identify certain outcomes as overwhelming likely when—compared to the results of our neighbour-modulated fitness model—they are very unlikely (or even impossible).

Our discussion is also important because, as noted, Maynard Smith's heuristic has been used to predict outcomes of the Sir Philip Sidney game, which is

[6] Or, depending how one thinks about the relationship between a model and its interpretation, we could say that our model correctly describes the fictional population it attempts to give predictions for, while models using the heuristic claim be talking about one fictional population (where organisms are relatives and so are likely to have the same traits) but end up giving predictions for a different population. The heuristic would then be thought of as giving incorrect predictions in the sense that it gives predictions for the wrong fictional population.

[7] Thanks to an anonymous reviewer for helping us to more clearly state the sense in which the heuristic makes incorrect predictions.

used for studying the emergence of communication among relatives in situations where their interests conflict. As Birch ([2013]) notes, 'This is not to say that the game does not shed any light on the evolution of signalling; but if it does, this can only be because the miscalculations of inclusive fitness it embodies turn out to be harmless'. We provide a new model of the Sir Philip Sidney game that does not rely on these miscalculations and that incorporates correlations between types to show how the predictions differ from those obtained under the heuristic. We see that some of the previous insights from Bergstrom and Lachman ([1997]) and Huttegger and Zollman ([2010]) are preserved (for example, that costs are no longer necessary to sustain honest communication when organisms are highly related), but other conclusions are not supported.

## 3 The Sir Philip Sidney Game

As mentioned, the Sir Philip Sidney game is one of the central games in the animal communications literature, taking up substantial portions of influential books on the subject (Maynard Smith and Harper [2004]; Searcy and Nowicki [2005]). This game involves two individuals, a sender and a receiver. The sender can be one of two types, 'healthy' or 'needy', and initially only the sender is aware of its underlying type.[8] The sender can select to send a signal or remain mute. If the sender signals, it bears a cost of $c$. The receiver then has the option of transferring resources to the sender. This transfer is costly for the receiver, but significantly benefits the sender. If the receiver transfers the resource, its payoff drops from 1 to $1 - d$ (where $d > 0$). While both healthy and needy sender types welcome the transfer, needy types benefit more than healthy types from the transfer. In the absence of a transfer, needy and healthy types receive a payoff of $1 - a$ and $1 - b$, respectively (where $a > b$). If a transfer is made, then both types of senders attain a payoff of 1. This game is displayed in extensive form in Figure 1 and the four sender and receiver strategies are summarized in Table 1.

If sender and receiver are unrelated, honest communication is not possible since the receiver does best to never transfer resources to her counterpart. Yet communication is a real possibility if the agents are related. As is commonly done, one can introduce a relatedness parameter, $k \in [0, 1]$, which allows for communication by partially (if not completely) aligning the interests of sender and receiver. Following (Maynard Smith [1991]), an individual in this game would then receive its own payoff plus the relatedness parameter

[8] More complicated versions of the Sir Philip Sidney game representing, for instance, the child's health as a continuous variable can be found in (Bergstrom and Lachmann [1997]). We stick with this basic set-up introduced by Maynard Smith ([1991]) and later explored by Huttegger and Zollman ([2010]).

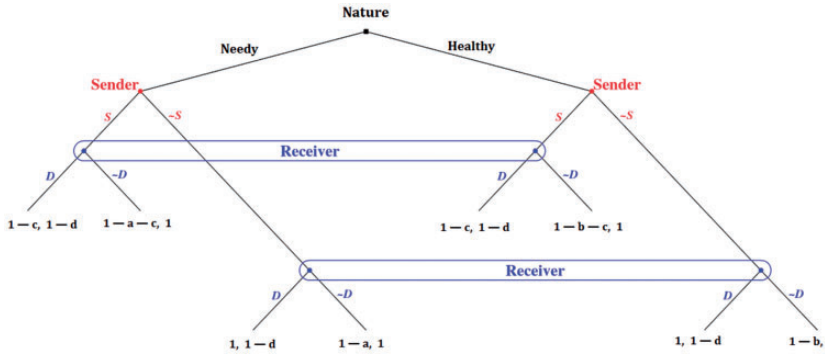


**Figure 1.** Extensive form of the Sir Philip Sidney game.

**Table 1.** Sender and receiver strategies

| Sender strategy | Receiver strategy |
|---|---|
| S1: Only signal when healthy | R1: Transfer only if signal |
| S2: Only signal when needy | R2: Transfer only when no signal |
| S3: Always signal | R3: Always transfer |
| S4: Never signal | R4: Never transfer |

times its counterpart's payoff.[9] It is easy to see how the inclusion of this relatedness parameter aligns the interests of sender and receiver. For example, in the extreme case where $k = 1$ the parties never disagree with regards to whether the receiver should donate or not.

There are a number of equilibria in this game. At a so-called separating equilibrium the receiver correctly infers the sender's type from its signalling behaviour. For instance, at (S2, R1) senders only signal if they are needy and receivers only transfer to needy individuals. An alternative separating equilibrium exists in which senders signal only if they are healthy and receivers transfer when they do not receive a signal (S1, R2).[10] Note that the only difference between these two signalling arrangements is whether needy or healthy individuals signal. In both cases transfers are only made to the needy individual. Furthermore, for both of these signalling arrangements, substantial signalling costs are needed to ensure stability for a wide range of parameter values.[11]

[9] Some refer to this new game (with payoffs modified to incorporate relatedness) as the Sir Philip Sidney game, but we will refer to the original game (with unmodified payoffs) as the Sir Philip Sidney game and to the change in payoff structure as Maynard Smith's heuristic.

[10] This is an equilibrium if $a \geq kd - c \geq b$ and (S2, R1) is an equilibrium when $a \geq c + kd \geq b$ and $a \geq d/k \geq b$.

[11] For instance, Bergstrom and Lachmann ([1997]) calculate that the minimum cost necessary to sustain the (S1, R2) equilibrium is $c = b - kd$.

Such costs are often required in order to ensure that not all sender types have incentive to signal. The crucial role cost plays in the Sir Philip Sidney game is consistent with the handicap principle, which, briefly put, contends that honesty in conflict of interest cases typically requires significant costs attached to signalling behaviour.[12]

There also exist so-called pooling equilibria in which no information is transferred from sender to receiver. For instance, the strategy pair 'never signal' and 'never donate' is stable when $d > k(ma + (1-m)b)$, where $m$ is the probability that the sender is needy. Similarly, the pooling equilibrium in which senders never signal and receivers always donate is stable when $d < k(ma + (1-m)b)$.

Huttegger and Zollman ([2010]) explore the evolutionary dynamics of the Sir Philip Sidney game. Through the use of computer simulation they estimate the basin of attraction (a measure of how likely it is that the population will evolve to a particular state) for the various equilibria. In particular, they use the two-population discrete-time replicator dynamics, a standard model of biological evolution. While details regarding the replicator dynamics are spelled out in the Appendix, it is worth noting here that the two-population version of the replicator dynamics employed by Huttegger and Zollman ([2010]) assumes random assortment, meaning senders (children) and receivers (parents) are randomly paired to play the Sir Philip Sidney game. They incorporate the parent–offspring relationship into their analysis solely by the use of the relatedness parameter, $k$. In what follows, we will explain how their results differ from results based on a correct calculation of fitness.

## 4 Model

We now develop a model of the Sir Philip Sidney game using neighbour-modulated fitness. While Huttegger and Zollman ([2010]) utilize the two-population replicator dynamics to study the Sir Philip Sidney game, we instead develop a class-structured one-population model. The model begins with an infinite population of 'parents', Each parent is endowed with both a sender and receiver strategy. We refer to this sender–receiver strategy combination as their 'total strategy' (TS), and note that there are sixteen possible total strategies. So, for example, if the TS of an agent is S1–R3, the agent only

[12] Some formulations of the handicap principle require that the cost of signalling be different for different types of senders (that is, that signals be differentially costly). We follow Zollman ([2013]) and work with a broader reading of the handicap principle. On this formulation all sender types may pay the same cost to signal and honesty is possible due to the fact that some sender types benefit more than others from the receiver's response. Note that signal cost is still a necessary ingredient of the handicap principle, for without signal cost none of the senders will be disincentivized from signalling, no matter how paltry the benefit.

signals when healthy if in the sender role and always transfers when in the role of receiver.

Each round of the model begins with all parents asexually reproducing to create a single child each. Children inherit the TS of their parent with probability $1-\mu$. With probability $\mu$, the child 'mutates' with equal probability to one of the other fifteen TS types. After reproduction, parents and children play the Sir Philip Sidney game as displayed in Figure 1. Children take the role of sender while parents inhabit the role of receiver. We then calculate the average payoff for each TS type (see the Appendix). Payoffs in this framework should be interpreted as affecting the likelihood an individual survives to the next time period (and thus can reproduce again). These average payoffs are then plugged into the discrete-time replicator dynamics to determine the new composition of the population.[13] This completes one generation, or round of the simulation, and the frequencies of types in the population become the new frequencies of parents in the next generation.

It is important to note that this model does not capture all the important biological details of the situation we've described. In particular, it assumes that organisms reproduce asexually when in fact many (or most) organisms of interest in the animal communications literature reproduce sexually. There are a few reasons we have elected to set up our model in this fashion. First, our aim is to contrast the results of our neighbour-modulated fitness model with previous results produced using Maynard Smith's heuristic, many of which are produced for high values of $k$ that cannot be achieved with sexually reproducing organisms.[14]

Additionally, we aim to show that different predictions arise due to the way relatedness is incorporated into the formal analysis. Therefore, our main alteration to past work on the Sir Philip Sidney game can be summed up as such: instead of altering the payoffs of the underlying strategic interaction using a 'relatedness' parameter, we incorporate relatedness into our account by having offspring inherit a strategy from its parent and interact with said

[13] Briefly, the replicator dynamics states that the frequency of agents utilizing strategy $x_i$ in the next generation $(t+1)$ is equal to $\frac{F_i x_i(t)}{F}$, where $F_i$ is the average payoff for agents using $x_i$, $F$ is the average payoff for the population as a whole and $x_i(t)$ refers to the proportion of agents using $x_i$ in generation $t$. Note that in our model average payoffs ($F_i$) are determined by the mutation rate $\mu$ as well as the overall composition of the population. More on the ways our model departs from the standard replicator dynamics model in the Appendix.

[14] Much more conflict of interest between relatives can be generated with organisms reproducing sexually because in these populations relatedness is generally close to 1/2, whereas relatedness will be close to 1 in the simulations we discuss. The evolutionary predictions for sexually reproducing populations are more complicated, involving what is called a 'hybrid equilibrium' (Huttegger and Zollman [2010]). Since many of Bergstrom and Lachmann's ([1997]) and Huttegger and Zollman's ([2010]) results are regarding population very high relatedness, our point can be made more clearly with a model involving asexually reproducing organisms. What happens in the Sir Philip Sidney game when there is sexual reproduction is the subject of future study.

parent.[15] That is, since previous results were obtained using a model with asexual reproduction, we also use a model with asexual reproduction in order to allow a direct comparison between the model when relatedness is incorrectly incorporated and when relatedness is correctly incorporated. This fits with our goal, described in Section 2, of comparing models with the same assumptions, where one model correctly calculates fitness and the other calculated fitness incorrectly via the heuristic, in order to show that the heuristic provides incorrect predictions for the idealized population it claims to describe. In addition to asexual reproduction, both models assume: an infinite population, that payoffs from the game are the sole determinate of evolutionary success, there are no sources of noise, evolution proceeds in discrete-time-steps, and so on.

While there are obvious benefits to setting up the model in this way (namely, it allows us to better understand the ways in which the heuristic misleads) there are limitations when it comes to investigating various other claims typically made in the literature about the evolution of communication among relatives. For instance, we focus on predictions for high values of relatedness, where Maynard Smith's ([1991]) observation was that the conflict of interest preventing honest communication is attenuated. We do not look at whether signals must be costly in order for honest signalling to be stable in cases where organisms have conflicting interests. Generally, these conflicts of interest only arise in Maynard Smith's ([1991]) model when relatedness is lower than what can be reasonably achieved in our model. A more complicated model would be needed to properly investigate these lower values of relatedness. Thus we sacrifice a full investigation of the Sir Philip Sidney game for clarity and ease of comparison between models employing the heuristic and our approach involving correct calculations of fitness.

## 5 Results

Our goal in this section is two-fold. First, we explore the implications of the model discussed in the previous section. We find that for even sizable mutation rates honest communication is often the only evolutionarily significant outcome and that signal cost is not necessary to stabilize honesty among highly related organisms. Second, we contrast our results with previous work relying on Maynard Smith's heuristic. Overall, reliance on Maynard Smith's heuristic paints a very different picture of when information transfer is possible when

[15] Grafen ([1979]) achieves something similar in his model, where an organism will interact with its own type with probability $r$ and will interact with another type randomly drawn from the population with probability $1-r$. Since there is a small mutation rate in our model, $r$ will change as the population composition changes and so Grafen's ([1979]) model would give predictions that are slightly off. See (Rousset [2002]) for an explanation of why mutation causes relatedness to change with the population composition.

compared to our model involving neighbour-modulated fitness, significantly understating the likelihood of honest communication.

We begin with the extreme case in which the mutation rate, $\mu$, is zero. In the absence of mutations organisms interact with a clone. That is, in the absence of mutation parents and offspring are always of the same type, meaning $P(T_j|T_i)=1$ and $P(T_j|N_i)=0$, corresponding to the case in which $k=1$. When organisms only interact with their own type, what determines the evolutionary success of a TS is how well it fares against itself. This means that selection is not frequency dependent as the fitness of a type is not affected by other types. It is easy to show that either S1–R2 (only signal when healthy and transfer only when no signal) or S2–R1 (only signal when needy and transfer only when signal), corresponding to the two separating equilibria, will go to fixation for a wide variety of parameters. In general these separating strategies are favoured by evolution when the costs associated with signalling are small and the costs associated with making a transfer is significantly greater than the benefit healthy types receive from a transfer. Which of the two separating equilibria is reached depends on how likely it is that offspring are needy (for example, if it is unlikely that offspring are needy, then the population will reach the equilibirum where offspring only signal when needy, since they will have to pay the signal cost less often). See the Appendix for the exact conditions under which information transfer evolves when $\mu=0$.

These preliminary results already depart from the findings of Bergstrom and Lachmann ([1997]). In the model explored in (Bergstrom and Lachmann [1997]), a pooling arrangement where senders never signal and receivers simply best respond to the proportion of needy and healthy types will always be a stable equilibrium, even when the agents are fully related (that is, when $k=1$). As mentioned above, on our neighbour-modulated fitness model, separating is often the only evolutionarily significant arrangement when agents play with clones (corresponding to the case in which $k=1$). Pairs of agents using pooling strategies do on average worse than those using separating strategies, meaning a population at the pooling arrangement can be easily invaded by a mutant using a separating strategy.

To see this more clearly, consider the case in which $a=31/32$, $b=9/32$, $d=1/2$, $c=1/100$, $m=4/10$ and the mutation rate is set to zero (corresponding to $k=1$). Bergstrom and Lachmann ([1997]) and Huttegger and Zollman ([2010]) predict two stable evolutionary outcomes are possible. The familiar separating equilibrium in which only needy types signal and receivers donate only upon receipt of a signal is evolutionarily significant, but so is the pooling equilibrium where senders never signal and receivers always donate. This is an equilibrium due to the fact that senders do best to not signal so as to avoid any signalling costs, and the frequency of needy senders is such that receivers prefer always making a donation to never making a donation.

Yet on our model this 'pooling equilibrium' is not a real possibility for the above parameters. Recall that when $\mu = 0$ selection is not frequency dependent so the separating strategy invades because it does best against itself. The TS S4–R3 (never signal, always transfer) when paired against itself secures an average payoff of 0.75. The separating strategy S2–R1 yields a payoff of 0.8136 when playing against a clone. Thus a population of individuals utilizing S4–R3 is not stable for the introduction of a lone agent utilizing S2–R1 will quickly drive the natives out of the community.

These pooling outcomes are additionally problematic because, under the replicator dynamic used with Maynard Smith's heuristic, such outcomes often have a sizable basin of attraction.[16] In other words, not only does the approach taken by Maynard Smith indicate pooling outcomes are stable, it also predicts that for large swaths of parameter space these arrangements will be the likely outcome of an evolutionary process.

To get a better sense of this, consider Figure 2. As is evident, high levels of information transfer are predicted when relatedness is incorporated into the model via a class-structured population using neighbour-modulated fitness. Yet under Maynard Smith's approach, the pooling equilibrium is attainable and has a sizable basin of attraction. When there is a 20% chance a sender is needy, for example, around 85% of simulation runs of the two-population replicator dynamics result in the pooling equilibrium.

Similarly, Maynard Smith's heuristic can also overstate the likelihood of information transfer by incorrectly identifying certain separating arrangements as stable when in fact signalling systems are in fact unstable once fitness is correctly accounted for in the mathematical model. Consider the numerical example used above except the cost associated with signalling is now 13/32. According to the heuristic, for $k = 1$ there exists a separating equilibrium where senders only signal when needy and receivers only transfer upon receipt of a signal (S2–R1) as well as a pooling equilibrium where senders never signal and receivers always transfer resources (S4–R3). Yet when paired against a clone, the pooling strategy S4–R3 outperforms the separating strategy S2–R1 (payoffs are 0.75 and 0.735, respectively). Thus while Maynard Smith's inclusive fitness heuristic suggests a separating equilibrium is possible under these parameters, the neighbour-modulated fitness approach does not allow for any information transfer.

Thus far we have considered the extreme case in which $\mu = 0$. When $\mu > 0$, information transfer is still a real possibility. Figure 3 shows the proportion of

[16] Outcomes our neighbour-modulated fitness model deems to be evolutionarily significant cannot be identified by Maynard Smith's heuristic as unstable when relatedness is 1. For $k = 1$, individuals only have incentive to change their strategy when doing so increases the combined payoff of sender and receiver. Thus any arrangement that is out of equilibrium will not maximize the joint payoff of the two individuals.

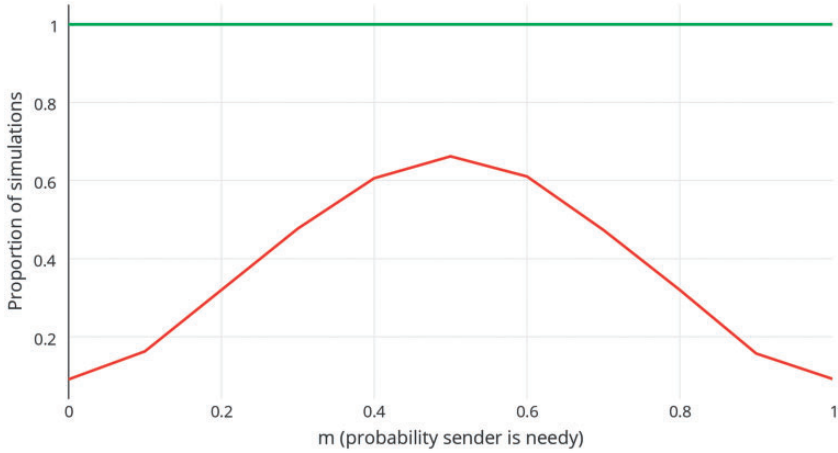


**Figure 2.** Simulation results for Sir Philip Sidney game for various values of $m$. Proportion of simulations that ended with populations using a signalling system for neighbour-modulated fitness model (light) and inclusive fitness model (dark) for values $a = 17/32$, $b = 15/32$, $d = 1/2$, $c = 1/100$, $\mu = 0$, and $k = 1$.

the population one can expect to use the signalling system, as $m$ changes. As Figure 3 illustrates, honest communication prevails even when mutations occur relatively frequently (1% of the time).[17] Note that for extreme values of $m$ (the probability a sender is needy) a smaller proportion of the population uses separating strategies (S1–R2 or S2–R1). This is due to the fact that pooling strategies fare better when $m$ is close to zero or one. This makes intuitive sense—in those cases where the vast majority of agents are needy (healthy), receivers who simply treat senders as though they were needy (healthy) do relatively well.

Furthermore, our results are interestingly not significantly affected by the cost associated with signalling. That is to say, as long as the value of $c$ is such that S2–R1 or S1–R2 still fare better against their own strategy (which they most often interact with for low mutation rates) than alternative strategies, the inclusion of minor signalling costs does not significantly reduce the proportion of the community utilizing signalling strategies. This result holds for even moderate mutation rates, and the exact relationship between signal cost and the evolution of separating equilibria is shown in Figure 4. Once again, our findings depart from results predicated on Maynard Smith's heuristic. Huttegger and Zollman ([2010]), for instance, find that increased signal cost reduces the size of the basin of attraction of the separating equilibrium. This is

[17] For extremely high mutation rates, such as 50%, the end state of the population is an even mixture of all sixteen types. This result holds for more moderate (but still unrealistically high) mutation rates as well, such as 20%.

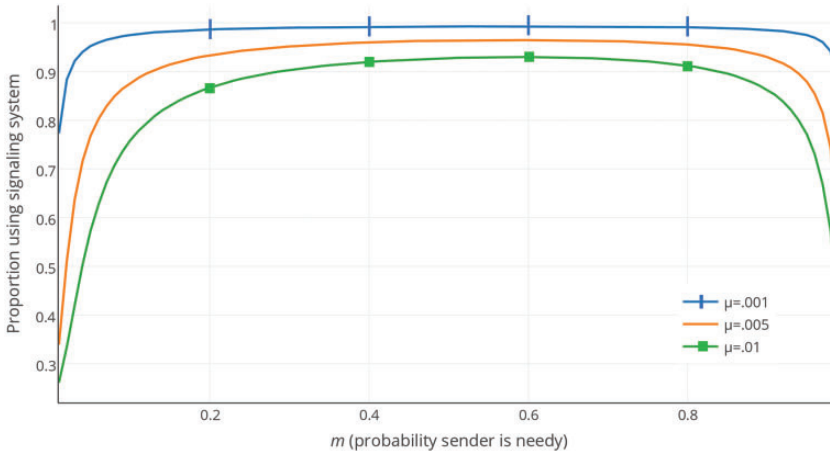


**Figure 3.** Frequency of separating strategies (S1–R2 or S2–R1) in the population for various values of $m$ for parameters $c = 0$, $a = 31/32$, $b = 9/32$, $d = 1/2$, and $\mu = 0.001$, $\mu = 0.005$, and $\mu = 0.01$.

due to the fact that under the replicator dynamics selection is frequency dependent, and an increase in the cost associated with signalling reduces the likelihood of information transfer in a rather straightforward fashion. In particular, when receivers are initially unresponsive to the signalling behaviour of senders, high signal cost heavily penalizes senders utilizing separating strategies, pushing the population toward the pooling equilibrium in which senders never signal. Note that this explanation depends on the fact that the heuristic incorporates relatedness into the game via alteration of the payoffs, representing something like the degree of common interest between organisms, rather than via correlations between types.

We have explained the results of our model in terms of neighbour-modulated fitness, but we can still provide some *post hoc* explanations in terms of inclusive fitness, even without formulating the exact fitness calculations for each type. For instance, when a parent always donates a resource it affects its own fitness by $-d$ and its offspring's fitness by $ma + (1 - m)b$. We weight this second term by $k$, which with a small mutation rate is close to 1.[18] This means, when $[ma + (1 - m)b]k - d > 0$ parents will transfer the resource in the model we've provided above. This has the form of Hamilton's rule, more commonly seen written as:

$$BR - C > 0,$$

[18] When the mutation rate is low, organisms tend to interact with their own type: $P(T_j|T_i)$ is high and $P(T_j|N_i)$ is low, so $k$ is high (although the actual value of $k$ will depend on the population composition (Rousset [2002]).

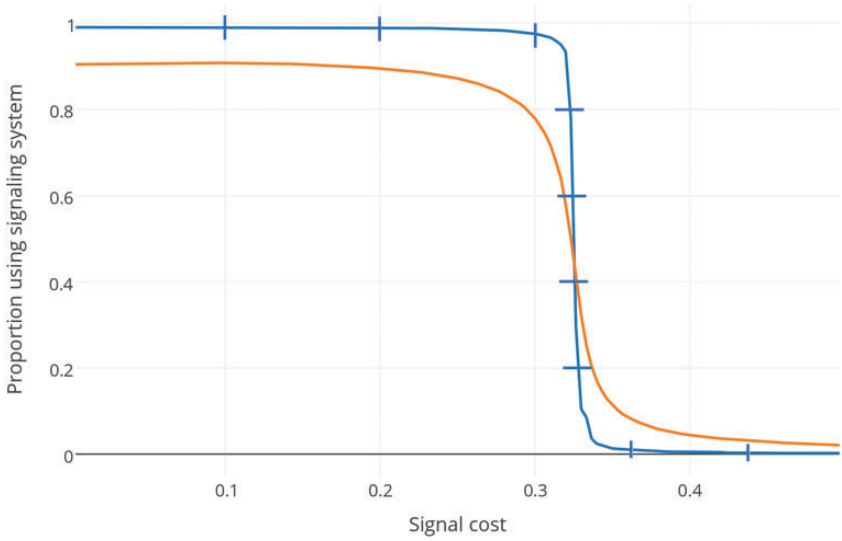


**Figure 4.** Frequency of separating strategies (S1–R2 or S2–R1) in the population for various values of $c$ (signal cost) and parameters $a = 31/32$, $b = 15/32$, $d = 1/2$, $m = 4/10$, and $\mu = 0.001$ (dark), and $\mu = 0.01$ (light).

where $B = ma + (1 - m)b$ is the benefit bestowed on the relative, $R = k$ is relatedness, and $C = d$ is the cost of bestowing the benefit. In other words, although our model used neighbour-modulated fitness we can still explain certain results in terms of relatedness having an important causal effect on the outcomes using the inclusive fitness framework. It is not the use of the inclusive fitness framework that leads to the difference between our results and those obtained by Huttegger and Zollman ([2010]), but instead the use of a heuristic that incorrectly incorporates relatedness.

## 6 Conclusion

A number of points are in order. First, full information transfer is possible even when there is no cost associated with signalling.[19] In other words, information transfer does not require significant signal costs when the parties are sufficiently related to align their interests. Yet this insight was originally predicated on a family of models that incorrectly incorporated relatedness into the game-theoretic model. We have preserved this basic finding while modelling parent–offspring interactions in a way that captures the fact that they tend to have the same phenotype. This is important because it can help explain why

[19] Note that previous studies not incorporating relatedness such as (Wagner [2012]; Martinez and Godfrey-Smith [2016]) have only shown some level of information transfer is possible when signals are cost-free.

empirical biologists have often had difficulty registering the existence of significant signal costs in nature—cost is simply not needed to stabilize honesty among kin.

However, while the basic insight arising from these models is correct and important for the study of animal communication, the predictions they make are incorrect for many parameter values—they do not match those obtained using explicit fitness calculations. For instance, in contrast to what Huttegger and Zollman ([2010]) find, as long as signalling is advantageous cost will not significantly affect the likelihood of honest communication evolving. Additionally, as the probability of offspring being needy either increases or decreases away from 0.5, our model predicts significantly more information transfer than Huttegger and Zollman ([2010]).

As noted earlier, the inclusive fitness framework is considered by some to be indispensable to the study of evolution. Yet despite its central position in evolutionary theory, models using inclusive fitness are not immune to criticism, as evidenced by the recent debate. Our analysis further reveals that even some commonly used and influential heuristics may be problematic when thinking about inclusive fitness. Compared to our model that incorporates correlations between types, predictions made on the basis of this heuristic give drastically different characterizations of the evolutionary dynamics of the Sir Philip Sidney game. Thus while such heuristics may conveniently simplify the task of evolutionary analysis, they should be used with both care and caution.

## Acknowledgements

Thanks to Brian Skyrms, Simon Huttegger, Kevin Zollman, Jonathan Birch, Patrick Forber, Kim Sterelny, Elliott Sober, Jack Justus, Wes Anderson, Bruce Glymour, and three anonymous referees. Thanks also to the audiences at Philosophy of Biology at Madison 2016, the 2016 Biennial Meeting of the Philosophy of Science Association, and the Social Dynamics seminar at UC Irvine.

## Appendix

### A1 Details about the Model

Recall that in each round of the model all 'parents' produce one child and play the Sir Philip Sidney game with this child. With probability $\mu$ the child mutates with equal probability to one of the other fifteen types. To determine how the population changes over time, we must first calculate the average payoff associated with each type. We first separately calculate the average payoff for offspring and parents.

The average payoff associated with offspring employing TS $j$ is determined as follows: Let $A^o_{jj}$ be the payoff an offspring utilizing $j$ receives against parent

utilizing $j$. Likewise, $A^o_{ji}$ refers to the payoff offspring utilizing $j$ receives when interacting with parent utilizing $i$. Let $B_j$ refer to the total proportion of offspring utilizing TS $j$. Thus $B_j = (1 - \mu)x^t_j + \sum_{i \neq j} \frac{\mu x^t_i}{n-1}$, where $x^t_j$ is the proportion of parents at round $t$ utilizing strategy $j$ and $\mu$ is the chance of mutation. Note that the first term in this expression refers to offspring with parents utilizing $j$, while the second term covers offspring with parents not utilizing $j$. Thus $(1 - \mu)x^t_j/B_j$ refers to the proportion of offspring utilizing $j$ with parents that also employ $j$. Similarly, $\frac{\mu}{(n-1)} x^t_k/B_j$ refers to the proportion of offspring using $j$ with parents using $k$. Putting these together, we can now calculate the average payoff for offspring utilizing $j$ ($P^o_j$):

$$\frac{1}{B_j}\left[(1 - \mu)x^t_j A^o_{jj} + \sum_{k \neq j} \frac{\mu}{n - 1} x^t_k A^o_{jk}\right].$$

Calculating the average payoff to parents utilizing $j$ is a bit more straightforward. Let $A^p_{jj}$ be the payoff parents receive against an offspring also utilizing $j$ while $A^p_{ji}$ is the payoff to parent when paired with an offspring using $i$. Given that the chance of mutation is $\mu$ and when mutations do occur the mutant has an equal chance of being any of the remaining fifteen types, average payoff to parent of type $j$ is ($P^p_j$): The average payoff to parent of type $j$ is ($P^p_j$):

$$(1 - \mu)A^o_{jj} + \sum_{i \neq j} \frac{\mu A^p_{ji}}{n - 1}.$$

Note that since each parent has exactly one child, the proportion of individuals utilizing TS $j$ is the proportion of parents utilizing $j$ ($x^t_j$) plus the proportion of children utilizing $j$ ($B_j$). The average payoff associated with an individual utilizing TS $j$ ($P_j$) is thus:

$$\frac{1}{B_j + x^t_j}\left(x^t_j P^p_j + B_j P^o_j\right).$$

To determine how the population changes over time we appeal to the discrete-time replicator dynamics. Under the discrete-time replicator dynamics, the proportion of individuals of type $j$ in the next time period is equal to the proportion of individuals in the current time period that are of type $j$ multiplied by the average fitness of type $j$ divided by the average fitness of the population as a whole. Thus for us, the proportion of individuals utilizing TS $j$ in the next time period is:

$$(x^{t+1}_j + B_j)\frac{P_j}{2P},$$

where $P$ is the average payoff across all types ($\frac{1}{16}\sum_{i=1}^{16} P_i$).

## A2 Conditions for Full Information Transfer

When the mutation rate is zero, the strategy that goes to fixation is the strategy that does best against a clone. We consider the conditions under which a separating strategy (S1–R2 or S2–R1) does best against a clone. It can be easily shown that of the twelve TS that never constitute an equilibrium, none of these can do better than the best pooling strategy when paired with a clone.

> Case 1: Pooling strategy S4–R3 outperforms pooling strategy S4–R4 when paired with a clone.

This occurs when $(1-m)b + ma > d$. Separating strategy S2–R1 does better than S4–R3 against a clone when $(1-m)(d-b) - mc > 0$. Similarly, separating strategy S1–R2 outperforms S4–R3 when $d-b-c>0$. In other words, when the cost associated with signalling is small and the cost associated with making a transfer is significantly greater than the benefit healthy types receive from a transfer information transfer is favoured.

> Case 2: Pooling strategy S4–R4 outperforms pooling strategy S4–R3 when paired with a clone.

This occurs when $(1-m)b + ma < d$. The separating strategy S2–R1 does better than S4–R4 against a clone when $a-c-d>0$. Likewise, separating strategy S1–R2 outperforms pooling strategy S4–R4 when $m(a-d)-(1-m)c>0$. These conditions suggest information transfer is possible when the benefit needy types receive from a transfer swamps the cost to the receiver as well as the cost of signalling.

Justin P. Bruner
*Australian National University*
*Canberra, Australia*
*and*
*University of Groningen*
*Groningen, The Netherlands*
*justinpbruner@gmail.com*

Hannah Rubin
*University of Groningen*
*Groningen, The Netherlands*
*and*
*University of Notre Dame*
*Notre Dame, IN, USA*
*hannahmrubin@gmail.com*

## References

Abbot, P., Abe, J., Alcock, J., Alizon, S., Alpedrinha, J. A. C., Andersson, M., Andre, J.-B., van Baalen, M., Balloux, F., Balshine, S., Barton, N., Beukeboom, L. W., Biernaskie, J. M., Bilde, T., Borgia, G., Breed, M., Brown, S., Bshary, R., Buckling, A., Burley, N. T., Burton-Chellew, M. N., Cant, M. A., Chapuisat, M., Charnov, E. L., Clutton-Brock, T., Cockburn, A., Cole, B. J., Colegrave, N., Cosmides, L., Couzin, I. D., Coyne, J. A., Creel, S., Crespi, B., Curry, R. L., Dall, S. R. X., Day, T., Dickinson, J. L., Dugatkin, L. A., El Mouden, C., Emlen, S. T., Evans, J., Ferriere, R., Field, J., Foitzik, S., Foster, K., Foster, W. A., Fox, C. W., Gadau, J., Gandon, S., Gardner, A., Gardner, M. G., Getty, T., Goodisman, M. A. D., Grafen, A., Grosberg, R., Grozinger, C. M., Gouyon, P.- H., Gwynne, D., Harvey, P. H., Hatchwell, B. J., Heinze, J., Helantera, H., Helms, K. R., Hill, K., Jiricny, N., Johnstone, R. A., Kacelnik, A., Kiers, E. T., Kokko, H., Komdeur, J., Korb, J., Kronauer, D., Kümmerli, R., Lehmann, L., Linksvayer, T. A., Lion, S., Lyon, B., Marshall, J. A. R., McElreath, R., Michalakis, Y., Michod, R. E., Mock, D., Monnin, T., Montgomerie, R., Moore, A. J., Mueller, U. G., Noë, R., Okasha, S., Pamilo, P., Parker, G. A., Pedersen, J. S., Pen, I., Pfennig, D., Queller, D. C., Rankin, D. J., Reece, S. E., Reeve, H. K., Reuter, M., Roberts, G., Robson, S. K. A., Roze, D., Rousset, F., Rueppell, O., Sachs, J. L., Santorelli, L., Schmid-Hempel, P., Schwarz, M. P., Scott-Phillips, T., Shellmann-Sherman, J., Sherman, P. W., Shuker, D. M., Smith, J., Spagna, J. C., Strassmann, B., Suarez, A. V., Sundström, L., Taborsky, M., Taylor, P., Thompson, G., Tooby, J., Tsutsui, N. D., Tsuji, K., Turillazzi, S., Ubeda, F., Vargo, E. L., Voelkl, B., Wenseleers, T., West, S. A., West-Eberhard, M. J., Westneat, D. F., Wiernasz, D. C., Wild, G., Wrangham, R., Young, A. J., Zeh, D. W., Zeh, J. A. and Zink, A. [2011]: 'Inclusive Fitness Theory and Eusociality', *Nature*, **471**, p. E1.4.

Archetti, M. [2009a]: 'Cooperation and the Volunteer's Dilemma and the Strategy of Conflict in Public Goods Games', *Journal of Evolutionary Biology*, **22**, pp. 2192–200.

Archetti, M. [2009b]: 'The Volunteer's Dilemma and the Optimal Size of a Social Group', *Journal of Theoretical Biology*, **261**, pp. 475–80.

Bachman, G. C. and Chappell, M. A. [1998]: 'The Energetic Cost of Begging Behaviour in Nestling House Wrens', *Animal Behaviour*, **55**, pp. 1607–18.

Bergstrom, C. T. and Lachmann, M. [1997]: 'Signalling among Relatives, I: Is Costly Signalling Too Costly?', *Philosophical Transactions of the Royal Society of London B*, **352**, pp. 609–17.

Birch, J. [2013]: 'Review: Samir Okasha and Ken Binmore Evolution and Rationality: Decisions, Cooperation, and Strategic Behaviour', *British Journal for the Philosophy of Science*, **64**, pp. 669–73.

Birch, J. [2016]: 'Hamilton's Two Conceptions of Social Fitness', *Philosophy of Science*, **83**, p. 848.

Brilot, B. O. and Johnstone, R. [2003]: 'The Limits of Cost-Free Signaling of Need between Relatives', *Proceedings of the Royal Society London B*, **270**, pp. 1055–60.

Frank, S. A. [2013]: 'Natural Selection, VII: History and Interpretation of Kin Selection Theory', *Journal of Evolutionary Biology*, **26**, pp. 1151–84.

Godfray, H. C. J. [1995]: 'Signaling of Need between Parents and Young: Parent–Offspring Conflict and Sibling Rivalry', *The American Naturalist*, **146**, pp. 1–24.

Godfrey-Smith, P. [2009]: 'Abstractions, Idealizations, and Evolutionary Biology', in A. Barberousse, M. Morange and T. Pradeu (*eds*), *Mapping the Future of Biology*, Springer, pp. 47–56.

Grafen, A. [1979]: 'The Hawk–Dove Game Played between Relatives', *Animal Behavior*, **27**, pp. 905–7.

Grafen, A. [1982]: 'How Not to Measure Inclusive Fitness', *Nature*, **298**, pp. 425–26.

Grafen, A. [1984]: 'Natural Selection, Kin Selection and Group Selection', in J. R. Krebs and N. B. Davies (*eds*), *Behavioural Ecology*, Oxford: Blackwell.

Hamilton, W. D. [1964]: 'The Genetical Evolution of Social Behavior, I and II', *Journal of Theoretical Biology*, **7**, p. 17.

Hamilton, W. D. [1970]: 'Selfish and Spiteful Behaviour in an Evolutionary Model', *Nature* **228**, pp. 1218–20.

Hines, W. G. S. and Maynard-Smith, J. [1979]: 'Games between Relatives', *Journal of Theoretical Biology*, **79**, pp. 19–30.

Huttegger, S. and Zollman, K. [2010]: 'Dynamic Stability and Basins of Attraction in the Sir Philip Sidney Game', *Proceedings of the Royal Society London B*, **277**, pp. 1915–22.

Johnstone, R. [1998]: 'Efficacy and Honesty in Communication between Relatives', *The American Naturalist*, **152**, pp. 45–58.

Johnstone, R. and Grafen, A. [1992]: 'The Continuous Sir Philip Sidney Game: A Simple Model of Biological Signaling', *Journal of Theoretical Biology*, **156**, pp. 215–34.

Martinez, M. and Godfrey-Smith, P. [2016]: 'Common Interest and Signaling Games: A Dynamic Analysis', *Philosophy of Science*, **83**, pp. 371–92.

Marshall, J. A. R. [2015]: *Social Evolution and Inclusive Fitness Theory: An Introduction*, Princeton: Princeton University Press.

Maynard Smith, J. [1978]: 'Optimization Theory in Evolution', *Annual Review of Ecology and Systematics*, **9**, pp. 31–56.

Maynard Smith, J. [1991]: 'Honest Signaling, the Philip Sidney Game', *Animal Behavior*, **42**, pp. 1034–5.

Maynard Smith, J. and Harper, D. [2004]: *Animal Signals*, Oxford: Oxford University Press.

McCarty, J. P. [1996]: 'The Energetic Cost of Begging in Nestling Passerines', *The Auk*, **113**, pp. 178–88.

Nowak, M. A. [2006]: 'Five Rules for the Evolution of Cooperation', *Science*, **314**, pp. 1560–63.

Nowak, M. A., Tarnita, C. E. and Wilson, E. O. [2010]: 'The Evolution of Eusociality', *Nature*, **466**, pp. 1057–62.

Nowak, M. A., Tarnita, C. E. and Wilson, E. O. [2011]: 'Nowak *et al.* Reply', *Nature*, **471**, pp. E9–10.

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**, 473–82.

Queller, D. C. [1992]: 'Quantitative Genetics, Inclusive Fitness, and Group Selection', *The American Naturalist*, **139**, pp. 540–58.

Rousset, F. [2002]: 'Inbreeding and Relatedness Coefficients: What Do They Measure?', *Heredity*, **88**, pp. 371–80.

Rubin, H. [2018]: 'The Debate over Inclusive Fitness as a Debate over Methodologies', *Philosophy of Science*, **85**, pp. 1–30.

Searcy, W. and Nowicki, S. [2005]: *The Evolution of Animal Communication: Reliability and Deception in Signaling Systems*, Princeton, NJ: Princeton University Press.

Skyrms, B. [2002]: 'Altruism, Inclusive Fitness, and the Logic of Decision', *Philosophy of Science*, **69**, pp. S104–11.

Taylor, C. and Nowak, M. A. [2007]: 'Transforming the Dilemma', *Evolution*, **61**, pp. 2281–92.

van Veelen, M. [2009]: 'Group Selection, Kin Selection, Altruism, and Cooperation: When Inclusive Fitness If Right and When It Can Be Wrong', *Journal of Theoretical Biology*, **259**, pp. 589–600.

Wagner, E. [2012]: 'Deterministic Chaos and the Evolution of Meaning', *British Journal for the Philosophy of Science*, **63**, pp. 545–75.

West, S. A. and Gardner, A. [2013]: 'Adaptation and Inclusive Fitness', *Current Biology*, **23**, pp. R577–84.

Zollman, K. [2013]: 'Finding Alternatives to the Handicap Principle', *Biological Theory*, **8**, pp. 127–32.