Unlike Agents: The Role of Correlation in Economics and Biology

Hannah Rubin

Abstract

While there are many important similarities between evolution in biology and learning in economics, we should be cautious when importing ideas from one evolutionary context to the other. I will argue that there is a lack of caution is behind the tendency to think of measures of correlation (e.g., 'relatedness') as akin to attitudes of economic agents (e.g., as capturing how much an organism or agent 'values' or 'cares about' a social partner), leading to use of unreliable heuristics and misunderstandings in biology, as well as to misuse of biological results in economics.

1 Introduction

There are striking similarities between natural selection in biology and decision making in economics; both are, in a sense, optimizing processes. As Elliott Sober puts it: "Just as the (objectively) fittest trait evolves, so the (subjectively) best action gets performed" (Sober, 1998, p. 409). Borrowing of ideas and techniques between the two fields has been enormously beneficial and has lead to mathematical frameworks that describe behavior in both fields. For instance, evolutionary game theory can describe evolution of social behaviors both in the biological context describing natural selection (Maynard Smith, 1982) and in the context of economics, sociology, anthropology, etc. describing (boundedly) rational agents learning and updating their strategies (Axelrod, 1984).

However, there are disanalogies between evolution and learning as well: things like genetic constraints generally do not make sense in the context of rational choice, and biological agents do not generally imitate their most successful neighbor, as economic agents often do. Ideally, we ought to outline the limitations of borrowing between fields so we can safely enjoy the benefits while avoiding potential pitfalls of being mislead. While many limitations of the borrowing of ideas between economics and biology have been noted,[1] I will identify an overlooked type of limitation and outline the negative consequences of not paying attention to it.

[1] For instance, because of how traits can be genetically encoded, there limitations to viewing 'mother nature' or selection itself as an economic agent choosing traits to climb peaks of an adaptive landscape (Okasha, 2018).

Specifically, I will argue that there are pitfalls that arise from importing agential descriptions into biology in order to describe measures of 'relatedness' as a measure of 'common interest' between organisms or the degree to which one organism 'cares' about its social partner's reproductive success. While this provides a helpful way to give intuitive explanations for why certain social behaviors are beneficial, there are limitations to the borrowing of the ideas like 'common interest' from the economic context. Below, I will briefly introduce the relevant biological concepts, then argue that overlooking these limitations has lead to at least three problems in biology and economics: reliance on an unreliable heuristic calculation of inclusive fitness, misuse of biological results in economics, and incorrect conclusions regarding the necessity of inclusive fitness for understanding the appearance of design.

2 Inclusive fitness, relatedness, and correlation

Inclusive fitness is seen as essential to explaining the evolution of *social behavior*, where how well an organism does, in terms of its reproductive success, depends both on the trait it has and the traits of its social partners. In the inclusive fitness framework, one looks at the effects an organism has on other organisms' reproductive success, taking into account its 'relatedness' to those organisms, rather than just looking at the organism's own reproductive success. The idea that relatedness between organisms can help explain social behaviors has been part of evolutionary theory since Darwin, but the theory of inclusive fitness introduced by Hamilton (1963, 1964) showed how precisely to take relatedness into account (see Dugatkin, 2007, for a historical overview). To get a picture of how the inclusive fitness framework allows us to view the consequences of social traits in a different way, we can look at how to calculate inclusive fitness and the related concept of neighbor-modulated fitness (which corresponds more closely to our standard notion of fitness as measuring an organism's reproductive success).

Roughly, the neighbor-modulated fitness of an organism is calculated by adding up the number of offspring the organism is expected to have. If we wanted to calculate the neighbor-modulated fitness of, for example an altruist[2] and non-altruist, we would look at all the effects on the focal organism's fitness: whether it pays the cost c and whether it receives the benefit b. The neighbor-modulated fitness of an altruist is

$$-c + P(A_j|A_i) \cdot b$$

This captures the fact that an altruist always pays a cost and has some chance of receiving the benefit. This chance is given by $P(A_j|A_i)$, which is the probability that the organism's social partner (labeled organism j) is an altruist, given that

[2]For simplicity, we can think of an altruistic behavior as one where there is some cost to the focal organism (in terms of fitness – decreasing its survival probability or reproductive output) and some benefit to another. Of course, there are other sorts of behaviors we might care to explain, some of which will be discussed below.

the focal organism (organism i) is an altruist. The neighbor-modulated fitness of a non-altruist is

$$P(A_j|N_i) \cdot b$$

A non-altruist does not pay a cost, but still has some chance of receiving the benefit (which is the probability the social partner is an altruist, given that the focal organism is not). These calculations tell us the expected reproductive success of each type of organism, corresponding to our standard notion of fitness.

Inclusive fitness is an alternative mathematical framework in which fitness calculations track the offspring *caused by* a particular organism, rather than tracking the offspring an organism actually has. The offspring caused by the organism are then weighted according to a 'relatedness' parameter. What relatedness is will be discussed more just a moment, but for now we can note that it measures how likely it is that organisms will have the same trait (or the same genes). Then, the inclusive fitness of an altruist is

$$-c + Rb$$

An altruist affects its own fitness by $-c$ (it pays a cost) and its social partner's fitness by b (it provides a benefit, which we then weight by relatedness, R). The fitness of the non-altruistic trait is zero because it does not affect its own fitness or the fitness of its social partner in any way, relevant to our trait of interest.[3]

The inclusive fitness framework might initially seem counter-intuitive, so it is helpful to mention a basic observation: in general, a trait will increase in frequency when organisms with that trait have more offspring than the average organism in the population. Inclusive fitness gives us the information to determine whether a trait will increase in frequency 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.

It is important to emphasize, for the purposes of this paper, that relatedness is a measure correlation between types. Specifically, R measures how likely it is that the focal organism and its social partner share genetic material, relative to the rest of the population. More specifically, the relatedness of a focal organism (organism i) to its social partner (organism j) is:

$$R = P(A_j|A_i) - P(A_j|N_i)$$

or, the probability the social partner is an altruist given the focal organism is, minus the probability the social partner is an altruist given the focal organism is not.[4]

Though relatedness is a measure of correlation,[5] it is also often described as "a measure of the extent to which... the focal individual values its social

[3]Technically, these fitness calculations include a baseline non-social fitness component, which is omitted here because it is the same for both inclusive fitness and neighbor-modulated fitness.

[4]For a discussion of when this definition of relatedness is equivalent to other common definitions of relatedness derived from the Price equation, see Rubin (2018).

[5]This is something that is agreed upon by inclusive fitness theorists. See, e.g., Marshall (2015), and references therein.

partners..." (West and Gardner, 2013, p. R578). The idea behind this description is that if we are thinking about a focal organism wanting to pass on its genes, and relatedness is telling us how likely it is that the social partner has these same genes, we can think of relatedness as measuring how much a focal organism cares about its social partner's fitness. This is supposed to be in contrast to neighbor-modulated fitness, where the probabilities of interacting with like individuals measures the extent to which "social partners have a similar disposition for altruism" (p. R578).

This interpretation of relatedness as how much a focal organism values its social partners is analogical, or a way of helping us understand how this term, which is a measure of correlation between types, could be used to explain the evolution of altruistic behaviors. However, as I will argue, this reliance on agential language has led to slippage and confusions regarding the transportability of concepts between economic and biological contexts. As an example of this, consider how Kevin Zollman describes the role relatedness plays in the evolution of honest communication in cases where there are conflicts of interest, e.g. between parents and offspring. In this case, 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, however, 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" (2013, p. 130). Instead, since it is well-known that correlations between traits can allow altruism to evolve, he proposes that we look to solutions using correlation and notes that "Relatedness might, beyond inclusive fitness, introduce additional correlation" (p. 131).

Zollman is not alone in contrasting relatedness and correlation in this way; in fact, as I will argue, similar tendencies to construe relatedness as akin to an attitude of an economic agent have caused problems in both biology and economics. First, I will argue that taking this interpretation of relatedness too seriously is a big reason for the reliance on heuristic methods of calculating inclusive fitness which are known to be unreliable. Second, I will argue that it has also lead to misinterpretation in economics of how relatedness might provide an 'exchange rate' for one person's fitness to another's. And finally, I will argue that it is behind a widespread misconception that we need inclusive fitness in order to view social behaviors as adaptations.

3 An unreliable heuristic

The 'simple-weighted-sum' (SWS) method of calculating inclusive fitness, famously used by Maynard Smith ([1991]), says that one can, heuristically, calculate the inclusive fitness of an organism by adding its own payoff and its relative's payoff, weighted by a relatedness parameter, R (sometimes written as k).[6] This heuristic is extremely common, especially in the animal commu-

[6] Compare to the definition given in section 2. Calculating inclusive fitness is often described as first stripping an organism's fitness of all the fitness effects from others, and then adding

nications literature (see, e.g. Johnstone and Grafen (1992); Johnstone (1998); Nowak (2006); Taylor and Nowak (2007); Archetti (2009a,b).)

However, it is generally agreed that this is an incorrect definition (see, e.g. (Grafen (1982); Grafen (1984), Skyrms (2002); Nowak et al. (2010); Okasha and Martens (2016); Birch (2016)). For instance, this heuristic has a well-known problem with double-counting. Say we have two relatives, organism A and organism B, which interact and both have trait X. Under the SWS heuristic, when we calculate the fitness of trait X 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 R) when we take into account B's contribution to the fitness of the trait. We similarly double-count B's fitness.

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.

In fact, the heuristic is often seen as preferable to explicit calculation of inclusive fitness. When payoffs are additive – i.e., 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, like we did in section 2) – the heuristic correctly identifies the Nash equilibrium of a game. Further, the heuristic is easy to generalize to games where payoffs are not additive. 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)). In addition, it can be shown that in a game with non-additive payoffs, the heuristic at least allows one to calculate necessary, but not sufficient, conditions for something to be an equilibrium.[7] 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.

So, the SWS heuristic 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. It would be fine to use this heuristic if we were to be careful in the conclusions we draw and restrict ourselves to identifying possible equilibria.

---

the fitness effects the organism confers on its relatives (Hamilton (1964)). By contrast, the SWS heuristic does not strip away anything and adds in all the social partner's offspring.

[7] In the late 1970s, the usefulness of the heuristic was debated in the context of the hawk-dove game and it was determined that heuristic sometimes gives the correct equilibrium predictions, and in other cases it lets you calculate necessary, but not sufficient, conditions for something to be an equilibrium. See Maynard Smith (1978), Grafen (1979), and Hines and Maynard Smith (1979) for this debate, and Bruner and Rubin (2020) for an overview of the conclusions.

However, the heuristic is not generally used in this sort of restricted way in equilibrium analysis and is often further used in models with evolving populations to give a picture of how evolution in a population is expected to go. In dynamic analyses, the problems with the misinterpretation of relatedness become even more extreme, including both problems with inaccurate predictions and with interpretations and explanations of the evolutionary process.

To see this more clearly, let us look at the Sir Philip Sidney game, which is used in the animal communications literature to investigate biological evolution of communication between relatives. The SWS heuristic is commonly used in this context, following Maynard Smith (1991). In this game, there are two players: a sender and a receiver. The sender can be in one of two states, needy or healthy, each with some probability. Which state the sender is in is known by the sender but not the receiver. The sender has two options to try to communicate with the receiver: send a signal or not. The receiver then observes whether the signal was sent and decides whether or not to donate a resource to the sender. Donation is costly as the receiver is giving up something of value. A healthy individual will benefit from a donation, though not as much as someone in need. Individuals in a population have one strategy for when they are in the sender position (a rule for when they will signal) and one for when they are in the position of the receiver (a rule for when they will donate); we can call a combination of sender and receiver strategies a 'total strategy', which captures what an individual will do in each of their roles.

In this game, the receiver would not ordinarily have any incentive to donate, but the sender always wants the receiver to donate. This is one reason why it is generally thought that relatedness, or, at least partially aligning the sender's and receiver's interests, is important to analyzing this game. Additionally, (if it is the case that receivers generally only donate to needy individuals) senders have incentive to always try to convince the receiver they are needy. Again, we need some alignment of interests in order for 'honest' communication to evolve.

Consider the case where $R = 1$. In a model making use of the SWS heuristic, this translates into the organism caring as much about their social partner's payoff as they do about their own. By contrast, with relatedness appropriately conceived of a measure of correlation, $R = 1$ means that there is perfect correlation, or, that organisms always interact with someone that has the same total strategy as them. In the model with the heuristic, offspring will only signal when it is sufficiently likely their parent will be responsive to signal (that is, if there is a high enough frequency of organisms in the parent population with the strategy to only donate when they get the signal) and parents will only employ this strategy if its sufficiently likely the offspring are communicating honestly. See the red line in figure 1, which shows outcomes for a model using the SWS heuristic.[8] By contrast, when thinking about interactions with perfect correlation, this form of explanation is inappropriate; talking about likelihoods of outcomes does not make sense in this context. Because organisms are always

[8]See Bruner and Rubin (2020) for details ans descriptions of the parameter values chosen. The SWS model is the same as Huttegger and Zollman (2010).

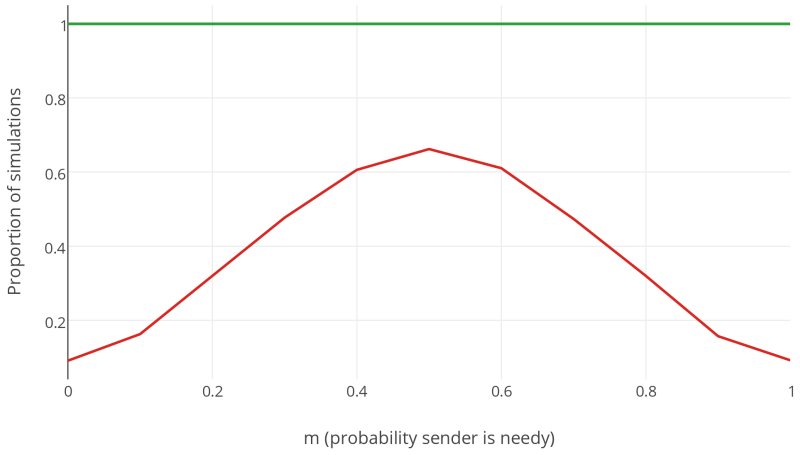

Figure 1: Proportion of simulations which ended with populations using an honest signaling system for a model using the SWS heuristic (red) and a model which incorporates relatedness as a measure of correlation (green).

interacting with others that has the same strategy, whatever total strategy does best against itself is the one that will evolve, as shown by the green line in figure 1.[9]

Thus, we are given two very different pictures of how likely honest communication is for $R = 1$: if relatedness measures correlation, it is the only predicted outcome, but with the SWS heuristic, it may be very unlikely in some cases. So, the heuristic gives a inaccurate picture of which outcomes should be expected and what sorts of explanations are allowed because it ignores the fact that relatedness should measure correlation. Of course, perfect relatedness is merely the case where these issues are most apparent; a similar argument can be given for why the heuristic mis-predicts likelihoods of outcomes in cases where $R < 1$.

In sum, we have demonstrated the first problem that comes from importing ideas which are better suited for learning and decision making: thinking of relatedness as how much an organism cares about its social partner's reproductive success masks problems with and perpetuates use of an unreliable heuristic. Of course, these models using the heuristic might accurately model some scenarios in, e.g., cultural evolution, if we are considering the evolution of honest communication where the people involved care about the others they interacting with, because they have altruistic preferences or something along those lines. However, in calculating reproductive success, we ought not to replace a measure of correlation with a measure of common interest.

[9] Again, see Bruner and Rubin (2020) for further details, including a description of the particular model used and a defense of why it is the appropriate model to contrast with Huttegger and Zollman (2010)'s model using the SWS heuristic.

4 Exchange rates

We will now turn to economics, where I will argue that thinking of relatedness as measuring 'caring' rather than correlation has lead to the misuse of biological results. Until now, we have been speaking of the 'payoffs' in a game as representing fitness effects arising from sort sort of social interaction. Here, we will be speaking of payoffs in terms of 'utility', in addition to their ability to represent fitness effects. In economics, payoffs generally represent what the decision maker cares about, i.e. utility they assign to outcomes, which may be different if we are thinking about economic agents versus biological entities. For instance, all a gene 'cares' about is reproducing itself, while an agent may care about another persons' well-being, fairness, etc. Of course, when we start including things like one agent caring about another, we need some idea of how much; we need some measure of how one person exchanges their own good against another person's.

Coming up with, measuring, and/or explaining these exchange rates – or understanding how *interpersonal comparisons* of utility are made – is central to much of economics, political philosophy, ethics, etc., including Ken Binmore's account of the evolution of our notions of justice. As Binmore explains:

> I don't suppose anyone but the most diehard of neoclassical economists denies that we actually do have standards for making interpersonal comparisons of utility, and that these are widely shared within a particular society... It is easy to guess that the origins of the capacity lie in the need for members of an extended family to recognize how closely they are related to each other (Binmore, 2005, p. 28).

That these interpersonal comparisons have their basis in recognition of family seems plausible enough on the face of it, but the argument Binmore gives rests on confusions about the role relatedness plays in the biological context.

Binmore's discussion of how kin relations gave rise to the exchange rates used in interpersonal comparisons of utility makes use of a game called the prisoners' dilemma. The prisoners' dilemma can be described as a choice between being altruistic or non-altruistic. The payoffs agents get for choosing altruism or non-altruism, depending on the trait of their social partner, are summarized in table 1a. In this game, the rational choice is always to choose *not* to be altruistic: if your social partner is an altruist you get a payoff of b rather than $b - c$ and if your social partner is not an altruist you get a payoff of 0 rather than $-c$. Before getting into Binmore's specific account, it will be useful to recount how rational decision making differs from evolution in this game, and give some terminology to help describe how economic agents could reason about correlation.

First, and most simply, when interactions in a population are random, the evolutionary prediction will be the same as the rational choice for the game: evolution will lead to a population of non-altruists, just as if the organisms were rational agents choosing their traits in order to maximize their fitness. However, interactions in a population are not always random. If there is sufficient correlation between types, the population will evolve to become composed entirely

	Altruist	Not
Altruist	$b-c$, $b-c$	$-c$, b
Not	b, $-c$	0, 0

(a)

	Altruist	Not
Altruist	2, 2	0, 3
Not	3, 0	1, 1

(b)

Table 1: Prisoners' dilemmas. (a) a prisoners' dilemma arising from a choice of whether or not to be altruistic, and (b) the prisoners' dilemma used by Binmore (2005), which is equivalent to assuming $b = 2$ and $c = 1$, which a background fitness of 1.

of altruists, as altruists receive the benefits from others sufficiently more often to outweigh the cost that they are paying. One might think that a rational actor should somehow take the correlation between types into consideration when deciding between traits: one should choose to be an altruist because one would be more likely to receive the benefit from interacting with another altruist. However, decision makers should not generally take correlations into account.

To see why this is a case, let us think about the prisoners' dilemma played with a twin, as discussed by Skyrms (1994) and Sober (1998). You and your twin are apprehended by the police for committing a crime, taken into separate rooms, and each offered a deal from the police: you will get a reduced sentence if you turn state's witness and offer up evidence against your twin. You can each remain silent or turn state's witness. This relates to an altruistic action in the biological case: there is some cost to altruistically remaining silent (you could have reduced your sentence had you turned state's witness) and it also produces some benefit (withholding information means your twin can only be convicted of a lesser crime, for which the sentence is shorter).

Since you and your twin are held in separate rooms, you cannot influence each other's decisions. However, being twins, you believe that you are alike and that your choosing altruism is very good evidence your twin will as well. This means you calculate the conditional probability of their choosing to be altruist given that you yourself choose to be altruist to be high. If it is high enough, you might even choose to be altruistic because you believe the likelihood you will receive the benefit is sufficiently high so as to outweigh the cost you pay.

But the decision to be altruistic is irrational - it yields a worse payoff for no matter what your twin does. There is no reason for you to take correlations into account when evaluating the payoff consequences of your actions, as your actions will not affect anything your twin does. This sort of decision making has been referred to as 'magical thinking', as it seems to assume actions magically affect probabilities we know they cannot affect (Skyrms, 1994). This is the same sort of reasoning that occurs when talking about Newcomb's problem[10] and other related decision problems. Thus, bringing in the decision-theoretic

[10] An actor decides between only taking box B – an opaque box with either $1,000,000 or nothing depending on what a reliable (or perfectly accurate, in some variations) predictor has predicted – or also taking box A – a transparent box with $1,000 in it. If the predictor has predicted the actor will take one box (only B), there will $1,000,000 in box B. If they predicted the actor will take two boxes (both A and B), box B will be empty. Since B is already either empty or filled with money, the actor's choice cannot influence its contents.

distinction between evidential decision theory and causal decision theory will help us understand what is going on here, as it helps us understand differing intuitions in these other decision problems.

First, evidential decision theory tells decision makers to evaluate actions based on their 'news value,' or which action provides evidence that good outcomes will occur. This accounts for the intuition that one should choose to be altruistic in the above prisoners' dilemma: being an altruist is a sign that the other person will also be an altruist. However, many philosophers think that this is problematic. We should choose actions based on their consequences, yet evidential decision theory ignores the difference between cases where there is a solely evidential (or correlational) relationship between an act and an outcome and cases where there is a genuine casual relationship.

Causal decision theory, by contrast, only takes into account the causal consequences of an action and will choose an action based on its efficacy, or which outcomes it will produce. So, evolutionary predictions and rational decision making come apart: a rational decision maker should not generally take correlations between types into account, yet the evolutionary prediction necessarily takes correlations into account in calculations of fitness. (This point will be important again in section 5.)

However, there are many factors one might consider that could change an agent's utility function – the payoffs they assign to the different outcomes. In particular, if one cares about the success of their interactive partner, one may care not just about one's own time in jail, monetary gain, reproductive success, etc. Often, having other-regarding preferences is enough to transform the prisoner's dilemma into a game called the prisoners' delight, where the socially beneficial action is also the rational choice (e.g. table 2). Questions then arise: how do people come to have these other-regarding preferences, and how do they weight their own good versus the good of their interactive partner. According to Binmore, the ability to perform these weightings arises from our ability to exchange our own success against the success of our genetic relatives; relatedness provides an exchange rate of our utility against our social partner's.[11]

Binmore begins his argument by considering a prisoners' dilemma with a twin. He states: "When relatives play a game, the payoffs need to be identified with their inclusive fitnessess rather than their individual fitnesses" (Binmore, 2005, p. 104-5), leading him to calculate payoffs in this game by adding together both payoffs in the corresponding cell from the original prisoners' dilemma in table 1b to arrive at the payoffs in table 2. That is, Binmore is using the SWS heuristic described in section 3 to transform the payoffs of the game. Obviously, this representation of the prisoners' dilemma with a twin differs from Skyrms' and Sober's discussion of how correlation is accounted for in the prisoners' dilemma. Further, if we are thinking of relatedness as a measure of correlation, using it to transform the payoff table ought to strike one as problematic. This is because, as in standard in game and decision theory, probabilities of

[11]There are other ways to talk about relatedness as an exchange rate, e.g. Frank (1998), which are not being argued against here.

	Alt.	Not
Alt.	4, 4	3, 3
Not	3, 3	2, 2

Table 2: Binmore's prisoners' dilemma with a twin

receiving outcomes are not included in the evaluation of outcomes. They are, instead, used to calculate the expected utility of some strategy or decision – following the causal decision theorists' reasoning, one multiplies utilities associated with possible outcomes by the probability of each outcome occurring, in a stage of analysis after the payoffs for the game have already been fixed.

However, it does not seem that Binmore is thinking of relatedness as a measure of correlation, He is instead identifying relatedness with caring about a social partner. In fact, he explains the transformation of payoffs as the same kind that occurs when a game is played between lovers who care about each other's outcomes (Binmore, 2005, p. 108). He then proceeds to say that, since this is a game played with a twin, we ought only to look at the payoffs along the diagonal, where both players are playing the same strategy, and model the strategic scenario as a one-person game because the twins do not choose independently, i.e. their actions are perfectly correlated (Binmore, 2005, p. 108-9). If relatedness is meant to measure correlation, why would we both use it to calculate the payoffs then to determine the likelihoods of receiving those payoffs? Even an evidential decision theorist would not argue we should take correlations into account twice.

Binmore seems to be thinking along the same lines as the quote introduced at the start of this paper, that "Relatedness might, beyond inclusive fitness, introduce additional correlation" (Zollman, 2013, p. 131), which misconstrues the nature of relatedness. This is not to say a project like Binmore's is doomed to fail. One could instead tell a story about how exchange rates emerge through interactions among family members. However, relatedness does not provide us with such an exchange rate separately from its role as a measure of correlation. That sort of thinking likely arises from describing relatedness in agential terms, as a measure of common interest.

5 Maximizing agents

Finally, I will argue that using agential language to describe relatedness hides problems with arguments regarding the usefulness of inclusive fitness for understanding the adaptive value of social behaviors, which rest on its role in thinking of organisms as maximizing agents. The *maximizing agent analogy*, or the *heuristic of personification*, asks us to imagine what trait an organism would choose if they were an agent attempting to maximize their fitness. While there are many types of evolutionary analysis that do not require the maximizing agent analogy, e.g. predicting changes in gene frequencies, it is seen as an

important for preserving Darwin's insight that selection leads to the appearance of design. As Okasha (2018) puts it, thinking about selection is not the same as thinking about adaptation and "[t]he 'fit' of an organism at which Darwin marvelled is about an organism having traits that benefit it, or further is biological goal; the agential idiom is hard to avoid here" (p. 50).[12]

So, we often think of organisms as acting 'as if' they are agents choosing traits to maximize their fitness, and therefore understand those traits as furthering biological goals. When it comes to social behavior, however, the maximzing agent analogy often cannot be straightforwardly applied. As described in the previous section, rational decision makers ought not to take correlations into account, but these correlations affect evolutionary outcomes. The traits we observe are not those that an organism 'trying' to maximize fitness would choose.[13]

Hamilton (1964, 1970) proposed inclusive fitness as a quantity that organisms are selected to maximize. It has since become a standard assumption that inclusive fitness is necessary in order to make sense of the appearance of design when it comes to explaining social behaviors. For instance, it is common to state that: "inclusive fitness... is a quantity that natural selection tends to cause individuals to act as if maximizing, just as Darwinian fitness tends to be maximized in the non-social case" (Grafen, 2009, p. 3137). Or, more explicitly stated, if we are going to think of organisms as maximizing agents "... doing so requires inclusive fitness" (West and Gardner, 2013, p. R579). This idea is so influential that biology students are commonly taught the principle that that natural selection leads to organisms acting as if maximizing their inclusive fitness (Grafen, 2006, p. 559).[14]

Though the usefulness of inclusive fitness does not wholly depend on allowing this sort of agential thinking, its role in the maximizing agent analogy has been seen as a major factor in explaining its popularity. As Okasha et al. (2014) put it:

> The popularity of the inclusive fitness concept in evolutionary biology arises because it allows social behaviour, even when it is individually costly, to be understood from the perspective of an individual organism 'trying' achieve a goal, thus preserving Darwin's insight that selection will lead to the appearance of design in nature. (p.28)

Some claim that inclusive fitness is the only major development in our understanding of adaptations since Darwin proposed the theory of natural selection (West et al., 2011, p. 233) or that recent criticisms of inclusive fitness are irrelevant because inclusive fitness is the only concept of fitness that can play this role in explaining the appearance of design (West and Gardner, 2013, p. R582).

[12] He also argues that agential thinking has advantages over talk about functions. See Okasha (2018, ch. 1) for details.

[13] Skyrms (1994); Sober (1998) both give extended arguments for this conclusion

[14] For instance, the following, as well as many others, all express this basic idea: Birch (2016); Gardner (2009); Grafen (2006, 2009); Okasha et al. (2014); Okasha and Martens (2016); Queller (2011); West et al. (2011); West and Gardner (2013).

How would inclusive fitness allow us to use the maximizing agent analogy, when other concepts of fitness do not? West and Gardner (2013) explain:

> The individual does not, in general, have full control of its neighbour-modulated fitness, as parts of this are mediated by the actions of her social partners. However, the individual does have full control of inclusive fitness, as this is explicitly defined in terms of the fitness consequences for itself and others that arise out of its actions (p. R579).

Queller (2011) also notes: "This focus on what the actor can control allows us to tie into the long biological tradition of thinking of actors, or their genes, as agents. Additionally, it tells us that these agents should appear to be trying to maximize inclusive fitness" (p. 10792).

The basic argument is this: organisms are in control of their inclusive fitness because they are in control of whether they confer the benefit on their social partner, but organisms are not in control of their neighbor-modulated fitness because they are not in control of whether their social partner confers a benefit on them. That is, neighbor-modulated fitness explains the evolution of altruism in terms of 'statistical auspiciousness', or altruism happening to correlate with advantageous social neighborhoods. From a neighbor-modulated fitness point of view, if the organism could choose not to be altruistic, while keeping its social environment fixed, it would always stand to gain by doing so (Birch, 2016).

Using the terminology described in the previous section, we can say that the proponents of this *indispensability argument* are causal decision theorists: a rational actor, or maximizing agent, should only take into account the causal consequences of their decisions, not things like correlations or 'statistical auspiciousness'. Or, as West and Gardner (2013) put it: if we are going to say natural selection leads organisms to appear as if they are trying to maximize their fitness, the concept of fitness we use must be under the organism's complete control, "meaning that it is determined only by the traits and actions of the focal organism. This is because organisms can only appear designed to maximise something that they are able to control." (p. R579).

I will argue in this section that the intuitive appeal of the indispensability argument, and the reason it is so widely accepted, relies on us thinking in terms of the agential interpretation of relatedness as a measure of how much an organism cares about it social partner's reproductive success. While a goal of inclusive fitness maximization is clear when thinking of relatedness in terms of caring, it is less clear how the maximizing agent analogy is supposed to play out when thinking of relatedness as a measure of correlation.

It is important to be clear upfront what I am *not* arguing. I am not arguing that those providing this indispensability argument are misunderstand what relatedness is, only that there is a subtle switch from talking about correlation to valuing social partners, and that this hides the role that correlation plays in the maximizing agent analogy. I am also not arguing that genealogical relatedness is unimportant to understanding the evolution of social behaviors, or that inclusive fitness is not optimized by selection, as will become clear below. Finally, I am

not arguing that biologists should not interpret experimental results in terms of organisms maximizing inclusive fitness. This may in fact be the most intuitive way to understand many effects of social interaction. Rather, I am arguing that the importance of inclusive fitness as *necessary* or *the only way* to understand design in the context of social evolution is overstated and entangled with the problems described in the previous two sections.

I will make my argument in two stages below. First, I argue the idea of an organism's full or complete control is not sufficient to establish that inclusive fitness is required in order to think of organisms as maximizing agents when studying social evolution. Second, I will focus more in depth on the idea of an organism choosing a trait 'as if' they were trying to maximize some sort of fitness. I will argue that there are different ways to describe the context in which this choice is made, that not all of these decision-making contexts lead to inclusive fitness as a unique goal, and that the decision-making contexts in which inclusive fitness is a unique goal are not clearly useful for understanding the appearance of design. I will then conclude the section with some discussion of what can be said about the usefulness of the inclusive fitness framework in the maximizing agent analogy.

5.1 Full control

Recall that a concept of fitness is under an organism's full control when "it is determined only by the traits and actions of the focal organism." (West and Gardner, 2013). That is, the focal organism (and no other organism) must be causally responsible for the fitness consequences of the trait under consideration. For instance, the fitness consequences of an altruistic trait would include both cost and benefit terms terms, and both of these would need to be caused by the focal organism. It seems intuitive to say that the production of the benefit is under the organism's control in the inclusive fitness calculation, but not in the neighbor-modulated fitness calculation, since in inclusive fitness we count the benefit the organism produces, whereas in neighbor-modulated fitness we count the benefit the organism receives.

However, considering actions performed by other organisms is not generally thought to be a problem for the viewing organisms having a 'goal' of maximizing fitness. An example from Sober (1998) will demonstrate the point. We can think about what trait an agent would want to have if they were a zebra and choosing between being a fast or slow runner. The agent would choose to be fast, maximizing fitness by escaping predators, and we can reason that natural selection will lead to a population of fast zebras. Zebras are either fast or slow runners and which trait a zebra has determines how likely it is that it will be eaten by a predator (thus determining its fitness). In this example, the predators are considered part of the environment. They are out there eating and not eating certain zebras. As long as the trait a zebra has causally influences the likelihood that the eating or not eating will be directed towards them, we have no problem seeing their fitness as under their control.

It is the same in the case of social behavior as long as we remember that

the social partners are just part of the environment, out there exhibiting or not exhibiting altruistic behavior. As long as the focal organism's trait influences the likelihood that the altruistic behavior is directed toward it, we should not have any problem saying the its fitness is determined by traits and actions under its control. Perhaps because the social environment includes organisms with similar traits, it is tempting to think of the organism's social partners as choosing to whether or not to bestow benefits and thus think of the benefit being under the social partner's control.[15] However, it is important to remember that, in using the maximizing agent analogy, the organism's social partners are considered part of the environment, not agents in their own right. (Our focal organism is not really an agent either, but we are pretending it is in using the analogy.) The focal organism's interaction with them is just like any other interaction with the environment.

So, whether or not the organism actually produces the benefit ought not to factor in to our judgements about whether a benefit term is under an organism's control. For inclusive fitness, then, what matters for our judgements about the benefit term Rb is not whether b is caused by the traits and actions of the organism, but whether R is. But, of course, relatedness is not, in general under an organism's control. To say that it is assigns the organism casual control over something that is merely correlational. However, claims that the terms in inclusive fitness are under the organism's control focus on the costs incurred and benefits conferred, rather than on relatedness, which is generally described (in this context) as a measure of caring. Here, the agential language hides its true nature as a measure of correlation, possibly leading us to unwittingly confuse correlation with causation in talking about the Rb term in inclusive fitness as being something under the organism's control.

5.2 Decision-making context

One might object to the argument in the previous section on the grounds that I have misconstrued the decision-making context, and that relatedness is supposed to be held constant as we imagine the organism choosing a trait to maximize their fitness, since relatedness is part of the 'social context' or 'environment' of the trait. That is, one might argue, there is good reason to only focus on c and b because those result from the choices of our hypothetical maximizing agent, and there is something about the decision making context for social behaviors that makes the analogy to the non-social case inapt.

To be sure, it is not always clear what type of choice we are meant to envisage our focal organism making when using the maximizing agent analogy. This section will consider three types of choices, or three decision-making contexts, which I believe cover the possibilities that those putting forth the indispensability argument could have in mind: 1. the organism chooses a trait, as well as the associated underlying genetics, with fixed relatedness, 2. the organism

[15] As Lynch (2017) discusses (in the context of heritability debates), the presence of another agent who can be assigned causal responsibility can affect causal attributions (p. 36).

chooses a trait, as well as the associated underlying genetics, without fixed relatedness, and 3. the organisms chooses a phenotype, but not the associated underlying genetics. In looking at the details of the decision-making context, we will see that each context either fails to uniquely pick out inclusive fitness as a maximization goal, or is of unclear use for understanding adapations.

1. Choosing a phenotype and the associated underlying genetics, with fixed relatedness. In this case, we imagine our focal organism deciding its own personal phenotype, assuming that choice is accompanied both by whatever underlying genetics are associated with the trait and by the probabilities of interacting with different social partners. In other words, if the organism chooses to be an altruist, it will have a probability $P(A_j|A_i)$ of interacting with another altruist and if it chooses not to be altruistic it has probability $P(A_j|N_i)$ of interacting with an altruist. So, relatedness is held fixed as opposed to, e.g., the traits of the social partners being held fixed. That is, in this case, an organism choosing an altruistic trait makes it more likely that they interact with altruists (assuming $R = P(A_j|A_i) - P(A_j|N_i)$ is positive). In other words, we might say that the organism's decision casually influences the likelihood that they will interact with another altruist.

This is a case where it makes sense to say that the fitness consequences of the organism's choice of trait are under an organism's control. From the decision maker's point of view, their choice of trait causally influences the likelihood that they will interact with an altruist. If it is sufficiently likely that they will interact with an altruist, the rational decision is the altruistic trait, in line with the evolutionary outcome. But, both inclusive fitness and neighbor-modulated fitness work equally well for the basis of the decision maker's choice because they are both quantities that are under the organism's control. For neighbor-modulated fitness, the fact that an organism receives the benefit with a certain probability is under its control, and for inclusive fitness the fact that it is likely their social partner will share their genes is something that is under its control. This is a similar point to that made by Rosas (2010): "if controlling assortment is the clue to controlling inclusive fitness and if the organism can be credited with it, the organism controls inclusive fitness and neighbor-modulated fitness in one move" (p. 8).

Note that this reasoning holds even if the altruistic trait is conditionally expressed, as in common, for example, in explanations of worker sterility or reproductive helping. In these cases, we treat the trait in question as conditional: 'give help if such and such conditions hold', where those conditions will hold with probability p. An example of relevant conditions would be a case where an individual is stronger than another, and can take the role of a reproductive. If that condition is met, will the weaker organism stay and altruistically help the stronger reproduce?[16] A focal organism is then choosing whether it would

[16]This is in comparison to, for example, treating the relevant choice as between helping or reproducing. See Queller (1996) for a discussion of why it important to consider the fitness of each trait separately, i.e. choosing whether or not to accept the reproductive role when in the position to do so is one trait choice, and choosing to help or leave when not the reproductive is another. Mixing the two traits together leads to seemingly paradoxical results.

help if the appropriate condition is met, and the consequence of choosing the altruistic trait is that there is some chance they will have to pay a cost c to confer a benefit b on their social partner.

For neighbor-modulated fitness, organisms only receive benefits when their social partner expresses the altruistic behavior (which happens when the social partner is an altruist and the conditions are met for it) and they only pay the cost when they themselves express the behavior. For inclusive fitness, an altruist only pays the cost to confer a benefit on a genetic relative when the conditions are met. So, all that happens to our initial calculations is that every term is weighted by p, the probability of the relevant conditions being met.[17] The reasoning employed by our maximizing agents does not change when the expression of the altruistic behavior is conditional. If, as is true in this decision making context, $P(A_j|A_i)$ and $P(A_j|N_i)$, and therefore R remain constant regardless of the focal organism's choice, maximizing neighbor-modulated fitness and inclusive fitness are equivalent goals.

2. Choosing a phenotype and the associated underlying genetics, without fixed relatedness. In this decision-making context, we keep fixed the traits of the organism's potential social partners and let the organism decide only its own personal phenotype (while assuming that the choice of phenotype comes along with whatever underlying genetics are associated with the trait). In this case, the probability of interacting with an altruist does not change based on the focal organism's choice of trait. In other words, $P(A_j|A_i) = P(A_j|N_i)$ and so relatedness is zero. While there may be correlations in the population as a whole, the decision maker's choice does not affect *their* probability of interacting with an altruist, so (in the vein of causal decision theory) they should not take these correlations into account when making a decision – the rational choice here is not to be altruistic, whether the organism uses inclusive fitness or neighbor-modulated fitness as the measure it is trying to optimize.

One might object to this narrow focus on relatedness as correlation for a single trait; surely whole genome relatedness is the important measure to consider? That is, the choice of altruistic/non-altruistic trait is independent of the rest of the organism's genetic makeup and so if whole genome relatedness is still high, we can still think of the organism as increasing their reproductive success indirectly through increasing the reproductive success of their genetic relatives. Whole genome relatedness is certainly relevant to a lot of reasoning surround-

[17] More explicitly, we can calculate the neighbor-modulated fitness based on this probability:

$$\text{NMF}(A_i) = P(A_j|A_i)pb - pc \tag{1}$$

$$\text{NMF}(N_i) = P(A_j|N_i)pb \tag{2}$$

The conditional altruism trait is better when $p[(P(A_j|A_i) - P(A_j|N_i))b - c] > 0$, or $p[Rb - c] > 0$. Since the inclusive fitness of this conditional altruism is $p[Rb - c]$, the conditions for choosing altruism are the same regardless of which fitness concept we use. See Frank (1998) chapter 6 for similar calculations where the traits under consideration are tendencies of being sterile.

These calculations, of course, assume that condition (weak or strong) is independent of whether or not an organism is an altruist. If condition and trait are not independent, we could include a term that measures the influence of trait on condition (or vice versa), but this would not affect the basic conclusions we draw.

ing social behavior, e.g. in determining whether there would be selection for a mutation at an unlinked locus which suppressed altruistic tendency.

However, it is less clear how whole genome relatedness is relevant to understanding altruism as an adaptation in the context of the maximizing agent analogy. In other words, it is hard to see how we are explaining altruism as an adaptation, because in this case behavior is decoupled from inheritance – the altruistic behavior increases the reproduction of some organisms, but those organisms do not tend to have the altruistic gene more often, relative to the population average. If the point is to connect the outcomes of selection with the appearance of design, looking at whole genome relatedness when we know relatedness for the trait in question is zero breaks this connection.

3. Choosing a phenotype but not the associated underlying genetics. In this case, we might say then that the organisms choice does not determine A_i or N_i, but rather *do*-A_i or *do*-N_i. In this case, the choice of trait is totally irrelevant to the social partners' traits or genes. This means that for neighbor-modulated fitness, the probability of receiving b for a focal organism does not depend on their choice, i.e., $P(A_j|do\text{-}A_i) = P(A_j|do\text{-}N_i)$, meaning the rational choice is *do*-N_i, to not be altruistic. Plausibly, only altering behavior would not affect any genetic relatedness between organisms, so we might think of holding fixed both $P(A_j|A_i)$ and $P(A_j|N_i)$, while also holding fixed whether the organism in question has the A or N genotype. This means that for inclusive fitness, R remains constant whatever the organism chooses, meaning they compare $Rb-c$ with 0, and choose to *do*-A_i.

So, it seems that, in this decision-making context, we can think of inclusive fitness as the goal of a maximize agent, where we cannot think of neighbor-modulated fitness in the same way.[18] But this get us the conclusion that inclusive fitness allows us to see how evolution leads to the appearance of design? The problem is that imagining an organism in this decision-making context is not of obvious use for understanding the outcomes of natural selection, which describes changes in gene frequencies or traits over time.

Similar to the issues with using whole genome relatedness when we know relatedness for the trait of interest is zero, we again have a case where behavior is decoupled from inheritance – the altruistic behavior may increase some organ-

[18]For a more explicit argument, we can calculate the neighbor-modulated and inclusive fitness for organisms who choose *do*-A_i or *do*-N_i. Since these choices are independent of genetics, we calculate the probability of receiving a benefit depending on whether the focal has the underlying A or N genotype, with probability $P(A_i)$ or $P(N_i)$:

$$\text{NMF}(do\text{-}A_i) = [P(A_j|A_i)P(A_i) + P(A_j|N_i)P(N_i)]b - c \tag{3}$$

$$\text{NMF}(do\text{-}N_i) = [P(A_j|A_i)P(A_i) + P(A_j|N_i)P(N_i)]b \tag{4}$$

Comparing these two, we can see that the neighbor modulated fitness of *do*-A_i is always c less than *do*-N_i, and so the rational choice is not to be altruistic. By contrast, the inclusive fitness of the altruistic action is $Rb-c$, where:

$$R = [P(A_j|A_i) - P(A_j|N_i)]P(A_i) + [P(N_j|N_i) - P(N_j|A_i)]P(N_i) \tag{5}$$

and of the non-altruistic action is 0. Assuming relatedness is high enough, the rational choice is to be altruistic. (Note that often $P(A_j|A_i) - P(A_j|N_i) = P(N_j|N_i) - P(N_j|A_i) = R$.)

isms' reproductive success, but those organisms do not tend to pass on altruistic genes. Here is one consequence that demonstrates the oddity of the maximizing agent analogy in this case: we would conclude that it is evolutionarily advantageous for organisms without altruistic genes to be altruists. These organisms can maximize their inclusive fitness by providing relatives with the benefit b, it just so happens that they are helping their relatives pass on non-altruistic genes. While we might conclude that, in some sense, altruism is advantageous, it is unclear how this helps us say anything about the *evolution* or *adaptive value* of altruism.[19]

5.3 Maximizing fitness

It is worth repeating that I am not claiming proponents of the indispensability argument misunderstand what relatedness is. Rather, I am suggesting that in providing the intuitive gloss of relatedness as a measure of caring, they push relatedness to the side in considering the consequences of an organism's traits in order to focus on the costs and benefits. In doing to, correlation is allowed to sneak in, disguised as part of (or a weighting of) the benefits to social partners, which are considered to be wholly under the organism's control. Describing relatedness as how much an organism cares about its social partner's reproductive success masks difficulties with using inclusive fitness in the maximizing agent analogy and makes the indispensability argument appear much more straightforward than it actually is: the organism chooses whether to pay a cost c to a confer a benefit b on a social partner depending on how much that value that social partner's reproductive success, R.

However, when we try to give the same argument while describing relatedness as a measure of correlation, things become more complicated. The previous section gave a characterization of when we can expect inclusive fitness to be an organism's unique maximization goal, in terms of how we conceptualize the decision-making context of our focal maximizing agent organism. I then argued that those cases in which inclusive fitness is a unique goal are cases where a choice of trait is decoupled from inheritance, which seems to make the maximizing agent analogy ineffective for reasoning about adaptations or the appearance of design. Proponents of the indispensability argument may defend the usefulness of the maximizing agent analogy in this kind of decision-making context, but it initially seems at odds with how the analogy is usually conceptualized (in terms of attempting to explain ultimate, rather than proximate causes of traits) and would leave us with a more complicated story to tell than the simple, intuitive maximization of $Rb - c$ we get when relatedness is thought of as a measure of how much the organism cares about their social partner's reproductive success.

[19] This is similar to a point made by Okasha (2018, ch. 1), who argues that agential thinking surrounding decisions arising from flexible behavior is aimed at proximate causes (or, providing explanations in terms of physiology or the environment), whereas the maximizing agent analogy aims at ultimate causes (providing evolutionary explanations).

6 Conclusion

It is tempting to think of relatedness in agential terms, as a measure of 'common interest' or 'caring'. If an organism's goal is to pass on its genes, and its social partner is likely to share those genes, then in some sense the organism 'wants' the social partner to succeed or 'cares' about its reproductive success. Relatedness measures how likely it is that organisms share genes, relative to the rest of the population, so it does, in a sense, give as idea of the degree of common interest between them. Noticing this can be useful as a way to intuitively understand how correlation affects the evolution of social behaviors.

However, we can take this reasoning too far and supplant or replace our understanding of relatedness as a measure of correlation with an understanding of it in these agential terms. I have argued that a lack of caution in reasoning about relatedness has lead to reliance on unreliable heuristics (section 3) and misunderstandings (section 5) in biology, as well as to misuse of biological concepts in economics (section 4). Instead, we ought to acknowledge the limitations in borrowing ideas from economics like 'degree of common interest' and importing them into the biological context. These ideas are appropriate to incorporate into evolutionary dynamics when talking about learning among (boundedly) rational agents, but incorporating them into biological evolution can lead us astray.

References

Archetti, M. (2009a). Cooperation and the volunteer's dilemma and the strategy of conflict in public goods games. *Journal of Evolutionary Biology 22*, 2129–2200.

Archetti, M. (2009b). The volunteer's dilemma and the optimal size of a social group. *Journal of Theoretical Biology 261*, 475–480.

Axelrod, R. (1984). *The Evolution of Cooperation*. Basic Books.

Binmore, K. (2005). *Natural justice*. Oxford university press.

Birch, J. (2016). Hamilton's two conceptions of social fitness. *Philosophy of Science 83*(5).

Bruner, J. P. and H. Rubin (2020). Inclusive fitness and the problem of honest communication. *The British Journal for the Philosophy of Science 71*(1), 115–137.

Dugatkin, L. A. (2007). Inclusive fitness theory from darwin to hamilton. *Genetics 176*(3), 1375–1380.

Frank, S. A. (1998). *Foundations of social evolution*. Princeton University Press.

Gardner, A. (2009). Adaptation as organism design. *Biology Letters 5*(6), 861–864.

Grafen, A. (1979). The hawk-dove game played between relatives. *Animal behaviour 27*, 905–907.

Grafen, A. (1982). How not to measure inclusive fitness. *Nature 298*(29), 425–426.

Grafen, A. (1984). Natural selection, kin selection and group selection. In J. Krebs and N. Davies (Eds.), *Behavioural Ecology* (2 ed.). Oxford: Blackwell Scientific Publications.

Grafen, A. (2006). Optimization of inclusive fitness. *Journal of Theoretical Biology 238*(3), 541–563.

Grafen, A. (2009). Formalizing darwinism and inclusive fitness theory. *Philosophical Transactions of the Royal Society of London B: Biological Sciences 364*(1533), 3135–3141.

Hamilton, W. D. (1963). The evolution of altruistic behavior. *The American Naturalist 97*(896), 354–356.

Hamilton, W. D. (1964). The genetical evolution of social behavior i and ii. *Journal of Theoretical Biology 7*, 1–16.

Hamilton, W. D. (1970). Selfish and spiteful behaviour in an evolutionary model.

Hines, W. G. S. and J. Maynard Smith (1979). Games between relatives. *Journal of Theoretical Biology 79*(1), 19–30.

Huttegger, S. and K. Zollman (2010). Dynamic stability and basins of attraction in the sir philip sidney game. *Proc. R. Soc. B 277*, 1915–1922.

Johnstone, R. (1998). Efficacy and honesty in communication between relatives. *The American Naturalist 152*, 45–58.

Johnstone, R. and A. Grafen (1992). The continuous sir philip sidney game: a simple model of biological signaling. *Journal of Theoretical Biology 156*, 215–234.

Lynch, K. E. (2017). Heritability and causal reasoning. *Biology & Philosophy 32*(1), 25–49.

Marshall, J. A. (2015). *Social Evolution and Inclusive Fitness Theory*. Princeton University Press.

Maynard Smith, J. (1978). Optimization theory in evolution. *Annual Review of Ecology and Systematics 9*, 31–56.

Maynard Smith, J. (1982). *Evolution and the Theory of Games*. Cambridge university press.

Maynard Smith, J. (1991). Honest signaling, the philip sidney game. *Animal Behavior 42*, 1034–1035.

Nowak, M. A. (2006). Five rules for the evolution of cooperation. *science 314*(5805), 1560–1563.

Nowak, M. A., C. E. Tarnita, and E. O. Wilson (2010). The evolution of eusociality. *Nature 466*(26), 1057–1062.

Okasha, S. (2018). *Agents and Goals in Evolution*. Oxford University Press.

Okasha, S. and J. Martens (2016). Hamilton's rule, inclusive fitness maximization, and the goal of individual behaviour in symmetric two-player games. *Journal of evolutionary biology*.

Okasha, S., J. A. Weymark, and W. Bossert (2014). Inclusive fitness maximization: An axiomatic approach. *Journal of theoretical biology 350*, 24–31.

Queller, D. C. (1996). The measurement and meaning of inclusive fitness. *Animal Behaviour 51*(1), 229–232.

Queller, D. C. (2011). Expanded social fitness and hamilton's rule for kin, kith, and kind. *Proceedings of the National Academy of Sciences 108*(Supplement 2), 10792–10799.

Rosas, A. (2010). Beyond inclusive fitness? on a simple and general explanation for the evolution of altruism. *Philosophy & Theory in Biology 2*.

Rubin, H. (2018). The debate over inclusive fitness as a debate over methodologies. *Philosophy of Science 85*(1), 1–30.

Skyrms, B. (1994). Darwin meets the logic of decision: Correlation in evolutionary game theory. *Philosophy of Science 61*(4), 503–528.

Skyrms, B. (2002). Altruism, inclusive fitness, and the logic of decision. *Philosophy of Science 69*, S104–111.

Sober, E. (1998). Three differences between deliberation. *Modeling rationality, morality, and evolution* (7), 408–422.

Taylor, C. and M. A. Nowak (2007). Transforming the dilemma. *Evolution 61*(10), 2281–2292.

West, S. A., C. El Mouden, and A. Gardner (2011). Sixteen common misconceptions about the evolution of cooperation in humans. *Evolution and Human Behavior 32*(4), 231–262.

West, S. A. and A. Gardner (2013). Adaptation and inclusive fitness. *Current Biology 23*(13), R577–R584.

Zollman, K. (2013). Finding alternatives to the handicap principle. *Biological Theory 8*, 127–132.