Title
The Explanatory Value of Inclusive Fitness for Evolutionary Theory

Author
Rubin, Hannah

Publication Date
2017

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA,
IRVINE

The Explanatory Value of Inclusive Fitness for Evolutionary Theory

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Philosophy

by

Hannah Rubin

Dissertation Committee:
Professor Simon Huttegger, Chair
Distinguished Professor Brian Skyrms
Assistant Professor Cailin O'Connor

2017

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

While at the University of California, Irvine, I have been supported by a Social Science Merit Fellowship from the School of Social Science. I was supported by a National Science Foundation Grant (No. EF 1038456) in the Summer term of 2013 administered by Simon Huttegger, as well as a National Science Foundation Grant (No. 1535139) in the 2015-2016 academic year administered by Cailin O'Connor. I would like to thank The University of Chicago Press for granting permission to reprint material in chapter 3, originally published in *Philosophy of Science*.

My committee members deserve many thanks for their help with this thesis and encouragement throughout my graduate student career. I could not ask for a better advisor than Simon Huttegger. He has always offered sound advice and made helpful suggestions, while still letting me develop my ideas on my own. From meeting with me before class presentations to help me through mathematical formalisms in my first year, to taking time to check in with me even when overseas, Simon has been there for me. Brian Skyrms, in addition to being a source of knowledge and wisdom, has been incredibly supportive. His stamp of approval for any project is invaluable. Cailin O'Connor has been a mentor to me since I started graduate school, always taking time to give helpful advice and feedback, whether it was comments on papers or just chatting over coffee. From the beginning, she made me feel like my ideas were welcome contributions to the academic community.

I would like to thank the many other professors, at UCI and Mizzou as well as from other universities, who helped me along the way: Jeff Barrett, Jim Weatherall, Kevin Zollman, Kyle Stanford, Sean Walsh, Louis Narens, Don Saari, Jonathan Birch, Paul Weirich, Peter Markie, Zac Ernst, and Jorge Pacheco. I would also like to thank my friends and fellow graduate students for their feedback and support during graduate school: Justin Bruner, Ben Feintzeig, Sarita Rosenstock, Mike Schneider, Elliott Wagner, Jenny Herrera, Bennett Holman, and Aydin Mohseni.

Most importantly, I would like to thank my family for their endless support and encouragement.

CURRICULUM VITAE

Hannah Rubin

EDUCATION

Doctor of Philosophy in Philosophy **2017**
University of California, Irvine *Irvine, CA*

Master of Arts in Social Science **2015**
University of California, Irvine *Irvine, CA*

Bachelor of Arts in Philosophy, Religious Studies **2011**
University of Missouri *Columbia, MO*

ABSTRACT OF THE DISSERTATION

The Explanatory Value of Inclusive Fitness for Evolutionary Theory

By

Hannah Rubin

Doctor of Philosophy in Philosophy

University of California, Irvine, 2017

Professor Simon Huttegger, Chair

At the heart of evolutionary theory is the concept of 'fitness', which is, standardly, an organism's reproductive success. Many evolutionary theorists argue, however, that to explain the evolution of social traits, such as altruism, we must use a different notion of fitness. This 'inclusive fitness', which includes the reproductive success of relatives, is seen as indispensable for studying social evolution. Recently, however, both biologists and philosophers have critically scrutinized its significance. My thesis explores the explanatory value of inclusive fitness, while attempting to resolve significant conceptual confusions. I argue that although inclusive fitness is not necessary for evolutionary explanations, it can nonetheless provide an extremely useful way of conceptualizing the evolutionary process.

Chapter 1

Introduction

At the heart of evolutionary theory is the concept of 'fitness', which is, standardly, an organism's reproductive success. Many evolutionary theorists argue, however, that to explain the evolution of social traits, such as altruism, we must use a different notion of fitness. This 'inclusive fitness', which includes the reproductive success of relatives, is seen as indispensable for studying social evolution. Recently, however, both biologists and philosophers have critically scrutinized its significance. My thesis explores the explanatory value of inclusive fitness, while attempting to resolve significant conceptual confusions.

1.1 History and Use of Inclusive Fitness

The idea that relatedness between organisms can help explain their social behaviors has been part of evolutionary theory since Darwin. For instance, eusocial organisms, in which there is a division of reproductive labor where some organisms do not reproduce, might seem problematic for a theory which posits that traits in a population exist because they help organisms survive and reproduce. However, Darwin noted that this puzzle "...disappears

when it is remembered that selection may be applied to the family, as well as the individual and may thus gain the desired end" ((Darwin, 1959, p. 204), cited in (Dugatkin, 2007, p. 1375)).

The idea that kinship could explain the evolution of social traits was further developed during the modern synthesis. When asked whether he would lie down his life for his brother J.B.S. Haldane famously replied "two bothers or eight cousins," alluding to the fact that 50% of one's genetic material is identical by descent with one's siblings and 12.5% with one's cousin. Beyond this famous quip, Haldane, as well as Ronald Fisher and Sewall Wright (the three founders of population genetics) all discussed the importance of relatedness in explaining the evolution of behaviors that would seem puzzling for evolutionary theory. For instance, Fisher explained the evolutionary advantage of insects being distasteful to their predators: although the distastefulness of the insect can only affect the actions of the predator when the insect is eaten, the death of the insect dissuades the predator from eating nearby insects which tend to be its relatives. These genes then spread in the population due to the increased survival of the insects possessing them. Other self-sacrificial behaviors, like altruism (in the biological sense of performing acts which decrease the fitness of an organism's own fitness and increase the fitness of another or other organisms), can spread through evolution in similar ways (Dugatkin, 2007).

The mathematical formulation of inclusive fitness was first introduced by Hamilton (1964) in order to help explain the evolution of social traits, especially traits evolving via kin selection.[1] In calculating inclusive fitness, one looks at the effects organisms have on other organisms' reproductive success, rather than just looking at the organism's own reproductive success. These effects are then weighted by the 'relatedness' of the organism to those organisms whose fitness it affects. This is in contrast to what is often called 'neighbor-modulated fitness' which is calculated by summing up all the affects on the focal organism's fitness in order to arrive

[1]See Dugatkin (2007) for more details on the history of inclusive fitness theory and an argument for why a mathematical model of the evolution of altruism via kin selection did not arise during the modern synthesis.

at its expected number of offspring. These two ways of calculating fitness will be further explained in section 2.2.2. One of the most famous results associated with inclusive fitness is Hamilton's rule, discussed in section 3.2, which offers a simple way to weigh the costs and benefits of social behaviors in order to predict whether selection will favor the behavior.

Since its conception, inclusive fitness has been extraordinarily useful in producing a new results and insights. For instance, it has helped to give new, intuitive explanations of a variety of traits including altruism, eusociality, parental care, and genomic imprinting (Grafen, 1984; Marshall, 2015, and references therein). Inclusive fitness and Hamilton's rule have also been debated since their introduction. For instance, Cavalli-Sforza and Feldman (1978) and Karlin and Matessi (1983) argued that results derived in the inclusive fitness framework fail to give correct predictions, or that they fail to give exact predictions, in certain cases. Inclusive fitness has also been defined incorrectly in many cases, which has lead to debate over results within the theory and critiques of the conceptual complexity involved in calculating and measuring inclusive fitness.

Following a recent article by (Nowak et al., 2010), the debate surrounding inclusive fitness has surged. These authors argue that inclusive fitness is less general than 'standard' evolutionary theory and that it provides an inadequate framework for describing evolutionary change. These claims, and others made within the article, prompted an enormous response (see (Abbot et al., 2011) for instance, which is signed by over 130 people). In addition to defending inclusive fitness against criticisms, some proponents further claim that the criticisms miss the point of inclusive fitness as an indispensable tool for providing explanations (West and Gardner, 2013).

1.2 Roadmap

This thesis will look at the explanatory value of the inclusive fitness framework for evolutionary theory. Chapter 2 criticizes the widespread assumption in biology that inclusive fitness is essential to explaining the evolution of social traits. Many argue that the *heuristic of personification*, the idea that organisms can be viewed as rational decision makers trying to maximize their fitness, breaks down when analyzing social evolution. When organisms tend to interact with those like them, this 'correlation' between interacting individuals can impact the evolution of social traits. A decision maker taking these correlations into account, however, confuses correlation with causation. Many evolutionary biologists argue that if we think of organisms as trying to maximize their inclusive fitness rather than standard fitness, we can regain the connection between rational choice and evolution. I show, however, that this reasoning is flawed: rather than allowing us to ignore correlations, inclusive fitness calculations merely hide them where they are hard to see, namely, in a relatedness' term.

The fact that inclusive is not necessary for evolutionary explanations based on the heuristic of personification does not mean that inclusive fitness is not useful for evolutionary theory. In fact, the next two chapters defend inclusive fitness against some important criticisms and show how it can provide an incredibly useful way of conceptualizing the evolutionary process when relatedness is a key evolutionary factor.

Chapter 3 analyzes the recent debate surrounding inclusive fitness, which, at face value, is about whether inclusive fitness calculations are as general as standard calculations of fitness. I demonstrate that much of the debate is best understood as being about the orthogonal issue of using abstract versus idealized models. Most proponents of inclusive fitness make use of abstractions, which achieve simplicity by ignoring features of the evolutionary situation. Critics propose using alternative frameworks with idealized models, which simplify by having features not true of any real population (e.g. being infinite in size), arguing that only these

models can properly represent evolutionary change. I prove, contra these claims, that when inclusive fitness is used in these idealized models, it is equivalent to the standard calculation of fitness favored by its critics. I then argue that although inclusive fitness is less useful in idealized models, it nonetheless helps conceptualize evolutionary processes when relatedness is a key evolutionary factor.

Chapter 4 provides a case-study of a situation where inclusive fitness can help us better understand the evolutionary process. Namely, I look at gene-culture co-evolution, the interaction between biological evolution and cultural development in humans. We have evidence that the average relatedness in human groups declined over time: humans went from interacting in small kin groups to larger groups which included non-family members. Using this information about relatedness, I provide an inclusive fitness model of the evolution of altruistic traits where gene-culture co-evolution produces cultural groups with different degrees of altruism. This variation in degrees of altruism, many argue, is necessary to explain the broad spread of altruism throughout human cultures, but has not itself been previously explained.

Finally, chapter 5 concludes. We will see that while inclusive fitness is not necessary for evolutionary explanations, it is nonetheless very useful for conceptualizing evolutionary processes.

Chapter 2

Does Inclusive Fitness Save the Connection Between Rational Choice and Evolution?

2.1 Introduction

Inclusive fitness is seen by many as an indispensable part of evolutionary theory, and the only major development since Darwin proposed the theory of natural selection (Grafen, 2006; West et al., 2011). This is due to the assumption that it allows us to talk about adaptations (traits that are advantageous for an organism, which evolved and are maintained through natural selection) as traits that organisms would choose if they were rational decision makers:

> 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 to achieve a goal, thus preserving Darwin's insight that selection will lead to the appearance of design in nature." (Okasha et al., 2014, p. 28)

In fact, some have gone so far as to claim that the criticism of inclusive fitness is irrelevant because inclusive fitness is the only concept of fitness that can play this role in evolutionary theory:

> ...in order to challenge the inclusive fitness paradigm, it would be necessary to show that there is a more useful maximand (design objective) than inclusive fitness. None of the critiques of inclusive fitness have even suggested an alternative maximand, let alone compared their relative utility [12-20]. These critiques have missed that the major purpose of inclusive fitness is to provide an answer to the question of what organisms should appear designed to maximise. (West and Gardner, 2013, p. R582)

That is, (although these authors also argue that inclusive fitness does not have the drawbacks critics ascribe to it) they claim that it does not matter if inclusive fitness has certain drawbacks; it is indispensable for evolutionary theory because it only it allows us to view organisms as maximizing agents when talking about social behavior.

This chapter will argue against this claim. It will show that, depending on the situation, either inclusive fitness cannot save this connection between rational choice and evolutionary theory or it is not necessary to save it.[1] First, in section 2.2, I will discuss the *heuristic of personification*, which is meant to capture the connection between rational choice and evolution based on the idea that we can view organisms as if they were decision makers

[1]Let me note a quick motivation for thinking this is the case. Inclusive fitness and neighbor modulated fitness are, in many cases, provably equivalent descriptions of evolutionary change. (Neighbor-modulated fitness is equivalent to standard Darwinian fitness when payoffs are additive, as explained in footnote 3.) That is, in certain models of evolution, it can be proven that both of these fitness calculations give the same prediction for the direction of evolutionary change (see Birch, 2016, and references therein) and in other more idealized models it can be proven that these fitness calculations predict both the same direction and magnitude of evolutionary change (van Veelen, 2011). (Chapter 3 will also provide a new proof of the equivalence between these fitness calculations.) These mathematical equivalence results have been used to demonstrate that inclusive fitness is an acceptable method of calculating fitness, but it also means that there are no new predictions coming out of models using inclusive fitness. One implication of these mathematical results that seems to be overlooked is that this makes it hard to see how inclusive fitness is necessary to save the maximizing agent analogy: since inclusive fitness and neighbor-modulated fitness yield equivalent predictions for evolutionary change they are maximized under exactly the same circumstances. It is argued that these mathematical equivalent results do not undermine the case for inclusive fitness: they are irrelevant because it is the different logic being employed in inclusive fitness that makes it indispensable. (West and Gardner, 2013).

trying to maximize their fitness. I will also explain how inclusive fitness differs from the traditional definition of fitness. Then, in section 2.3, I will describe why the heuristic of personification is standardly thought to break down when talking about social evolution, and why many argue that inclusive fitness saves it. In section 2.4, I will expose the flaws in this reasoning in showing that it confuses correlation with causation in a non-obvious way. Section 2.5 will discussion some general lessons that can be drawn for issues in philosophy of science regarding the interpretation of models and section 2.6 concludes.

2.2 Fitness the Appearance of Design

Natural selection describes the differential survival and reproduction of organisms due to differences in the traits of those organisms. This differential survival and reproduction leads to organisms which appear to be designed to fit their environment, or to have traits which serve some functional role for the organisms. We call traits which have this sort of functional role adaptations and it is the goal of evolutionary theory to explain these adaptations.

Fitness is commonly defined as an organism's propensity to survive and reproduce.[2] People often also think of fitness as being connected to some sort of 'fitted-ness' to the environment: there is a match between the natural properties of an organism and its ability to thrive in the environment it finds itself in, where the environment is independent of the organism. While fitness is defined in terms of survival and reproduction, this connection to fitted-ness allows us to explain the appearance of design in nature: organisms appear adapted to a certain environment because natural selection leads organisms with a greater fitness to reproduce more often. Those traits which appear designed will spread throughout the population via natural selection. This helps to answer the question of why an organism has the fitness value we ascribe to it. That is, fitness is not defined as fittedness, but we try to connect it to

[2]Which definition of fitness which we use will not matter for the argument here.

fitnedd-ness in helping to explain and understand the outcomes of evolution.

2.2.1 The Heuristic of Personification

The heuristic of personification, which is also referred to as the individual as maximizing agent analogy, or the rational actor heuristic, or other similar names, provides a way to understand why organisms have the fitness they do. The heuristic expresses a connection between rational choice and evolution, which allows us to see a connection between the trait that evolves and the trait a rational decision maker would choose (the trait that seems best for the environment, or is most fit). We start with the observation that natural selection and rational choice are both optimizing processes: "Just as the (objectively) fittest trait evolves, so the (subjectively) best action gets performed" (Sober, 1998, p. 409).

Paraphrasing from Sober (1998), we can write the heuristic of personification explicitly as follows: a trait will evolve via natural selection, rather than some alternative traits, if and only if an agent rationally deliberating would decide to have the trait over the alternatives (p. 409). This heuristic often helps conceptualize the evolutionary process and allows us to talk as though organisms will evolve to behave as though they are trying to maximize their fitness. For instance, to use Sober (1998)'s example, 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, in order to escape predators, and so we can reason that natural selection will lead to a population of fast zebras.

2.2.2 Inclusive Fitness

This section will briefly introduce inclusive fitness in comparison to standard Darwinian fitness.[3] These two fitness calculations provide alternative ways of partitioning the causal structure of social interactions. A more concrete description of the equations used in both frameworks will be provided below.

Inclusive fitness and the related concept of neighbor-modulated fitness were first proposed by Hamilton (1964). Roughly, the neighbor-modulated fitness, or direct fitness, of an organism is calculated by adding up the number of offspring the organism is expected to have from some social interaction of interest. 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, which is a measure of how likely it is that the focal organism and its social partner share genetic material, relative to the rest of the population. What relatedness is will be discussed more in section 2.4, but for now we can note that it is generally thought of as a measure of the correlation between types in a population (Marshall, 2015).

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 that 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 see how many offspring organisms with that trait will have. Inclusive fitness gives us this information by telling us

[3]For the purposes here, we will only consider games with additive payoffs, where the causal effects of an organism on its social partner's fitness are the same irrespective of the type of its social partner (so we can just sum all these fitness effects up to determine an organism's fitness). In games with this property, standard Darwian fitness, which is described as an organism's propensity to survive and reproduce, is equivalent to what is referred to as 'neighbor-modulated fitness' in the inclusive fitness literature. For a discussion of games with non-additive payoffs, see for example Birch (2014b) and Birch and Okasha (2015). For a further explanation of neighbor-modulated fitness and what it means to require additive fitness components, see Birch (2016).

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.

We can calculate inclusive fitness for a focal organism, i, by looking at the effects from all its social interactions relevant to our trait of interest. When i interacts with other organisms, it affects its own fitness by some amount (s_{ii}) and the fitness of another organism, j, by some amount (s_{ij}). The genotype of organism i also predicts, to a certain extent, the genotype of the social partner j. This relationship is described by the relatedness r_{ij}. We can then calculate inclusive fitness as follows:

$$f_i = \sum_j r_{ij} s_{ij} \tag{2.1}$$

This fitness calculation gives us information about how the population will evolve. It tells us how many offspring are had by organisms with the trait of interest, and since offspring tend to be like their parents, this gives us information about how the composition of a population is expected to change. Note that, although it is sometimes described this way, inclusive fitness is *not* calculated by counting the number of offspring an organism has and then adding all the offspring its relatives have (weighted by relatedness).

Compare the inclusive fitness approach to the neighbor-modulated fitness approach, where we look at an organism, i, and add up the effects of its social interactions on its own number of offspring. The neighbor-modulated fitness of organism i is then calculated as follows:

$$f_i = \sum_j s_{ji} \tag{2.2}$$

where s_{ji} is the effect i's social interaction with j has on i's fitness.[4] This gives us information about how many offspring i is expected to have and, since i's offspring tend to be like i,

[4]Note that the definition of neighbor-modulated fitness looks formally different from inclusive fitness as fitness effects are unweighted, while the fitness effects in inclusive fitness are weighted by a relatedness parameter. This apparent asymmetry disappears at the population level when we calculating the fitness of organisms with a certain trait. See section 3.4.1 for a calculation of neighbor-modulated fitness at the population level. For more information on the calculations of these two types of fitness, see (Frank, 1998, p. 48-9) and Birch (2016).

about how the composition of a population is expected to change.[5]

2.3 Argument for the Indispensability of Inclusive Fitness

The argument that inclusive fitness is indispensable for evolutionary theory proceeds in two parts. First, it is argued that the heuristic of personification breaks down when using neighbor-modulated fitness to explain the outcomes of social evolution: we cannot think of organisms as acting as if they were trying to maximize their neighbor-modulated fitness, breaking the connection between rational choice and evolution. Second, it is argued that we can instead view organisms as acting as if they were trying to maximize inclusive fitness, saving the connection between rational choice and evolution. We will discuss these two parts in turn, then consider one way of formalizing the argument which shows that if a rational decision maker uses inclusive fitness as their measure of utility, then the heuristic of personification is left intact.

2.3.1 Breakdown of the Heuristic

The heuristic of personification can be problematic when applied to the evolution of social behavior, where the environment we are interested in is not something like a habitat, but rather the social environment of the organism (the type of organisms it is interacting with). A simple example to demonstrate this problem is the prisoners' dilemma, discussed by Sober (1998) and Skyrms (1994). In this game, an organism with the altruistic trait pays a cost c and bestows a benefit of b on its social partner, while an organism with the non-altruistic trait

[5]Technically, both inclusive fitness and neighbor-modulated fitness include a baseline non-social fitness component, so these calculations are the fitness effects of the social trait of interest.

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

Table 2.1: Prisoners' dilemma

does not pay the cost or bestow any benefit. The payoff, or utility, each type of organism would get from this type of interaction, depending on the trait of their social partner, is summarized in table 2.1. The utility function describing the how much an actor values each of the outcomes of the interaction is thought of, in the evolutionary context, as the effects on the organism's fitness (e.g. a positive utility represents the organism receiving some material benefit, such as a resource donation, which increases their expected number of offspring).

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 $-c$. 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 organisms without the altruistic trait, just as if the organisms were rational agents choosing their traits in order to maximize their fitnesses.

However, interactions in a population are not always random. There are often *correlations* between types, where organisms are more likely to interact with other organisms of their same type. So, altruists are more likely to interact with other altruists and non-altruists are more likely to interact with other non-altruists for a variety of reasons. For example, this could be because of 'greenbeard' effects, where altruists have some observable trait allowing them to recognize and preferentially interact with other altruists, or because organisms interact with their kin who tend to have the same inherited traits as them. If there is sufficient correlation between types, the population will evolve to become composed entirely of altruists. In this case, the heuristic of personification breaks down; the trait that evolves is not what a rational actor would choose.

	Altruist	Not
Altruist	.5	-1
Not	1.5	0

Table 2.2: Max and Moritz's prisoners' dilemma

	Altruist	Not
Altruist	.45	.05
Not	.05	.45

Table 2.3: Max's beliefs

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 to receive the benefits form interacting with another altruist. However, decision makers should often not take correlations into account. Here is an example provided in Skyrms (1994) to show why. Trouble-makers Max and Moritz are apprehended by the police and forced to play the following prisoners' dilemma. They are taken into separate rooms and offered a deal from the police: they will get a reduced sentence if they turn state's witness and offer up evidence against their partner in crime. Each can remain silent or turn state's witness. They know they will get reduced sentence for providing information, but their sentence will be increased if their partner offers information to the police.

This relates to the altruistic action in the biological case. For the ease of exposition, let us use some particular numbers. Say the cost of altruistically remaining silent is 1 (they could have reduced their sentence by one year had they turned states witness) and the benefit is 1.5 (withholding information means their partner in crime can only be convicted of a lesser crime, for which the sentence is 1.5 years shorter). Then table 2.2 represents the prisoners' dilemma played by Max and Moritz. Remember that Max and Moritz are held in separate rooms so cannot influence each other's decisions. We will look at this decision problem from Max's point of view, but the situation will be symmetric from Moritz's point of view. Max believes that Moritz and he are very much alike and so whatever he chooses Moritz will likely choose the same thing. Table 2.3 represents his beliefs over the likelihoods of the outcomes.

If we add up the probabilities in each column of table 2.3, we see that he assigns .5 probability to Moritz choosing the altruistic action of not turning state's witness (.45 + .05) and also probability .5 to him turning state's witness. This means that the expected payoff for Max of choosing to be altruistic is $.5(1) + .5(-.5) = .25$ and the expected payoff of choosing not to be altruistic is $.5(1.5) = .75$. Since the expected payoff of non-altruism is higher, the rational choice is to be non-altruistic.[6]

If we were to use the conditional probabilities (the probability that Moritz is altruistic given Max is, and so on), we would calculate expected payoff for Max of choosing to be altruistic to be $.9(.5) - .1(1) = .35$ and the expected payoff of choosing not to be altruistic to be $.1(1.5) = .15$. Since the expected payoff of being altruistic would be higher than the expected payoff of choosing not to be altruistic, we would conclude that Max should be altruistic and refuse to turn state's witness.

But the decision to be altruistic is irrational - it yields a worse payoff for Max no matter what Moritz does. And there is no reason for Max to use the conditional probabilities in evaluating the payoff consequences of his actions, as his actions will not affect anything Moritz does. This sort of decision making has been referred to as magical thinking, or voodoo decision theory, as it seems to assume actions magically affect probabilities we know they cannot affect. This is the same sort of reasoning that occurs when talking about Newcomb's problem and other related decision problems. The 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 about Newcomb's problem.[7]

[6]Note that the probability of Moritz taking each action does not need to be .5 for this reasoning to hold.

[7]Newcomb's paradox, or Newcomb's problem, describes a dilemma in decision theory presented by Nozick (1969). In this dilemma, a decision maker is shown a transparent box with one thousand dollars in it and an opaque box which either contains nothing or one million dollars. The agent must choose between taking just the opaque box and taking both boxes. The contents of the opaque box are determined by a predictor. If the predictor predicts the agent will choose to take one box, the opaque box will contain one million dollars and if the predictor predicts the agent will take both boxes, the opaque box will contain nothing. In this case, the decision to take both boxes is a sign that one is the sort of person the predictor would predict to take one box (and is therefore a sign that the opaque box contains a million dollars) but it does not causally influence the prediction since the decision is made after the money has been placed in the box. Thus, it is

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. 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 between the act and the desirable outcome. That is, a decision maker taking these correlations into account exhibits the sort of magical thinking' involved in choosing to take one box in Newcomb's problem: thinking that an action which is a sign of an expected outcome will cause the outcome to occur. Causal decision theory, by contrast, only takes into account the causal consequences of an action. Rather than making decisions based on the news value of an action, a causal decision theorist will choose an action based on its efficacy, or which outcomes it will produce. So the proponents of inclusive fitness must be causal decision theorists when arguing the heuristic of personification breaks down when considering social evolution: 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.

I have argued for causal decision theory here, although I will not have convinced proponents of evidential decision theory, and that is not the main point of the argument provided here. For the evidential decision theorist, I will at least show that the arguments from the proponents of inclusive fitness reveal them to be inconsistently switching between evidential and causal decision theory. Additionally, the general lessons for philosophy of science in section 2.5 will not depend on whether one ultimately believes evidential or causal decision theory is the better theory.

argued, the rational decision is to take both boxes, gaining the one thousand dollars from the transparent box and whatever has already been placed in the opaque box. See Weirich (2016) for further discussion.

2.3.2 Inclusive Fitness Saves the Heuristic

Hamilton (1964, 1970) proposed inclusive fitness as a quantity that organisms are selected to maximize. It has become a standard assumption that inclusive fitness is necessary in order to make sense of the heuristic of personification. "[I]nclusive 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).

West and Gardner (2013) explain that 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 satisfy two criteria. First, it must be the target of selection. They (and others) agree that both inclusive fitness and neighbor-modulated fitness satisfy this criterion. So, we will focus on the second criterion, that the terms in our fitness calculation 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." (West and Gardner, 2013, p. R579). That is, we want to explain adaptations in terms of their casual contribution to the organism's fitness.

As 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).

This sort of argument is quite common. For instance, Queller (2011) expresses the same

idea:

> Inclusive fitness points to cause-effect relations, specifically the various effects caused by the actor's behavior. 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 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.

Note that these authors appear to be causal decision theorists in this argument in claiming to evaluate traits in terms of the casual consequences of having the trait. There are two important features of this argument which are important to note here, which will figure into the argument in section 2.4. First, these authors are talking about the costs and benefits, but ignoring the fact that 'relatedness' also appears in the calculation of inclusive fitness. Second, it is unclear what is meant by 'keeping the social environment fixed'. However, before we get into the general argument against the claim that inclusive fitness is essential for providing evolutionary explanations, it will be helpful to go through a formalization of the argument. This will highlight some features of inclusive fitness calculations in order to help us begin to see why the general argument is wrongheaded

2.3.3 Inclusive Fitness as a Utility Function

Okasha and Martens (2016) claim that "in effect, these authors are arguing that inclusive fitness plays the role of a utility function in rational choice" (p. 473). Their approach to formalizing the argument is to first define a utility function (a description of how much the actor values each of the outcomes), find which trait a rational decision maker would choose, then ask if the rational decision is the same as the evolutionary prediction. "If so, we can conclude that evolution will lead organisms to behave as if trying to maximize the utility function in question" (Okasha and Martens, 2016, p. 475).[8]

So, for instance, the utility function could just be that an actor get utility based on the actual payoffs measuring the amount of reproductive success they have have. Then the payoffs are just as in table 2.1, and the rational choice is to not be altruist because a non-altruist always gets c more than an altruist. However, as we have already said, if there is sufficient correlation, this is not the evolutionary outcome. So organisms do *not* appear to maximize this utility function.

The utility function could also be based on the inclusive fitness of the traits. Using the prisoners' dilemma discussed above, we can calculate the inclusive fitness for each trait. When an organism has the altruistic trait, it affects its own fitness by $-c$ and its social partner's fitness by b, meaning that the inclusive fitness of an altruist is $rb - c$, where r is the relatedness of the organism to its social partner. When an organism does not have the altruistic trait, it does not pay any cost and does not benefit its social partner, so its inclusive fitness is 0. The inclusive fitness payoffs of the interaction are summarized in table 2.4 below.

[8]These authors do note that inclusive fitness is not the unique utility function that yields this agreement between rational choice and evolutionary theory and also argue that there are difficulties with the inclusive fitness in the heuristic of personification when payoffs are non-additive (see footnote 3 for a description of what it means for payoffs to be additive). We will discuss their formalization in order to show why the conclusion the biologists want does not follow, even in the simple case where payoffs are non-additive.

	Altruist	Not
Altruist	$rb - c$	$rb - c$
Not	0	0

Table 2.4: Prisoners' dilemma with inclusive fitness.

In this case, we can see that a rational agent would choose to be altruistic so long as $rb-c > 0$. This is exactly the same as the condition for altruism to spread via evolution. This condition for the spread of altruism is famously known as Hamilton's Rule, and will be discussed more in chapter 3. Generalizing from this example, Okasha and Martens (2016) show that organisms will behave as though they are trying to maximize their inclusive fitness.[9] Okasha and Martens (2016) conclude that "defining utility as inclusive fitness makes the rational actor heuristic valid... This supports the idea that evolution will lead organisms to appear as if trying to maximize their inclusive fitness, just as Hamilton originally argued." (p. 476)

2.4 Why the Argument Fails

Here, I will address what I went over in the previous section in opposite order. First I will argue that the procedure provided by Okasha and Martens (2016) does not show that inclusive fitness saves the heuristic of personification. Then, I will return to the more informal argument provided by many biologists and show that the reasoning is flawed. I will argue that we have to make sense of what decision problem we envisage the decision maker to be facing, and any way we conceive of the decision problem, inclusive fitness does not save the heuristic of personification.

[9]That is, assuming payoffs are additive.

2.4.1 Against the Formalized Argument

As noted in section 2.2.2, the relatedness r is a measure of how likely it is that the focal organism and its social partner share genetic material, relative to the rest of the population. That is, it is a way to capture the correlations between types in a population. More specifically, the relatedness of a focal organism to its social partner 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.[10]

Contrast this to the interpretation provided by West and Gardner (2013): "The r term is a measure of the extent to which the focal individual values its social partners..." (p. R578).[11] 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). However, this interpretation of relatedness is meant to analogical, or a heuristic way of helping us understand how this term, which is a measure of correlation, could be used to explain the evolution of altruistic behaviors.

So, let's replace the r term in table 2.4 with what we know relatedness is, yielding table 2.5. This payoff table should strike one as problematic: we calculate expected payoffs by multiplying payoffs by probabilities of receiving them, so we should not have these probabilities as part of the payoffs. For instance, in the top right cell of the payoff table, $P(A_j|A_i)$ and $P(A_j|N_i)$ appear in the payoff for an altruist. However, the fact that the the payoff appears in the right-hand column means the social partner is a non-altruist, so presumably we can set both of these probabilities to 0, rather than assuming there can be some positive value

[10] Why this is the right definition to use is shown in (Skyrms, 2002).

[11] The idea is that if we are thinking about a focal organism wanting to pass on their genes, and r is telling us how likely it is that the social partner has these same genes, we can think of r as measuring how much focal organism cares about its social partner's fitness.

of relatedness.

	Altruist	Not
Altruist	$[P(A_j\|A_i) - P(A_j\|N_i)]b - c$	$[P(A_j\|A_i) - P(A_j\|N_i)]b - c$
Not	0	0

Table 2.5: Prisoners' dilemma with inclusive fitness, expanded.

At any rate, if we are going to include terms in our payoff table that capture correlations in the population, we can provide terms which would lead a decision maker to choose according to maximizing neighbor-modulated fitness. To see this, consider table 2.6 which includes weights which measure the evidential relevance of the trait. For instance, if the diagonal weights are high, that means it is very likely that an organism's social partner will share its trait. For instance, if $P(A_j|A_i)/P(A_j)$ is high, this captures a situation where the social organism being an altruist makes it more likely that its social partner would be an altruist than if we just took an organism from the population at random.

	Altruist	Not
Altruist	$[P(A_j\|A_i)/P(A_j)] \cdot (b - c)$	$[P(N_j\|A_i)/P(N_j)] \cdot (-c)$
Not	$[P(A_j\|N_i)/P(A_j)] \cdot b$	$[P(N_j\|N_i)/P(N_j)] \cdot 0$

Table 2.6: Prisoners' dilemma with measures of evidential relevance

In this case we can compare the payoff an altruist receives to the payoff a non-altruist receives and see that a rational decision maker will choose the altruist trait whenever $rb - c > 0$, which is the exact same condition given for the rational decision maker choosing traits based on their inclusive fitness (and is the condition for the altruistic trait to spread through evolution).[12]

These weights representing evidential relevance measure how likely it it that my social partner is like me and so can also be interpreted as measuring how much I value my social partner

[12]Following Okasha and Martens (2016), we can calculate this condition by comparing the expected utility of the altruistic trait with the expected utility of the non-altruistic trait, where the decision maker takes into

being like me. This is a similar to what many economists a 'Kantian' utility function, as it represents the degree to which you act in accordance with a maxim you would will to be a universal law (Bergstrom, 1995, 2000). So, we can view an organism as maximizing the following utility function based on the payoffs in table 2.6:

$$u_A = k \cdot (b - c) + (1 - k) \cdot (-c)$$

$$u_N = (1 - k) \cdot b + k \cdot 0$$

with the term k (for 'Kantian-ness'), which gives more weight to the payoffs when you and your social partner are acting according to the same maxim.

Remember that the argument is not that the organisms actually act to maximize either of these utility functions – the Kantian utility function or the inclusive fitness utility function – this is just a claim that we can view them as trying to maximize a particular utility function. Each utility function is adequate for the use of the heuristic of personification, though each lends itself to a different interpretation of what a rational decision maker would be attempting to maximize when choosing a trait. The argument provided here is that we can find a utility function that lets us view organisms as neighbor-modulated fitness maximizers just like we can find a utility function that lets us view organisms as inclusive fitness maximizers. In other words, we cannot, based on Okasha and Martens (2016)'s formalization, conclude that inclusive fitness is indispensable for providing evolutionary explanations.

---

account the probability of interacting with each type (but not any correlations between types):

$$\begin{aligned}
&[P(A_j|A_i)/P(A_j)] \cdot (b - c) \cdot P(A_j) + [P(N_j|A_i)/P(N_j)] \cdot (-c) \cdot P(N_j) > \\
&\qquad [P(A_j|N_i)/P(A_j)] \cdot b \cdot P(A_j) + [P(N_j|N_i)/P(N_j)] \cdot 0 \cdot P(N_j) \\
&P(A_j|A_i) \cdot (b - c) + P(N_j|A_i) \cdot (-c) > P(A_j|N_i) \cdot b + P(N_j|N_i) \cdot 0 \\
&[P(A_j|A_i) - P(A_j|N_i)]b - c > 0 \\
&rb - c > 0
\end{aligned}$$

2.4.2 Against the General Argument

I will now step back from the formalization provided by Okasha and Martens (2016) and argue more generally against the argument for the indispensability of inclusive fitness. Recall that the basic argument for why inclusive fitness is needed for providing evolutionary explanations is that it only includes the causal consequences of the trait. That is, it includes things that are under the organism's control. If a rational decision maker were deciding based on inclusive fitness, they would choose to be altruistic under the same conditions that altruism would evolve. When using neighbor-modulated fitness, by contrast, the explanation for why certain traits evolve depends on whether or not the trait happens to correlate with a beneficial social neighborhood. From this it follows that, if a rational decision maker were deciding based on neighbor-modulated fitness, they would always want to choose *not* to be altruistic if they could do so without altering their social environment, breaking the connection between rational choice and evolution.

I have just argued that there is a way to view organisms as semi-Kantian in caring about how much their social partner is like them, which does not depend on looking at how a trait might happen to be correlated with beneficial social neighborhoods, and which leads to organisms acting as if they are maximizing neighbor-modulated fitness. Now I will move away from the idea of trying to find a utility function which could reconcile rational choice with evolutionary predictions, and go back to cases where an agent merely values the payoffs as measures of reproductive success. More specifically, I will look closely at the idea of 'keeping a social environment fixed'. I will distinguish between different interpretations of what this could mean, contrasting cases where the heuristic of personification fails (for both inclusive fitness and for neighbor-modulated fitness) and cases where the heuristic does not fail. In these cases where the heurisitc of personification remains a useful way to explain evolutionary outcomes, I will explain how neighbor-modulated fitness is just as adequate as inclusive fitness.

There are two obvious ways of thinking about keeping a social environment fixed. First, we could keep the organism's social neighborhood fixed, in that we keep fixed the traits of the organism's potential social partners. In this case, the probability of interacting with an altruist or non-altruist does not change depending on the trait of our focal organism. Second, we could consider cases where relatedness is fixed. In this case, the social neighborhood always consists of organisms who are likely to be similar to the focal organism.

The Heuristic of Personification Fails

Let us consider the first interpretation of keeping an organism's social environment fixed: we keep fixed the traits of the organism's potential social partners. As an example of such a biological situation, we might imagine that organisms interact with their siblings. If the organism is an altruist, its sibling is also likely to be an altruist. This is because one (or both) of its parents were altruists. However, whether or not the focal organism is an altruist has no *causal* influence on their sibling's being an altruist. This is because its parent already has whatever trait it has, independent of what trait the focal organism ends up having. If we imagine a decision maker choosing whether to be altruistic or not in a situation like this, the only thing they would be choosing is their own trait, not the trait of their social partner (their sibling), as in figure 2.1.

Let us consider a simple numerical example to demonstrate the point. Suppose that an organism will have 3 altruist and 7 non-altruist potential social partners, regardless of its trait.[13] So the probability that the focal organism interacts with an altruist is 3/10 and the probability they interact with a non-altruist social partner is 7/10. Since the focal organism's trait does not influence their social partner's trait, we can likewise calculate $P(A_j|A_i) = P(A_j|N_i) = 3/10$ and $P(N_j|A_i) = P(N_j|N_i) = 7/10$. While there may be

[13]This could be thought of as organisms our focal organism has equal probability of interacting with or the frequencies of types it interacts with over its lifetime, whichever interpretation makes more sense in the situation.

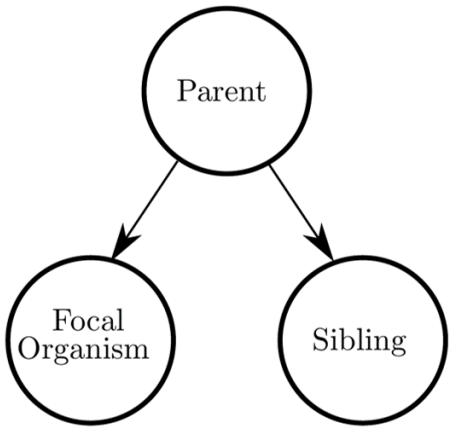


Figure 2.1: Interactions with siblings. Arrows represent causal influence.

correlations in the population as a whole, the decision maker's choice does not affect them, so they should not take these correlations into account when making a decision - the rational choice here is not to be altruistic.

Going through this in detail allows us to see clearly an important consequence for inclusive fitness calculations. Because a rational decision maker should use the above probabilities when making their decision, they should also think of r as being 0 when making their decision. This is because r is defined in terms of these conditional probabilities: $r = P(A_j|A_i) - P(A_j|N_i) = 3/10 - 3/10 = 0$. If the decision maker is only considering the causal influences of their actions, they will thus calculate inclusive fitness to be $rb - c = -c$. *Calculating IF with a positive relatedness confuses correlation with causation.*

This means that a decision maker deciding based on inclusive fitness calculations will conclude that the rational choice is not to be altruistic, just like the decision maker deciding based on neighbor-modulated fitness, or expected utility calculations. Here, the heuristic of personification fails, but it does so even if we think of organisms as attempting to maximize their inclusive fitness. For a causal decision theorist, neither neighbor-modulated nor inclusive fitness will save the heuristic. So, another way to think about where the argument for the indispensability of inclusive fitness goes wrong is that the proponents of inclusive fitness

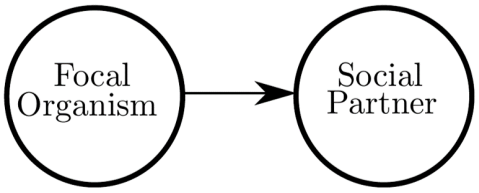


Figure 2.2: Kith selection. The arrow represents causal influence.

are causal decision theorists when arguing against the use of neighbor-modulated fitness in the heuristic of personification and evidential decision theorists when arguing for the use of inclusive fitness.

The Heuristic of Personification Works

Although, as mentioned previously, it is standardly argued that the heuristic of personification breaks down when there are correlations between types, this is not true in *every* case. That is, there are special cases where these correlations capture a causal influence the organism has on its social partner's trait. These types of cases fall under what Queller (2011) calls 'kith selection'. Examples of kith selection are some types of partner choice, where an altruist is more likely to choose an altruist social partner, based on reciprocity or some other mechanism that is not merely correlated with altruism.[14] So, the fact that the focal organism is an altruist makes it more likely that an altruist will seek them out. In other words, the organism's trait casually influences the likelihood that they will interact with another altruist. This situation is captured in figure 2.2.

We can illustrate this case like the last case, with a simple numerical example. Let us suppose that the organism's trait casually affects the likelihood that they will interact with an altruist such that $P(A_j|A_i) = P(N_j|N_i) = 8/10$. We see that the organism is causally

[14]So, for example, many types of greenbeard traits would not fall under this category, where the gene for the physical marker is merely correlated with the gene for altruism. Greenbeard effects would only count here if the altruistic action was somehow causally responsible for the marker, or if one and the same gene controlled both traits (ignoring possibilities for mutations), etc.

responsible for its degree of relatedness, which in this case is $r = P(A_j|A_i) - P(A_j|N_i) = 8/10 - 2/10 = 6/10$. However, importantly, the organism is also causally responsible for the conditional probabilities which figure into the calculation of neighbor-modulated fitness of altruism, $P(A_j|A_i) = 8/10$ and $P(N_j|A_i) = 2/10$.

So, here is a case where it makes sense to use the heuristic of personification 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, both the evolutionary outcome and the rational decision are the altruistic trait. But, here, both inclusive fitness and neighbor-modulated fitness work equally well for the basis of the decision maker's choice. That is, they are both quantities that are under the organism's control. For neighbor-modulated fitness, the fact that it gets the benefit with a certain probability is under its control, and for inclusive fitness the fact that it's likely their social partner will share their genes is something that is under its control.[15]

[15]One might wonder about the actual production of the benefit being under the organism's control. Clearly, in inclusive fitness we count the benefit the organism produces, whereas in neighbor-modulated fitness we count the benefit the organism receives. However, counting actions performed by other organisms as being under the focal organism's control is not generally thought to be a problem in fitness calculations. It is important to remember that the 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 for the purpose of our analogy.) Since the social partners are not to be considered agents, the focal organism's interaction with them is just like any other interaction with the environment.

For instance, consider a populations of moths, some of which have pigmentation that effectively disguises them from birds while others do not. The birds are considered the environment, not agents. They are out there eating and not eating certain moths. As long as the trait the moth has, in terms of its pigmentation allowing it to hide effectively or not, causally influences the likelihood that the eating or not eating will be directed towards them, we have no problem seeing the actor's control condition satisfied, even if the moth itself is not performing the action. 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, doing or not doing altruistic things, and as long as the focal organism's trait influences the likelihood that the altruistic behavior is directed toward them, we should not have any problem saying the actor's control condition is satisfied.

2.5 Discussion

We have shown that there is no support for the widespread assumption that inclusive fitness is necessary in order to view organisms as maximizing agents. We started with a discussion of the role of fitness in evolutionary explanations, and how the heuristic of personification is meant to connect our modern definition of fitness (which talks about chances of survival and reproduction) to the concept of 'fitted-ness' to the environment by explaining traits in terms of what a rational decision maker would choose. This is seen as preserving the notion that natural selection can explain the appearance of design in explaining how we can view evolved social traits as adaptations. I have argued that it is unproblematic to claim that inclusive fitness is maximized organisms act as if their utility function is inclusive fitness, but that we cannot conclude that inclusive fitness is necessary for evolutionary explanations which depend on the heuristic of personification.

What we have seen is that in the cases where maximizing agents should not consider correlations between types in choosing a trait, because they would be confusing a correlation with causation, inclusive fitness fails to save the heuristic of personification. In those cases where maximizing agents are warranted in considering correlations, because their trait will causally influence the trait of their social partner, we showed that one can just as easily use neighbor-modulated fitness. Depending on the situation, either inclusive fitness cannot save the heuristic of personification or it is not necessary to save it, and uncritically accepting inclusive fitness in order to avoid fallacious magical thinking leads to falling right back into magical thinking.

People often make this mistake, I argued, because of the tendency within the inclusive fitness literature to interpret the 'relatedness' parameter as a measure of how much the organism 'cares' about its social partner's fitness. This interpretation of relatedness was meant to analogical, or a heuristic way of helping us understand how this term, which is a measure of

correlation, could the evolution of altruistic behaviors. So, here is one way of understanding how proponents of this argument for the indispensability of inclusive fitness could possibly mistakenly commit this confusion between causation and correlation. What goes into the argument for the indispensability of inclusive fitness is something that was meant to be a heuristic way of understanding, which is taken too seriously as a literal expression of what relatedness is, then fed back into a new analogy (the heuristic of personification) and used to argue for the indispensability of inclusive fitness.

We can use this to draw some general lessons of interest in philosophy of science. In making a model, we provide a mathematical description and some interpretation of the terms that appear in this description to connect them to the biological situations into which they are meant to offer insight. We want our models to provide predictions, but we also want them to provide some sort of explanation, and we often use analogical reasoning to help provide more intuitive explanations. That is, some of the interpretations offer helpful ways to think about the terms in our models, such as thinking of relatedness as how much an organism cares about its social partner, but do not strictly connect the model to the world. We do not think the organisms in our models actually care about the fitness of their social partners, that was just a sort of analogical reasoning. What really connects relatedness to the biological situation at hand is the interpretation as a measure of correlation between types. It is important to be clear about these two roles interpretations of models can play to avoid confusions within the theory. One of these interpretations makes these theories do work, while the other seems to give us some sort of richer understanding, but is not to be taken too seriously.

2.6 Conclusion

We have just clarified and argued against a pervasive argument in one of the most influential research paradigms in evolutionary theory and shown that it rests on a misunderstanding of

what relatedness is, which leads to advocates of this argument unwittingly confusing correlation with causation. We have seen that inclusive fitness is not necessary for evolutionary explanations of social behavior which depend on the heuristic of personification. What does this mean for the inclusive fitness framework?

It does *not* mean that inclusive fitness is not useful for evolutionary theory. In fact, the next two chapters will show how inclusive fitness can provide an incredibly useful way of conceptualizing the evolutionary process when relatedness is a key evolutionary factor. Inclusive fitness not being necessary for evolutionary explanations means, instead, that we have to pay close attention to whether it has the drawbacks ascribed to it by critics in assessing its value for evolutionary theory. It is to those criticism that we now turn.

Chapter 3

The Debate Over Inclusive Fitness as a Debate Over Methodologies

3.1 Introduction

In recent years, an extensive debate has emerged over whether inclusive fitness is an adequate framework for evolutionary theory. Some authors argue that inclusive fitness calculations can be wrong (van Veelen, 2009), while others argue that it requires stringent assumptions and is less general than 'standard' natural selection (Nowak et al., 2010; Wilson, 2012; Allen et al., 2013). The response is that inclusive fitness calculations are not (merely by virtue of using the mathematical framework) susceptible to being wrong (Marshall, 2011) and do not require stringent assumptions like weak selection (Abbot et al., 2011; Marshall, 2015, etc.), additive payoffs (Queller, 1992; Taylor and Maciejewski, 2012; Birch, 2014b; Birch and Okasha, 2015, etc.), pairwise interactions (Taylor and Gardner, 2007; Abbot et al., 2011; Marshall, 2011, etc.), or special population structures (Taylor and Frank, 1996; Taylor and Gardner, 2007; Abbot et al., 2011; Taylor and Maciejewski, 2012; Marshall, 2015, etc.).

Critics of inclusive fitness often propose evolutionary game theory and/or population genetics as alternatives to the inclusive fitness framework (Traulsen, 2010; Nowak et al., 2010, 2011; Allen et al., 2013; Allen and Nowak, 2015). Often, the comparisons are made between very simple models in quantitative genetics, which abstract away from particular details of any given population, and more complex models arising out of population genetics, which often take into account more of the particular details. Here, we will look at how inclusive fitness can function in evolutionary game theory, which often makes idealizations rather than abstractions in order to achieve simple models. The difference between these two modeling strategies (using abstractions versus using idealizations) and how this relates to the inclusive fitness debate will be discussed more in sections 3.3.3 and 3.5. Looking at the way inclusive fitness can be incorporated into evolutionary game theory will help show where some of the disagreements about inclusive fitness arise and when inclusive fitness calculations might be expected to have the limitations ascribed to them by critics. It will also demonstrate how we can think of some parts of the debate as arising from different sides emphasizing different methodologies, rather than as disagreements over inclusive fitness as a way of calculating fitness.

First, I will discuss Hamilton's Rule, and important result in inclusive fitness theory, in connection to the Price equation and kin selection in section 3.2. Then, in section 3.3, I will discuss the debate that has arisen around the inclusive fitness framework, focusing on issues which can be understood as arising from the different sides of the debate emphasizing different methodologies. In section 3.4, I will discuss how models using both neighbor-modulated and inclusive fitness are connected and provide a simple example to demonstrate these connections. Section 3.5 will provide a few ways to think about these connections and explain how they can help us understand some issues in the inclusive fitness debate. Finally, section 3.6 concludes.

3.2 Hamilton's Rule, the Price Equation, and Kin Selection

Recall that Hamilton's Rule, famously associated with inclusive fitness, gives a condition for the increase of an altruistic behavior, where an organism performs an action that decreases its own fitness and increases the fitness of another. (An example of a model of the evolution of altruistic traits will be given in section 3.4.2.) It says simply that if the relatedness-weighted benefit of a trait exceeds its cost, then we should expect selection to favor that trait. That is, the trait is favored when:

$$rb - c > 0 \tag{3.1}$$

where b is the benefit to the focal organism's social partner and c is the cost to the focal organism.

Many results within the inclusive fitness framework, including Hamilton's Rule, are derived from the Price equation, which is a general description of evolutionary change. Let f be the fitness of a trait in the population, relative to the average fitness in the population. Then, the Price equation describes expected evolutionary change in the following way:

$$\dot{E}(p) = Cov(f, p) \tag{3.2}$$

We can think of p as the average phenotypic value of the population, although p can actually represent anything a modeler might want to keep track of: phenotypic value, genetic value, frequency of a trait, etc. $\dot{E}(p)$ is then the change in the average value. The covariance term measures how fitness changes with differences in phenotype.[1]

[1] There is sometimes a second term, $E_f(\dot{p})$, included which measures the fitness-weighted transmission bias, the difference between the phenotypic value of a parent and the average phenotypic value of their offspring. It is often assumed that $E_f(\dot{p}) = 0$, which is generally thought of as assuming there is no transmission bias. (Assuming that $E_f(\dot{p}) = 0$ is not exactly the same as assuming there is no transmission bias (van Veelen, 2005), but the details of what exactly it means to assume $E_f(\dot{p}) = 0$ are not crucial here.)

When fitness effects are additive, that is, when the fitness effects on the recipient do not depend on the recipient's genotype/phenotype and fitness effects from all an organism's social interactions can simply be added up, we can derive equations for both inclusive fitness and neighbor-modulated fitness from the Price equation.[2] These equations are discussed further in the appendix, but here we will look at Hamilton's Rule as derived from the Price equation. The (inclusive fitness version of) Hamilton's Rule is:

$$\dot{E}(g) > 0 \text{ when } \beta_{s_{i-i}p} \cdot \frac{Cov(p, g')}{Cov(p, g)} - \beta_{s_{ii}p} > 0 \tag{3.3}$$

When we can interpret the covariance between an organism's phenotype and its own fitness ($\beta_{s_{ii}p}$) as a 'cost' and the covariance between an organism's phenotype and its social partner's fitness ($\beta_{s_{i-i}p}$) as a 'benefit', we have Hamilton's Rule, where $r = \frac{Cov(p,g')}{Cov(p,g)}$. This measure of relatedness compares the covariance between a focal organism's phenotype, p, and its social partner's genotype, g', with the covariance between the focal organism's phenotype and its own genotype, g (Orlove and Wood, 1978). It is a measure of the degree to which the focal organism and its social partner are genetically related, or how likely it is that the fitness effects from a trait fall on organisms with the gene(s) encoding for the trait. Section 3.4 will discuss how this definition of relatedness connects to the definition of relatedness using conditional probabilities described in chapter 2.

Section 3.4.1 and the appendix discuss how inclusive fitness results derived from the Price equation are related to the replicator dynamics, using methods drawn from Page and Nowak (2002). Section 3.4.2 will discuss how this definition of relatedness matches up with the definition of relatedness we will use in game theoretic models. Section 3.5.2 will discuss versions of Hamilton's Rule which do not rely on the assumption of additive fitness components in relation to the results discussed here.

[2]The additivity of fitness effects requires satisfying these two conditions, which Birch (2016) refers to as *actor's control* and *weak additivity*. Actually, only the second condition is required to derive neighbor-modulated fitness while both are required to derive inclusive fitness. See Birch (2016) for a discussion of this.

Relatedness is often thought of as a measure of the average kinship between interacting organisms when talking about kin selection for a trait. However, as mentioned in chapter 2, it is widely acknowledged that r, and many methods for calculating r, can be thought of as general measures of correlation between types (Marshall, 2015). Sp, r can measure how likely it is that altruists interact with other altruists regardless of whether that correlation is caused by interacting with kin or by some other mechanism, such as a green-beard effect where altruists are able to recognize and preferentially interact with other altruists.

Because inclusive fitness is often used in describing traits that evolve via kin selection, the terms 'inclusive fitness' and 'kin selection' are sometimes used interchangeably. However, it is important to distinguish inclusive fitness from kin selection. Inclusive fitness is a method of calculating fitness, as described above. Kin selection, on the other hand, refers to the selection of a trait due to benefits falling differentially on relatives. Inclusive fitness is a mathematical framework used to describe evolution of a trait; kin selection is a mechanism by which traits can evolve (Hamilton, 1975; Grafen, 2007a, among others).

Some of the critiques of inclusive fitness models are aimed at showing that kin selection has been less important as an evolutionary force than many inclusive fitness theorists presume (see Wilson, 2012, for example). Other parts of the criticism are aimed at the mathematical framework of inclusive fitness itself, such as claims that there are mathematical difficulties with the calculations in inclusive fitness (Nowak et al., 2010; Traulsen, 2010; Wilson, 2012). This chapter will not discuss whether kin selection provides an adequate explanation of prosocial behavior. Instead, it looks at whether inclusive fitness can provide an adequate mathematical framework for use in evolutionary models. Kin selection is discussed only in considering how inclusive fitness can be used in models of traits evolving via kin selection. This focus will help us see which aspects of the debate are relevant to the inclusive fitness framework, and which pertain to kin selection explanations of the evolution of particular traits. Section 3.5 will discuss this further.

3.3 The Debate Surrounding Methods

There are several critiques levied against the inclusive fitness framework. This chapter will address a couple of particularly important critiques which, as we will see, can be understood in light of an emphasis on different modeling techniques: the critiques that inclusive fitness requires the assumption of weak selection and cannot provide dynamically sufficient models. Here, I will give a description of these critiques and a brief motivation for thinking of them as arising from different sides of the debate emphasizing different methodologies. Section 3.5 provides a more detailed argument for this conclusion using material that will be laid out in section 3.4.

3.3.1 Weak Selection

First, inclusive fitness has been critiqued for requiring the assumption of weak selection. In assuming that there is weak selection, we assume that gene frequencies are not changing or that the changes in gene frequencies are small enough to be ignored.[3] This assumption is used in various ways in inclusive fitness models: in employing estimation methods for calculating relatedness, in ignoring higher-order effects or certain types of population structure, etc.

It is easy to see why certain methods of estimating relatedness require weak selection. For example, unless very special conditions hold, estimating relatedness using pedigrees, or family trees, requires that selection is weak. If gene frequencies are systematically changing in the population, the relatedness of an organism to its siblings, for example, will change as the genetic composition of its siblings changes (Grafen, 1984). However, calculating relatedness does not, in general, require weak selection, and we can calculate how relatedness changes

[3]One way to achieve this in a model is to write down fitness as the sum of two components: $f = f_0 + \delta f_x$. One of these, f_0, is the 'background' fitness, the fitness organisms get from things that aren't related to the trait of interest. This is the same for all organisms. The fitness the organisms get from things related to the trait of interest, f_x, is then weighted by a parameter δ and as we take δ to zero, we approach the limit of weak selection. This is what Wild and Traulsen (2007) refer to as 'δ-weak selection'.

as gene frequencies change (Grafen, 1985; Birch, 2014a; Marshall, 2015).

The assumption of weak selection is also used because it allows one to ignore non-additive fitness effects. That is, the assumption of weak selection has been used to ignore things like synergistic effects (where organisms receive additional benefits from cooperation if they both cooperate) or the the effects of competition over resources. This is perhaps the more important use of the assumption of weak selection, as it allows one to separate the way an organism affects its own fitness (a self-effect) from the way it affects its social partner's fitness (an other-effect) in cases where the simplifying assumption of additive fitness components is false. Note that this critique also applies to neighbor-modulated fitness, as the fitness effects are similarly separated into components for self- and other-effect components. At some points in the debate, it seems that critics argue against the use of inclusive fitness (and neighbor-modulated fitness) because it requires weak selection in order to achieve the separation of fitness components. That is, without the assumption of weak selection, one is restricted to a special case in which fitness effects are additive, leading to the conclusion that inclusive fitness is less general than 'standard' natural selection (Nowak et al., 2010).

However, at some points it seems that critics want to claim that, whether or not fitness effects can be split into additive components, inclusive fitness calculations require weak selection. For instance, Nowak et al. (2010) claim that "...inclusive fitness theory cannot even be defined for non-vanishing selection; thus the assumption of weak selection is automatic" (SI 14). It is this second, stronger, claim that will addressed here. In section 3.4, the claim will be shown to be clearly false using modeling techniques from evolutionary game theory, one of the preferred frameworks of critics of inclusive fitness. Section 3.5 will then discuss how, if we read the debate as a debate about inclusive fitness theory as a set of methods rather than inclusive fitness theory as a framework for calculating fitness, we can make sense of this claim.

3.3.2 Dynamic Sufficiency

Inclusive fitness has also been criticized for not being able to provide dynamically sufficient models (Nowak et al., 2010; Wilson, 2012). In a dynamically sufficient model, information about the population at any particular time is enough to make predictions about the population at all future times. So, information about a population at some starting time is enough to be able to predict how the population will evolve at all future times. In a dynamically sufficient model, one can predict whether the population will reach an equilibrium, a state in which the population is no longer evolving, and what the population composition will be at the equilibrium should it reach one.[4] Critics of inclusive fitness argue that it cannot be used to describe the evolutionary trajectories or end points of evolution (Nowak et al., 2010, p. SI4).

One reason this criticism might be leveled against inclusive fitness is the general reliance on the Price equation, which is not dynamically sufficient.[5] More specifically, the Price equation itself is neither dynamically sufficient or insufficient (because it merely expresses a mathematical identity), but it can be either depending on what sort of model it is used with. When we do have a dynamically sufficient model, the Price equation will correctly describe evolutionary change in the model, but will not itself give any additional predictions (van Veelen et al., 2012).

Because many of the results in inclusive fitness theory, like Hamilton's Rule, are formulated in absence of a particular model, and because the focus is often on estimating the covariances rather than calculating them from an evolutionary model, we might not always get

[4]This chapter only deals with deterministic models, but stochastic models can also be dynamically sufficient. A stochastic model is dynamically sufficient when the information about the probability distribution over types at some starting time is enough to predict how the probability distribution will evolve at all future times and to predict the limiting distribution.

[5]Another reason, which will be discussed further in sections 3.3.3 and 3.5.2, is that many of the results which do not rely on the Price equation are focused solely on equilibrium analysis. See, for example, Taylor and Frank (1996).

dynamically sufficient models within the framework. These estimations of parameters will only predict the evolutionary outcome if they do not change over time, which is not the case when selection is frequency dependent (Nowak et al., 2010; Allen et al., 2013). However, as we will see in section 3.4, the regression methods often emphasized in inclusive fitness theory are intimately connected with the sort of dynamically sufficient models preferred by critics of inclusive fitness.

3.3.3 The Debate Over Methodologies

Critics of inclusive fitness then often propose population genetics or evolutionary game theory as alternative frameworks in which one can provide models that are dynamically sufficient and that do not require stringent assumptions like weak selection (Traulsen, 2010; Nowak et al., 2010, 2011; Allen et al., 2013; Allen and Nowak, 2015). It is not immediately clear how we should read this proposal, because although it is true that inclusive fitness tends to be used in quantitative genetics models (Frank, 2013) and is seen as primarily a quantitative method in spirit (Queller, 1992), it has been used in both game theoretic (Skyrms, 2002; van Veelen, 2009, 2011, etc.) and population genetics models (Rousset, 2002; Grafen, 2007b; Lehmann and Rousset, 2014, etc.). In fact, when Hamilton (1964) first proposed using inclusive fitness, he did so in the context of a population genetic model.

The methods used in quantitative genetics are designed to handle continuously varying traits, such as height or weight. In models of social behavior, a continuously varying trait could be the probability of performing an altruistic action. Models within quantitative genetics tend to emphasize simplicity and measurability. These models usually start with observations about phenotypes, or other easily measurable quantities, with few assumptions about the underlying genetics of a trait. This method of modeling involves *abstractions*, ignoring complicating details of the situation by merely leaving them out while still giving a description

that is literally true (Godfrey-Smith, 2009). The Price equation is often used within this approach. As mentioned in section 3.2, many of the common within inclusive fitness theory are derived from the Price equation.

By contrast, challenges to the inclusive fitness framework tend to come from population genetics (Frank, 2013, p. 1153). This is an approach that tends to start with specific assumptions (such as assuming we know the underlying genetics of a trait, the mutation rates, etc.), and make predictions based on these assumptions. Models within this approach tend to be dynamically sufficient, meaning that information about the population at any particular time is enough to make predictions about the population at all future times. The use of simplifying assumptions also means that these models make use of *idealizations* rather than abstractions. That is, they talk about populations which have features we know real populations do not have (e.g. infinite population size, no mutations, etc.) in order to provide a simple model. One way to think about models using idealizations is that they describe non-actual, fictional populations that we take to be similar to real populations in important ways (Godfrey-Smith, 2009). As mentioned, critics propose evolutionary game theory as an alternative to the inclusive fitness framework.[6] The replicator dynamics is often used within this approach. This dynamics requires many idealizing assumptions, which will be discussed in section 3.4.1.

The rest of this chapter will look more closely at the use of inclusive fitness in evolutionary game theory, focusing on the replicator dynamics. Since inclusive fitness is not as commonly used in evolutionary game theory, this will help us see the benefits and drawbacks of using inclusive fitness in highly idealized models. This chapter will also compare how inclusive fitness calculations can be used in evolutionary game theory with some of their uses in quantitative genetics. This comparison between the use of inclusive fitness within these two

[6]Evolutionary game theory and population genetics are sometimes seen as having distinct methods and other times they are seen as more or less continuous (Hammerstein and Selten, 1994, p. 953). They are loosely grouped together here because they are similar in that models within both approaches tend to start which specific assumptions and be dynamically sufficient.

traditions for studying evolution will be helpful in understanding key issues in the debate, since they represent extremes of methodologies using idealizations and abstractions: the replicator dynamics of evolutionary game theory is highly idealized, while the Price equation often employed in quantitative genetics uses only abstractions. We will see how some of the disagreement arises out of the sides of the debate emphasizing different methodologies and how this relates to arguments over the usefulness of Hamilton's Rule.

It is important to note that, while this distinction between abstract models in quantitative genetics and idealized models in evolutionary game theory is illuminating for the present purposes, it does not capture the full variety of modeling techniques within the two methodological traditions. There are evolutionary game theoretic models which make the assumption of weak selection in order to abstract away from genetic details and fail to be dynamically sufficient. For instance, Taylor and Frank (1996) employ a weak selection assumption, allowing them to approximate regression coefficients using partial derivatives, in order to use standard maximization techniques for finding evolutionary stable strategies (p. 28). This method can be used to derive 'approximate' versions of Hamilton's Rule, which will be described further in section 3.5.2.

This chapter will focus on the special case where fitness effects are additive. This is a starting point to examine how inclusive fitness can be calculated in idealized evolutionary game theoretic models and to see if there is any benefit to using inclusive fitness in this context. We will see that the assumption of weak selection is not essential to the calculation of inclusive fitness and that one can build dynamically sufficient models using inclusive fitness. There is, of course, further work to be done to see whether and how this can extend into the more complicated cases generally talked about in inclusive fitness theory. The relationship between these results and general versions of Hamilton's Rule, which do not require weak selection and do not assume additive fitness components, will be discussed in section 3.5.2. Note, however, that while the special case of additive fitness effects will not be

applicable to many traits of interest in the real world, it is an important special case which has been studied extensively in a variety of contexts even outside of the inclusive fitness framework (Eliashberg and Winkler, 1981; Chakraborty and Harbaugh, 2007; Maciejewski et al., 2014, among others).

3.4 Inclusive Fitness in Evolutionary Game Theory

Inclusive fitness and neighbor-modulated fitness are commonly viewed as 'formally equivalent' in that they yield the same predictions in terms of the direction of evolutionary change. That is, they give the same conditions for when a social trait is favored by evolution (see Birch, 2016, and references therein). This section will show that, in the special case discussed above, we can prove further that they also give the same prediction for magnitude of evolutionary change. Section 3.4.1 will prove that the two calculations of fitness are equivalent when used with the replicator dynamics, a standard model from evolutionary game theory. These results are then compared to more common calculations of inclusive fitness in the appendix, which proves the equivalence between the replicator dynamics and both the neighbor-modulated and inclusive fitness calculations derived from the Price equation. Then, section 3.4.2 provides a simple example to illustrate the connections between these fitness calculations.

3.4.1 Inclusive Fitness and Neighbor-Modulated Fitness in Evolutionary Game Theory

In evolutionary game theoretic models, the replicator dynamics is a standard model of the evolutionary process. Under this dynamic, if the fitness of a trait is greater than the average fitness of the population, the frequency of the trait will increase. The traits of interest

dictate behavior in some social interaction, so a trait's fitness is determined by how well it does against the other possible traits in the population (in addition to the population composition). If x_t is the frequency of the trait of interest, and $f_t(x)$ its fitness in a population of composition x, the replicator dynamics is governed by the following equation:

$$\dot{x}_t = x_t[f_t(x) - \bar{f}(x)] \tag{3.4}$$

where $\bar{f}(x)$ is the average fitness in the population. There are a number of assumptions involved in using the replicator dynamics, notably that the population size is infinite and there are a finite number of traits.

Since we are trying to see whether the trait of interest is favored, we can calculate the fitness of organisms which have the trait and the fitness of those that do not in order to have a full description of evolutionary change according to the replicator dynamics. As mentioned, we will look at the case where there are additive fitness effects. If we assume further that there are pairwise interactions, we can denote organism i's social partner as $-i$. In this case, we can write the neighbor-modulated fitness of the organisms with the trait of interest as

$$\begin{aligned} f_t(x) &= P(T_{-i}|T_i) \cdot (s_{ii} + s_{i-i}) + P(N_{-i}|T_i) \cdot s_{ii} \\ &= s_{ii} + P(T_{-i}|T_i) \cdot s_{i-i} \end{aligned} \tag{3.5}$$

where $P(T_{-i}|T_i)$ is the probability an organism with the trait will interact with another organism that has the trait and where $P(N_{-i}|T_i)$ is the probability an organism with the trait will interact with an organism that does not have the trait. Similarly, the neighbor modulated fitness of organisms without the trait of interest is

$$f_n(x) = P(T_{-i}|N_i) \cdot s_{i-i} \tag{3.6}$$

where $P(T_{-i}|N_i)$ is the probability an organism that does not have the trait will interact with another organism that does have the trait.

The inclusive fitness of organisms with the trait of interest is (using w for inclusive fitness

to distinguish it from neighbor-modulated fitness, f)

$$w_t(x) = s_{ii} + rs_{i-i} \tag{3.7}$$

and the inclusive fitness of not having the trait is 0, as in chapter 2. That is, the relatedness between interacting organisms, r, is defined as a difference in conditional probabilities (Skyrms, 2002; van Veelen, 2009; Okasha and Martens, 2016). As in chapter 2, 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:

$$r = P(T_{-i}|T_i) - P(T_{-i}|N_i) \tag{3.8}$$

This is a measure of the degree to which the focal organism's phenotype predicts it's social partner's phenotype.[7] Since genotypes (to a certain extent) predict phenotypes, this can also be thought of as a measure of genetic relatedness.[8]

If we start with the replicator dynamics with neighbor-modulated fitness as our measure of fitness, we can show that it is equivalent to using the replicator dynamics with inclusive

[7]For a demonstration that the assortment rate from Grafen (1979) commonly used in the replicator dynamics is equivalent to a covariance definition of relatedness derived from the Price equation, see (Marshall, 2015, chapter 5, note 1).

[8]Note that relatedness is *not* just the probability that the two organisms share the allele of interest. It is a measure of their genetic similarity relative to the genetic composition of the population as a whole. This is important because in studying altruism, for example, we want to know whether the benefits of altruistic acts fall on altruists sufficiently *more often* than they fall on non-altruists. That is, the benefits must fall on altruists rather than non-altruists with sufficient frequency to give them a reproductive advantage over non-altruists. We will see an example of how r depends on the population's genetic composition in section 3.4.2.

fitness as our measure of fitness:

$$\begin{aligned}
\dot{x}_t &= x_t[f_t(x) - \bar{f}(x)] \\
&= x_t[s_{ii} + P(T_{-i}|T_i) \cdot s_{i-i} - x_t(s_{ii} + P(T_{-i}|T_i) \cdot s_{i-i}) - x_n(P(T_{-i}|N_i) \cdot s_{i-i})] \\
&= x_t[s_{ii} - x_t s_{ii} + P(T_{-i}|T_i) \cdot s_{i-i} - x_t P(T_{-i}|T_i) \cdot s_{i-i} - (1 - x_t)P(T_{-i}|N_i) \cdot s_{i-i})] \\
&= x_t[s_{ii} + (P(T_{-i}|T_i) - P(T_{-i}|N_i))s_{i-i} - x_t s_{ii} - x_t(P(T_{-i}|T_i) - P(T_{-i}|N_i))s_{i-i}] \\
&= x_t[s_{ii} + r s_{i-i} - x_t(s_{ii} + r s_{i-i})] \\
&= x_t[w_t(x) - \bar{w}(x)]
\end{aligned}$$

That is, neighbor-modulated fitness and inclusive fitness are equivalent when used with the replicator dynamics, a standard model of evolution used in evolutionary game theory.[9]

The appendix shows further that, given the assumptions stated above, using the replicator dynamics is equivalent to the Price equation with either method of calculating fitness. That is, the following are equivalent descriptions of evolutionary change:

1. The replicator dynamics used with neighbor-modulated fitness
2. The replicator dynamics used with inclusive fitness
3. The Price equation used with neighbor-modulated fitness
4. The Price equation used with inclusive fitness

The equivalence between (1) and (3) is demonstrated in appendix A.1. The general strategy is the same as the one used in Page and Nowak (2002). First, show that the Price equation used with neighbor-modulated fitness (3) is descriptive of a population evolving according to the replicator dynamics used with neighbor-modulated fitness (1), then show that when there are a finite number of types (3) is also descriptive of a population evolving according (1). Using the same strategy, we can show that (2) and (4) are equivalent. This is done in

[9]For a discussion of the relationship between inclusive fitness and neighbor-modulated fitness in games that do not assume pairwise interactions, but with a constant relatedness, see van Veelen (2011).

appendix A.2. Note that these four ways of modeling evolutionary change are shown to be equivalent in that that they give the same prediction for both the direction and magnitude of evolutionary change. This goes beyond what is commonly meant by the claim that neighbor-modulated fitness and inclusive fitness are equivalent, which is that they give the same prediction for the direction of evolutionary change (see Birch, 2016, and references therein).

The next section provides a simple model using inclusive fitness in the context of evolutionary game theory. This simple illustrative example will let us see, in more concrete terms, the benefits and disadvantages of using inclusive fitness in such an idealized setting. Section 3.5 discusses how to understand these equivalences in the context of the inclusive fitness debate.

3.4.2 A Simple Model: Altruism with Haploid Siblings

This section will provide an idealized model using haploid siblings to show how one can dynamically model relatedness within the inclusive fitness framework when selection is not weak. We will assume that these organisms either have the altruistic trait or not, which is completely determined by whether or not they receive a certain gene from their parent. So that the relationship between this model and Hamilton's Rule is clear, we will assume that when an organism has the altruistic trait, it pays a cost c in order to bestow benefit b on its social partner. When an organism lacks the altruistic trait, it does not pay the cost and does not benefit its social partner. In this model, an organism's social partner is its sibling. Based on these assumptions, we can calculate the inclusive fitness of altruists to be:

$$f_a = -c + Rb \tag{3.9}$$

The inclusive fitness of non-altruists is 0 because they do not perform any action (relevant to our trait of interest) that affects their own or their social partner's reproduction. Thus altruism will spread when $rb - c > 0$.

Since the relatedness of haploid siblings is determined by the genetic material they receive from their common parent, we can let p be the frequency of altruists in the parent generation and use this to calculate relatedness among the offspring. We will also account for a small mutation rate μ in the calculation of relatedness. Once we rewrite the probabilities (according to the definition of conditional probability) so that they are easier to calculate from the assumptions of the model, we can calculate the relatedness of an altruist to its haploid sibling in the following way:

$$\begin{aligned} r &= P(A_{-i}|A_i) - P(A_{-i}|N_i) \\ &= \frac{P(A_{-i}\&A_i)}{P(A_i)} - \frac{P(A_{-i}\&N_i)}{P(N_i)} \\ &= \frac{p(1-\mu)^2 + (1-p)\mu^2}{p(1-\mu) + (1-p)\mu} - \frac{p(1-\mu)\mu + (1-p)(1-\mu)\mu}{p\mu + (1-p)(1-\mu)} \end{aligned}$$

Briefly, here is how to understand this calculation. The numerator of the first term is the probability of two haploid siblings both being altruists. Since there are two ways to get two altruistic offspring, we can calculate this as the probability the parent is an altruist (p) times the probability it has two offspring without mutations ($(1-\mu)^2$), plus the probability the parent is a non-altruist ($1-p$) times the probability it has two offspring which both have a mutation (μ^2). The denominator of the first term is then the frequency of altruists in the offspring generation. These offspring can come from an altruist parent without mutation or from a non-altruist parent with mutation. The second term is calculated similarly. The numerator is the probability that a focal non-altruist will have an altruist sibling: the probability that the parent is an altruist and the focal organism mutates while its sibling does not plus the probability the parent is a non-altruist and the focal organism does not mutate while its sibling does. This is divided by the frequency of non-altruists in the offspring generation.

Figure 3.1 shows how r will change when the population's composition changes.[10] In particular, it shows that relatedness decreases as the population becomes more uniform.[11] To see

[10] This graph was done with a mutation rate of $\mu = 0.1$, which is a fairly high mutation rate. This mutation rate was chosen in order to make the graphs more readable. Results similar to those described in this section can be obtained with much smaller mutation rates.

[11] For a demonstration of this is a more complicated setting, see Rousset (2002).

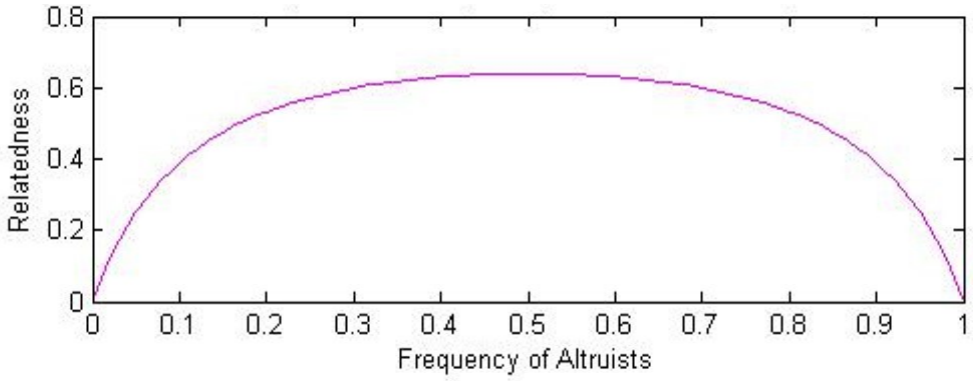


Figure 3.1: Relatedness graphed over the frequency of altruists in the parent population, for $\mu = 0.1$.

why this is the case, it is easiest to look at the extremes of $p = 0$ and $p = 1$. When $p = 0$, the parent population is entirely composed of non-altruists. In the offspring generation, altruists only exist because of mutation. The probability an altruist has an altruist sibling is just μ, the probability that their sibling also has a mutation. However, the probability that a *non-altruist* has an altruist sibling is also μ, the probability that their sibling has a mutation. So $r = P(A_{-i}|A_i) - P(A_{-i}|N_i) = 0$. Similar reasoning applies when $p = 1$. The parent population is composed entirely of altruists, so any non-altruists in the offspring generation arise through mutation. This means that although altruists are likely to have altruist siblings, non-altruists are equally likely to have altruist siblings. So although $P(A_{-i}|A_i)$ is high at $1 - \mu$, $P(A_{-i}|N_i)$ is also $1 - \mu$, and $r = 0$.

We can also calculate relatedness in this model using covariances or regressions. Since phenotypes in this idealized model are completely determined by genotypes (an organism with the altruistic gene is assumed to be an altruist), we can write:

$$r = \frac{Cov(p, g')}{Cov(p, g)} = \frac{Cov(g, g')}{Cov(g, g)} = \beta_{g'g} \tag{3.10}$$

For any population composition, we can perform a regression to calculate the value of r, and it will give the same value of relatedness as the probabilistic definition of relatedness. Figure 3.2 gives a way to visualize why this is the case. In this model, an organism's genetic value, g, is 1 if it has the gene for altruism and 0 if it does not. Thus there are four possible places

for data points on a graph of g versus g': the four corners of the graph. Then, when we do a regression of g on g', what matters is how many data points are in each of these locations. When the focal organisms' genetic value is 1, its social partner's genotype will on average be $P(A_{-i}|A_i)$. Similarly, when the focal organisms' genetic value is 0, its social partner's genotype will on average be $P(A_{-i}|N_i)$. As shown in figure 3.2, this is the intercept of the regression, and the regression coefficient is $\beta_{g'g} = P(A_{-i}|A_i) - P(A_{-i}|N_i)$.

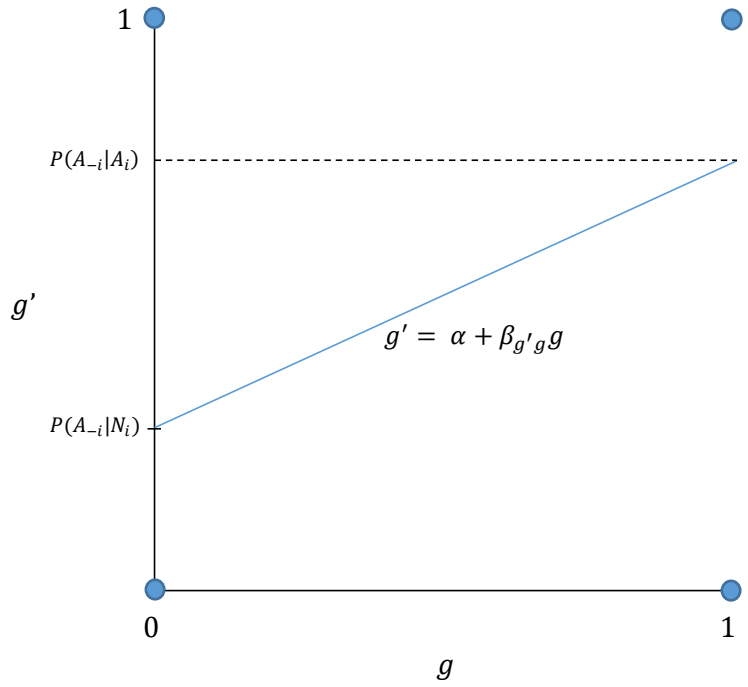


Figure 3.2: An illustration of $\alpha = P(A_{-i}|N_i)$ and $\beta_{g'g} = P(A_{-i}|A_i) - P(A_{-i}|N_i)$.

The inclusive fitness of altruists depends on r, so it also changes as the population composition changes. Figure 3.3 shows how the inclusive fitness of altruists compares with the inclusive fitness of non-altruists over the possible population compositions, for $b = 18$ and $c = 10$. Since relatedness drops off as the population becomes uniform, the inclusive fitness of altruists drops off as the population becomes more uniform. For many possible values of b, c, and μ this means that altruists will have a fitness advantage for some area around $p = 0.5$, but their fitness will drop below the fitness of non-altruists as the population becomes more uniform.

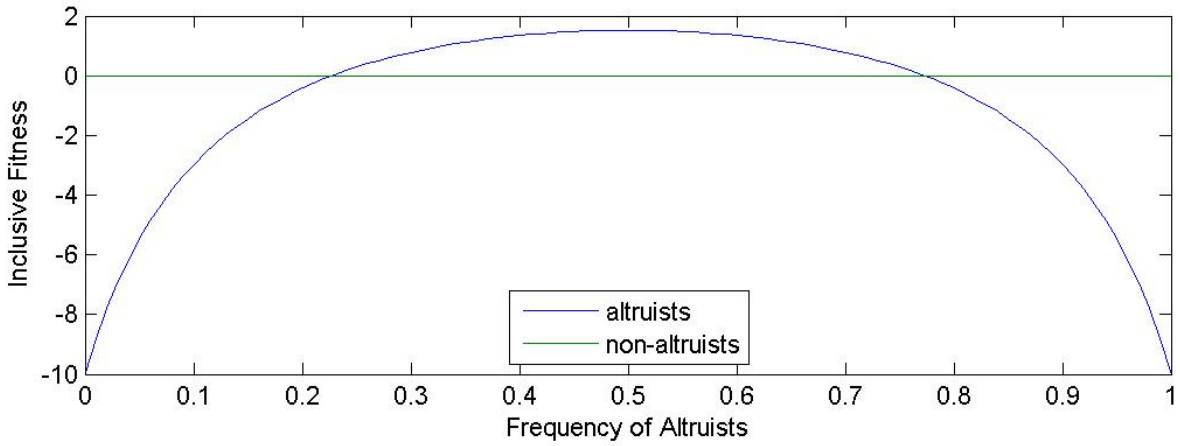


Figure 3.3: Inclusive fitness graphed over the frequency of altruists in the parent population, for $\mu = 0.1$, $b = 18$ and $c = 10$.

These calculations of relatedness and inclusive fitness can be used in a dynamic model where frequencies of genotypes are changing over time; we use these calculations with an appropriate dynamics to see how the population will evolve and to find the equilibria. For this model, we use the selection-mutation dynamics, which is just like the replicator dynamics except that there is an extra term that keeps track of mutations.[12]

Figure 3.4 shows the dynamical analysis of this model, using both inclusive fitness and neighbor-modulated fitness. Figures 3.4(a) and 3.4(b) show, respectively, how the inclusive fitness and neighbor-modulated fitness change as the population composition changes. Figures 3.4(c) and 3.4(d) show the evolutionary trajectories in the population, in terms of the change in frequency of altruists. When this change is positive (when the pink line is above the x-axis, which is represented by the black dashed line in figures 3.4(c) and (d)), altruists will increase in frequency. Likewise when the change is negative, altruists will decrease in frequency. Information about the magnitude of selective pressures is also represented; the further the pink line is from zero, the more selective pressure there is and the faster the

[12]With the selection-mutation dynamics, a population with two types will evolve according to the following equation: $\dot{x}_t = x_t[f_t(x) - \bar{f}(x)] + \mu(1 - 2x_t)$. Note that since this is the same as the replicator dynamics except for the mutation term, which does not depend on the definition of $f_t(x)$, we can prove that using neighbor-modulated fitness and inclusive fitness will be equivalent in the same way as in section 3.4.1.

population composition will change.

Comparing figures 3.4(a) and 3.4(b) shows that the two methods of calculating fitness do yield different numerical values of fitness. However, in comparing the evolutionary trajectory found using inclusive fitness in figure 3.4(c) with the trajectory calculated using neighbor-modulated fitness in figure 3.4(d), one can see that the choice between these fitness measures makes no difference for predicting the evolution of the population, either for the quantitative predictions of the amount of evolutionary change over time or the qualitative predictions about the evolutionary outcomes based on the model. That is, in this simple model, inclusive fitness and neighbor-modulated fitness both give us the same answer when we ask how much altruists will increase or decrease in frequency, across all possible population compositions.

We can also use either type of fitness calculation to find when the change in altruists is zero, when frequencies are not changing and the population is at an equilibrium. With the values of μ, b, and c chosen here, there are four equilibria, two of which are stable: one at about 1% altruists and one at about 75% altruists.[13]

3.5 Discussion

We can see from section 3.4 not only that inclusive fitness is perfectly well-suited for use in evolutionary game theory, but also that weak selection is not a necessary assumption for inclusive fitness calculations and that these calculations can be part of dynamically sufficient models. Some methods of calculating or estimating inclusive fitness may require stringent assumptions, but the calculations in general do not always require extra assumptions. How are we to understand this in the context of the debate over inclusive fitness?

[13] An equilibrium is stable when selective pressures will cause the population to return to the equilibrium if a small amount of drift changes gene frequencies in the population.

3.5.1 Inclusive Fitness with Idealized Models

Some of the disagreement over inclusive fitness can be understood as arising from two sides of the debate emphasizing different methodologies. Recall from section 3.3.3 that inclusive fitness is seen as fundamentally within the quantitative genetics tradition, while critics of inclusive fitness tend to favor population genetics or evolutionary game theory. This means that inclusive fitness theorists tend to favor models which make use of abstractions, leaving details out while still providing literally true general claims about evolution. By contrast, evolutionary game theory, one of the preferred frameworks of the critics of inclusive fitness, tends to provide highly idealized models, making many assumptions which we know are not true of any real population but which allow us to develop a simple model of a fictional population that we think is similar to the real world in important ways.

As section 3.4.1 showed, when there is an infinite population and a finite number of types, inclusive fitness calculations from quantitative genetics and evolutionary game theory are equivalent. Since quantitative methods are designed to handle continuously varying traits, assuming a finite number of types takes the methods out of the context in which they were developed and puts them into the context where dynamically sufficient models can be built. In doing so, we can get models with the kinds of properties valued by critics of inclusive fitness.

One way to think about this is that the regression methods developed by inclusive fitness theorists do not, in themselves, provide models with the properties the critics of inclusive fitness argue evolutionary models should have. However, we can formulate idealized models which are dynamically sufficient and which incorporate selection that is not weak. Then, when we abstract away from the particular details of genetic inheritance or population structure assumed by the simplified models, we arrive at the abstract equations based on the Price equation, which are often used in inclusive fitness theory. Section 3.4.1 (and the appendix)

showed how, when we make simplifying assumptions commonly made in evolutionary game theory, the replicator dynamics and the versions of the Price equation often used in inclusive fitness theory are equivalent descriptions of evolutionary change.

Section 3.4.2 gives an example of how the regression methods commonly used in inclusive fitness can be seen as abstract descriptions of models within evolutionary game theory. In this simple model, we can track how $\frac{Cov(p,g')}{Cov(p,g)}$ changes as the population evolves. Using covariances might seem a bit unnatural in this overly simplified case: because we can calculate the relatedness directly from the assumptions of the model, we do not need to estimate it using the methods of quantitative genetics.

In fact, one might wonder whether there is any benefit to be gained from inclusive fitness in this sort of simplified model. One of the main perceived benefits of inclusive fitness is that it allows modelers to track changes in traits rather than the genes encoding for these traits (which are very difficult to discover empirically) while accounting for genetics by using relatedness (which is often not too difficult to estimate in real populations) (Queller, 1992). Because we abstract away from the mechanisms of genetic inheritance and how the genes encode for the trait of interest, summarizing this with a 'relatedness' parameter, we can develop a phenotypic model that still incorporates genetics in a way that can be empirically easy to measure. That is, one can account for genetics without knowing or making assumptions about the actual underlying genetics of a trait. When we switch to an evolutionary game theoretic or a population genetics model, like the replicator dynamics, we generally then must make assumptions about what these underlying genetics are. We no longer use relatedness to estimate genetic assortment; we can calculate the level of assortment directly.

So, to a certain extent one might think that it is appropriate that the debate over inclusive fitness is a debate over methods: although we can use inclusive fitness in the highly simplified models of evolutionary game theory, in doing so we lose some of the main benefits of

the inclusive fitness framework. The statistical methods used in inclusive fitness make the framework particularly useful, although these methods may require weak selection to split fitness effects into additive components and do not provide dynamically sufficient models. Further, if we view the debate as being about the methods commonly used in quantitative genetics, we can see where these criticisms come from.[14] That is, since inclusive fitness has often been seen as fundamentally quantitative, and since one of the main benefits of inclusive fitness (incorporating genetics in a way that is easy to estimate in real populations) is generally tied up in the statistical methods arising out of quantitative genetics, it makes sense that the debate over inclusive fitness will be in part a debate over methods. However, status of inclusive fitness should not be decided by a debate over the use of methods for which inclusive fitness is seen as particularly beneficial.

The model in section 3.4.2 demonstrates that there can still be some benefit to calculating inclusive fitness rather than neighbor-modulated fitness even in models which are highly idealized, where the level of assortment can be calculated directly. The explanation given for why the population does not evolve to a population composed entirely of altruists was that relatedness drops off as the population becomes more uniformly altruistic. This sort of intuitive explanation is not readily available when using neighbor-modulated fitness. Because the terms describing how the benefits of altruism fall differentially on altruists are split between two different fitness calculations (one for the fitness of altruists and one for fitness of non-altruists), there is no parameter which systemically changes as the population composition changes that we can point to in order to explain why the fitness of altruists drops off as the population becomes more uniform.

[14]This is, of course, not to say that these are the only methods used in inclusive fitness theory, but that the critiques of inclusive fitness are often wrapped up in critiques of the statistical methods (see Allen et al., 2013, for instance).

3.5.2 The Use of Hamilton's Rule

Hamilton's Rule, the most famous result arising out of inclusive fitness theory, has been criticized for not being generally true, for not having any predictive power, and for being misleading in the absence of a particular model (Nowak et al., 2010). There is some truth to these claims. To get Hamilton's Rule in the form $rb - c > 0$, where c and b are interpreted as costs and benefits as described in section 3.2, one has to assume additive fitness components as we have been doing throughout this chapter. If fitness components are not additive, then the rule will not give a correct description of a condition for the spread of a trait. Additionally, if we only have enough information to estimate b, c, and r at a particular point in time, we cannot predict the evolutionary outcome. Further, if $rb - c > 0$ when we estimate these parameters, we might even be mislead into thinking that the population will eventually be entirely altruistic if we forget that the value of any of these parameters can change as the population evolves. However, inclusive fitness theorists will generally agree to this (see Marshall, 2015, for example), but maintain that Hamilton's Rule has both predictive and explanatory power, so it is not immediately clear where the disagreement lies.[15]

The distinction between idealizations and abstractions can again be helpful in understanding part of the dispute. In particular, why should we expect Hamilton's Rule to be true in general? Results derived within population genetics and evolutionary game theory are never true in general, as they rely on idealizations to achieve their simplicity. By contrast, Hamilton's Rule is seen as a general result that is applicable to any real population one might wish to study. This fits well with its prominent role in quantitative genetics, relying on abstractions rather than idealizing assumptions to help provide "the general principles of social evolution theory" (Marshall, 2015, p. xiv).

In this vein, there is emphasis on providing a version of Hamilton's Rule that is generally

[15]See (Marshall, 2015, chapter 6, note 9) for an example of an inclusive fitness model where parameters can change as the population evolves.

true. Hamilton's Rule can be given in a very general form in which we do not have to assume any particular population structure or additive payoff affects (Gardner et al., 2011). Birch (2014b) and Birch and Okasha (2015) describe this in detail, but we can think of the 'cost' and 'benefit' terms in the rule as statistical associations between an organism's fitness and its own genotype (a self-effect) and its social partner's genotype (an other-effect), respectively. This general version of Hamilton's Rule is true of any population. "In effect, this is because we are abstracting away from the complex causal details of social interaction to focus on the overarching statistical relationship between genotype and fitness" (Birch and Okasha, 2015, p. 24).

The question is then whether this version of Hamilton's Rule has any predictive power. It can have predictive power if its components can be understood causally instead of just statistically. That is, if the self-effect and other-effect terms can be interpreted as ways in which the focal organism causally contributes to its own and its social partner's fitness, we have a model that can be used to make predictions rather than just a statistical summary of evolution within a population. However, as Birch and Okasha (2015) explain, it is not entirely clear when a causal interpretation can be provided.

There are, however, a variety of different rules that go under the name 'Hamilton's Rule', each of which follows from different assumptions about the evolutionary process. We can describe these versions of Hamilton's Rule as falling into three categories. There are 'special' versions of the rule (where the b and c terms are interpreted as payoffs in a model) and 'approximate' versions (which provide marginal approximations of the general versions of the rule) in addition to the 'general' version described above (Birch and Okasha, 2015).

In the version of Hamilton's Rule in section 3.3.1, the b and c terms are interpreted as payoff from a game, or parameters in the model, so this can be thought of as a special version of Hamilton's Rule. The fact that we derived a condition $bR - c > 0$ for the spread of altruism depends on the particular payoff structure of the model. If there were non-additive

payoffs, we would have derived a different condition for the spread of altruism. Section 3.4 (and the appendix) illustrated how these general versions of Hamilton's Rule describe the special versions from particular models. As mentioned in sections 3.3.1 and 3.3.3, there are also approximate versions of Hamilton's Rule that require the assumption of weak selection to calculate relatedness or in order to split fitness effects into additive components. Thus, these rules abstract away from the particular payoff structure and so describe a wider range of cases than special forms of the rule. The assumption of weak selection, then, provides some restriction on the conditions under which approximate versions of Hamilton's Rule will apply, but allows us to give an approximately correct condition for the spread of a social behavior for arbitrary payoff structures. (See Birch and Okasha (2015) for more discussion.)

Note that both general and approximate versions of Hamilton's Rule apply for arbitrary payoff structures, but neither are dynamically sufficient. They instead allow us to perform a static analysis, comparing fitnesses at specific points in the evolutionary process (usually the points of interest are equilibria). Since this chapter has looked at how inclusive fitness is used in the replicator dynamics compared with approaches based on the Price equation, it has focused on the the contrast between abstract models in quantitative genetics and idealized models in evolutionary game theory. However, that the critics of inclusive fitness prefer dynamic models over these static modeling techniques is perhaps the more fundamental disagreement in the debate.

There is the additional issue of interpreting the r parameter in Hamilton's Rule. Although many inclusive fitness theorists recognize that r in inclusive fitness calculations can be thought of as a general measure of correlation, Hamilton's Rule is still usually presented as a condition for the evolution of a trait by kin selection. However, this is an additional opportunity for Hamilton's Rule to be misleading; a suggested biological or causal interpretation of the parameter might be unwarranted. Some criticisms seem to assume that Hamilton's Rule is only useful when r is a measure of kinship (Nowak et al., 2010). The

thought behind these sort of critiques of Hamilton's Rule seems to be that when r does not have an intuitive biological interpretation, it is not clear what explanatory power is gained from forcing terms into this particular inequality. The power of Hamilton's Rule then comes from using something like the statistical definitions of relatedness provided here and estimating relatedness using measures of kinship, like pedigrees.[16]

Since the statistical definitions of relatedness are historically explained and used as measures of kinship, adopting Hamilton's Rule as a starting point might seem to suggest an interpretation in terms of kin selection and may lead to theorists ignoring other mechanisms that generate assortment between types. Connecting the statistical and probabilistic definitions here is one way of emphasizing how the association between 'relatedness' and r is contingent: r just measures differences in conditional probabilities of interacting with certain types in the population. In this context, Hamilton's Rule might be thought of as a convenient mathematical description of the fact that there must be sufficient assortment between types in order for a trait such as altruism to evolve, a general point that has been made without the use of Hamilton's Rule (see Skyrms, 1996, for example). A fully specified (but idealized) model, like the one in section 3.4.2, can connect r in Hamilton's Rule to kinship, giving it a meaningful biological interpretation.

This is in line with one suggestion to avoid wrongly interpreting results in terms of kin selection, advanced by Taylor and Frank (1996) and Frank (2013), among others: formulate and analyze a model first, then afterwards use Hamilton's Rule to give an intuitive explanation of the results if appropriate. This allows us to set up the model with whatever mechanism of assortment we think is plausible, then use Hamilton's Rule if it helps illuminate important aspects of the causal structure.

[16]There are of course, other issues with applications of Hamilton's Rule aside from interpreting r in terms of kinship. Often in more biologically realistic models, in order to keep r defined in a way that is plausibly connected to relatedness, b and c become functions of r itself. These sorts of issues are dealt with by Frank (2013); Birch and Okasha (2015) among others.

3.6 Conclusion

While there can be benefits to using inclusive fitness, this does not mean that it is always beneficial to do so. Whether inclusive fitness or Hamilton's Rule should be used depends on the model or the population one is studying. Many of the issues involved in deciding whether to use these methods were not addressed here. This chapter has discussed the use of inclusive fitness in a special type of evolutionary model, in which pairwise interactions, additive fitness effects, and a finite number of types were assumed. In doing so, it focused the discussion on issues surrounding the different methodologies favored by the critics and proponents of inclusive fitness theory, in absence of conceptual and mathematical complexities that can arise in more complicated scenarios. Looking at this simple case helped to illuminate several features of the mathematical framework of inclusive fitness and the debate surrounding it.

While there may be difficulties with partitioning fitness effects into the form demanded by inclusive fitness when interactions become more complicated, we have seen that the specific causal partition used in inclusive fitness does not prevent one from building dynamically sufficient models nor does it require weak selection. Criticisms of inclusive fitness claiming that it requires these stringent assumptions are best thought of as criticisms of the types of quantitative methods generally used by inclusive fitness theorists. One can use inclusive fitness calculations in the sort of population genetic or evolutionary game theoretic models favored by these critics. In these models much of the advantage of using inclusive fitness, such as providing terms that can be easy to estimate empirically, disappears, but its power as an intuitive explanation of the evolution of social traits remains.

The next chapter will provide a more extensive example of the usefulness of inclusive fitness in conceptualizing the evolutionary process, looking at the the evolution of altruism in human populations. We have limited evidence on the conditions under which human populations evolved altruistic tendencies, but we do have important information about how relatedness

changed over time, making inclusive fitness calculations a useful tool.

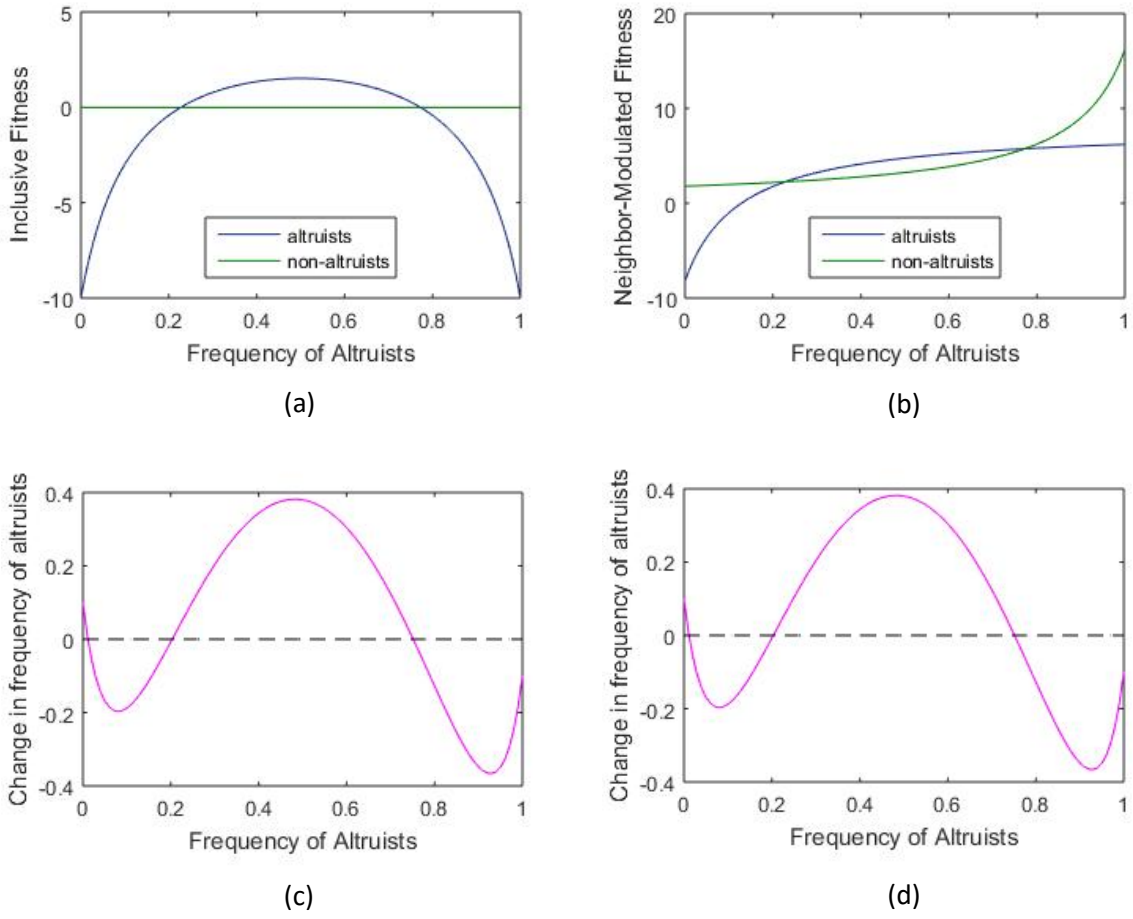


Figure 3.4: A comparison of inclusive fitness and neighbor-modulated fitness, for $\mu = 0.1$, $b = 18$ and $c = 10$. Comparing the calculation of inclusive fitness shown in (a) and neighbor-modulated fitness in (b) shows how the calculations of the two types of fitnesses differ. Comparing the change in the frequency of altruists found using inclusive fitness in (c) and neighbor-modulated fitness in (d) shows that the evolutionary trajectories are the same regardless of which calculation of fitness is used.

Chapter 4

Inclusive Fitness and the Evolution of Altruism in Human Groups

> People do more for their fellows than return favors and punish cheaters. They often perform generous acts without the slightest hope for payback ranging from leaving a tip in a restaurant they will never visit again to throwing themselves on a live grenade to save their brothers in arms (Pinker, 2003, p. 259).

4.1 Introduction

As Pinker expresses, altruistic behavior in human beings is widespread. Examples of altruistic behavior are everywhere we look. In experimental settings, people are given a sum of money, say $10, and can give any amount of this to an anonymous other participant in the experiment, any amount ranging from $0 to $10 without and fear of retribution for giving nothing. Yet, the majority of participants in these experiments will give at least something to the other person, many of them giving $5 out the $10 (Henrich and Henrich, 2007, p. 166). More systematically, people will conserve fuel, recycle, and cut down on consumption to protect the environment. Their contribution is individually costly in terms of time spent

and material sacrifice but brings no real tangible benefit to them personally as one person alone cannot affect climate change.

There have been many attempts to explain the emergence of these sorts of altruistic behaviors in human populations. This chapter will discuss one particularly influential line of reasoning in explanations of human altruism. The argument is that kin selection, or interactions between close genetic relatives, cannot explain human altruism. Instead, group selection can explain human altruism: in competition among groups, groups with more altruists will outcompete groups with fewer altruists because their members are doing better overall. I will argue that this explanation is incomplete in that it does not explain how these groups with various level of altruism arose. Using an inclusive fitness model, I show how kin selection can fill in the gaps of this argument.

That is, this model will describe, using inclusive fitness, how biological evolution can interact with forces of cultural evolution to produce groups with various levels of altruism. So, first, in section 4.2, I will discuss gene-culture co-evolution, which is way of talking about how the biological evolution of humans interacts with the evolution of our culture. In section 4.3, I will then explain why people reject kin selection as an explanation of altruism in humans and describe their alternative proposal, which is that group selection can explain human altruism. I will explain why this group selection explanation is incomplete, then, in section 4.4, I will provide a new model to show how kin selection can play in important role in the evolutionary history of altruism, although in a different way than has been previously proposed. As I will discuss in section 4.5, this model will show us how inclusive fitness neatly organizes around a parameter which changes systematically over time (relatedness) and provides and explanation in terms of that parameter.

4.2 Gene-Culture Co-Evolution

While humans evolved biologically like any other species, culture has also been incredibly important in explaining the success and the behaviors we see in modern human society. So, we have two types of evolution occurring at the same time: biological and cultural. Gene-culture co-evolution provides us a way to talk about how these two types of evolution occur and how they can interact with each other.

In biological, or genetic, evolution, traits are influenced by genes which are passed on from parents to offspring. Traits that are beneficial will generally increase in frequency because people with those traits will tend to survive and/or reproduce more often. In cultural evolution, traits are influenced by social learning and can be passed on in a variety of ways. So for example, they can be passed on from parents to their children; from any member of an older generation to a younger generation, or between members of the same generation.

Cultural evolution occurs when there is variation in behavior in a group of people and some of these behaviors are imitated more often than others. This could be because members of the group benefit from one behavior, others observe the behavior and the resulting benefits, then imitate the behavior. But, in social learning, people do not just always to pay attention to how beneficial a behavior is. It could be the case that one behavior is more prevalent than other, and people display some sort of *conformist bias*, meaning they adopt a behavior at least partially based on how common it is in the population. Similarly, individuals may exhibit some *prestige bias*, preferentially imitating group members who are most successful (whether or not the trait they are imitating is the source of that individual's success).

Of course these various forms of social learning are likely to some degree influenced by genetics that affect psychological predispositions. For example, if groups are in a relatively stable environment where, for example, a certain hunting technique remains successful generation after generation, then genes encoding for conformist bias are likely to evolve because it is

much faster and less costly to simply do what everyone else is doing than to figure out the optimal hunting technique for yourself (Boyd and Richerson, 1985; Henrich and Henrich, 2007). Just as genetics affect how culture evolves, so culture can also affect how genetics evolve. We will come back to a discussion of conformist bias in section 4.4 as it will be particularly important for the model presented here.

To demonstrate how gene-culture co-evolution is important here is one paradigmatic example of how it can occur. Before about 10,000 years ago, adult humans were not generally capable of processing the sugar in milk, called lactose. They could process the lactose as infants but after they were weened off their mother's milk they stopped producing the necessary enzyme to metabolize lactose. After animals like cows started being domesticated, there was a steady source of milk readily available. So, people who could process any of the nutrients in milk were favored by biological evolution. They had more nutrients in their diet so were able to survive and produce more offspring. This example is well supported by historical and genetic evidence. For instance, there is high frequency of people who can easily digest lactose in areas where dairying has been common for a long time and a low frequency in places where it has not historically been common (Feldman and Cavalli-Sforza, 1989; Aoki, 2001). This simple example shows how something that arose via cultural evolution affected our biological evolution, which is just one of the ways cultural and biological evolution can interact.

4.3 Common Explanation of Human Altruism

For traits like altruism, the story is a little less clear-cut than for things like lactose tolerance, and we will discuss some of the details why. As mentioned earlier, this is the common conclusion in the gene-culture co-evolution literature: kin selection cannot explain human altruism, but group selection can (Bowles and Gintis, 2002; Boyd et al., 2003; Bowles and

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

Table 4.1: Prisoners' dilemma

Gintis, 2011, among others).[1] As we have just seen however, the forces of cultural evolution can interact with the forces of genetic evolution. We will see in section 4.4 how kin selection as a force of biological evolution can actually interact with forces of cultural selection to explain the high levels of altruism we see in humans. First, though, we will go through this common line of reasoning. I will first explain why kinship has been largely dismissed in the gene-culture co-evolution literature. I will then explain how the group selection argument is supposed to explain the high levels of altruism in human populations. I will then argue that it is an incomplete explanation of human altruism and before showing how incorporating kin selection allows us to have a complete picture of how we might have evolved the high levels of altruism we see today.

4.3.1 Kin Selection

Kin selection describes the evolution of social behavior due to benefits falling on genetic relatives. As an example, consider the evolution of altruistic behaviors already discussed in chapters 1 and 2, and reproduced in table 4.1 here. When there is kin selection the organism pays the cost c in order to bestow a benefit b on a genetic relative who is likely to also have altruistic genes.

Although kin selection is a common explanation of altruism in non-human animals, it is often dismissed fairly quickly when talking about altruism in humans. So for example, Bowles and Gintis (2011) say that "... because one of the distinctive aspects of human cooperation

[1] These models also often assume there is some kind of punishment within groups for failure to exhibit a cooperative or altruistic trait, but this will not be relevant to the discussion here.

is that is extends far beyond the immediate family, we treat kin-based altruism only in passing" (p. 49). These authors argue that kin selection can only explain altruism toward immediate family members, and since what we are interested in explaining is widespread altruism toward non-kin, kin selection cannot be explanatorily helpful.

However, Bowles and Gintis (2011) are responding to a particular type of argument for the importance of kin selection, which they (and others) refer to as the 'big mistake hypothesis.' This is the argument that altruism evolved in humans at a time when we lived in small kin groups when, biologically, altruism was favored. In modern times, we are still altruistic because we have retained these genes for altruistic behavior even though they are no longer favored by evolution. This has become popularly known as the big mistake hypothesis because it implies that all our altruistic actions toward non-kin are just big mistakes - they are just misfirings of our desire to help kin in a world where we no longer primarily interact with kin.

This argument is often dismissed quickly for a couple of reasons. First, other primates can distinguish kin from non-kin when deciding whether to behave altruistically, so it seems unreasonable that humans would not be able to do so. Second, it is argued that kin selection is unlikely to be important for explaining altruism in human societies because human groups were too large at the time we think modern human society started to evolve (around the late Pleistocene). That is, if humans were in groups of 50 to 100 people, then they were not just interacting with close kin like parents and siblings. These large groups included many others individuals which a much lower relatedness, so being altruistic towards groups members in general would not be favored under biological evolution (Fehr and Henrich, 2003; Bowles and Gintis, 2011).

4.3.2 Group Selection

Instead, group selection is proposed as the explanation of altruism in modern humans. Although our understanding of group selection has deepened in recent years, the idea has been around since Darwin:

> Selfish and contentious people will not cohere, and without coherence, nothing can be effected. A tribe possessing... a greater number of courageous, sympathetic and faithful members, who were ready to warn each other, to aid and defend each other... would spread and be victorious over other tribes... Thus the social and moral qualities would tend to slowly advance and be diffused throughout the world. (Darwin, 1871, p. 156).

The basic idea is that groups whose members are altruistic will tend to out-compete other groups because they will more often survive things like environmental crises or attack by a predator.

More specifically, the current arguments for group selection explain that even though altruism might decrease within a group, groups with more altruists do better and reproduce more often, so overall altruists increase in frequency. Figure 4.1 provides an illustration of this concept. Altruists are represented by open circles, and non-altruists by filled-in circles. The group of all altruists survives to reproduce two new groups of altruists, the group with no altruists dies out, and the group with a majority altruists survives to reproduce one group. Although the frequency of altruists decreases within this last group, the frequency of altruists increases overall.

This argument assumes that we can think of groups as distinct entities which do things like survive and reproduce. For human groups, there are good reasons to think we can do this. First, we can think of these groups as distinct entities, coherent wholes, that survive over time. A group of altruists stays a group of altruists over time because importantly, within groups, there are norms for how to behave. In addition, there are forces of cultural evolution

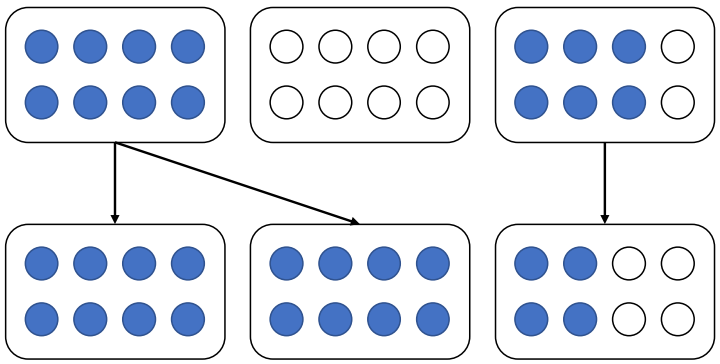


Figure 4.1: Group selection.

which keep groups coherent wholes over time. For instance, insider bias, the tendency to interact with people within one's own group, keeps groups as separate entities. Additionally, conformist bias reduces within group differences. Even if there is a fair bit of migration between groups (people leaving one group and joining another), a group of altruists will remain a group of altruists because new members generally conform to the norms of the group. So these groups can be thought of as discrete entities, and they can also reproduce: when individuals within a group thrive and increase in number, a group will eventually gets too large and splits into smaller groups. In this case, the altruist group grows in size to a point where eventually it splits into two groups (Bowles and Gintis, 2011, p. 50-2).

However, if we take this group selection argument seriously, and we think it provides a better explanation of altruism in humans than the kin selection argument, there is still work to be done in explaining how the groups with varying levels of altruism arose in the first place.

4.4 Gene-Culture Co-Evolution and Kin Selection

Now, I will provide a description of how it is possible to have kin selection be important in our evolutionary history without relying on the big mistake hypothesis, which, recall, says

that our altruistic actions are misfirings of our desire to help kin in a world where we no longer primarily interact with kin. Using inclusive fitness calculations, I will show how kin selection, as a force of biological evolution, can interact with forces cultural evolution in a model of gene-culture co-evolution.

4.4.1 Facts About Human Evolutionary History

We will start with three observations about human evolutionary history, then model a situation in which we take them to be important for how evolution occurs. That is, we will start with some facts about our evolutionary history and then provide an account of the evolution of altruism based on these facts.

Here are three facts about human evolutionary history:

1. Before the late Pleistocene humans lived in smaller kin groups, which eventually grew in size to become the larger groups talked about in the group selection argument (Tomasello et al., 2012).

2. The benefit of conformist bias increases as group size increases (**?**). The basic idea is that if 5 people exhibit a certain behavior that is not as reliable of an indication of its quality as it would be if a group of 50 people exhibit the behavior, because its more likely that the 5 people just arrived at the behavior by chance. However, the level of conformist bias might be different in different groups at any given time for a variety of reasons. Importantly, how fast groups adapt might vary; the behaviors influenced by conformist bias may be more or less crucial to their survival, or selective pressures on the groups might be more or less strong because they live in harsher versus more mild climates (Henrich and Henrich, 2007, 22).

3. As we have already discussed, relatedness decreases as group size increases because

there are more people in the group than just immediate family members.

4.4.2 Model with Inclusive Fitness

Before I get into more of the details of the model, here is a basic description of how the evolution of genes and strategies occurs. Here, I use the terms *phenotype* or *strategy* to refer to whether or not a person chooses to be an altruist, regardless of their genotype. In order to determine how populations would evolve, I ran computer simulations. In these simulations, people start out with a strategy depending on their genetics. That is, someone with altruist genes will very likely be an altruist. People then interact within their group multiple times before reproducing. While they are interacting, they see how well their strategy is doing and might change it if they can see another strategy is doing better or (when there is conformist bias) if another strategy is more prevalent in the population. This is a process of cultural evolution. During these interactions they also accumulate material payoffs. The greater the material payoff a person accumulates, the more offspring they have, and, thus, their genes (but not necessarily their strategy) will increase in the next generation. This is the process of biological evolution. We then start the process over with the next generation in which we have some cultural evolution then one instance of biological reproduction.

I will explain the parts of the model in the following order: starting conditions, cultural evolution, biological evolution, then how group size increases over time and how that effects certain things in the model.

Initially, group size is small to represent the situation where people interact within a small kin group. Within this group, there is a distribution of genotypes describing how many people have altruistic genes. A person with altruist genes is very likely to be penotypically, or culturally, altruist (to behave altruistically). As such, we must provide a measure of the heritability of the trait, a measure of how much the trait is influenced by genetics, h^2. We

have four phenogenotypes to track in our evolutionary model, representing the combination of genotype and phenotype for each individual.

Within each biological generation, individuals interact a number of times and undergo cultural evolution. For each interaction, each individual will perform an altruist action or not, depending on their phenotype. Then, after each interaction, the distribution of phenotypes (but not genotypes) evolves. This evolution occurs according to the discrete-time replicator dynamics, which captures the fact that if the value of altruism is greater than the value of non-altruism at time t, then during the next time period, $t + 1$, altruism will increase in frequency:

$$x_a(t+1) = \frac{x_a(t) \cdot v_a(x(t))}{\bar{v}(t)} \tag{4.1}$$

where $v_a(t)$ represents the value of altruism at time t, and $\bar{v}(t)$ represents the average value of all traits in the population (which is calculated by taking a weighted average of the values of altruism and non-altruism).

The value of altruism depends on $x_a(t)$, the frequency of altruists at time t, and the current level of conformist bias. So, if conformist bias is high, the value of altruism will depend mostly on how frequent it is in the population, and if conformist bias is low, then the value of altruism will depend mostly on the material payoffs. If $\mathcal{C}$ represents the level of conformist bias, and $f_a(x(t))$ the material payoff gained in that interaction period based on the current distribution of strategies in the population, then the value of altruism is:

$$v_a(x(t)) = \mathcal{C}x_a(t) + (1 - \mathcal{C})f_a(x(t)) \tag{4.2}$$

Remember that the level of conformist bias evolves over time, which I will return to shortly.

We have thus far described how cultural evolution occurs within one biological generation. The model tracks the evolution of the population over 1000 of these biological generations. Throughout each generation, genes (or the humans which possess them) accumulate material

payoffs based on the cultural interactions. Then, after the period of cultural evolution occurs, one instance of biological reproduction occurs. This genetic evolution occurs according to an equation very similar to the equation used for cultural evolution, but instead uses the fitness (the expected number of offspring, based on material payoffs) rather than the perceived value of a trait:

$$x_a(t+1) = \frac{x_a(t) \cdot f_a(x(t))}{\bar{f}(t)} \tag{4.3}$$

If altruists have a greater fitness, the genes for altruism will increase in frequency.

We can calculate fitness using inclusive fitness. In calculating inclusive fitness, we look at the benefits altruists confer on their relatives (and how related they are) and the cost altruists have to pay:

$$rb - c$$

Non-altruists do not pay any cost or confer any benefits so we set their fitness equal to 0. In order to more accurately represent human populations, we also include some overlap between generations, o. That is, not all the adults die when children are born.

It is now time to incorporate the important facts about human evolutionary history discussed in the previous section. As noted, group size increases over time (fact 1). This is incorporated into the model as having group size, N, increase by 1 every 10 biological generations.

As group size increases, the benefit of conformist bias increases (fact 2). As noted, how quickly the actual level of conformist bias will increase in the group depends on the strength of the selective pressures. This is incorporated into the model in the following way. For each group size, we have some optimal level of conformist bias:

$$\mathcal{C}_{optimal} = \frac{N}{N+100} \tag{4.4}$$

This equation for the optimal level of conformist bias is chosen to represent a general trend; the specific form it takes is unimportant. Figuring out the actual equation for what the

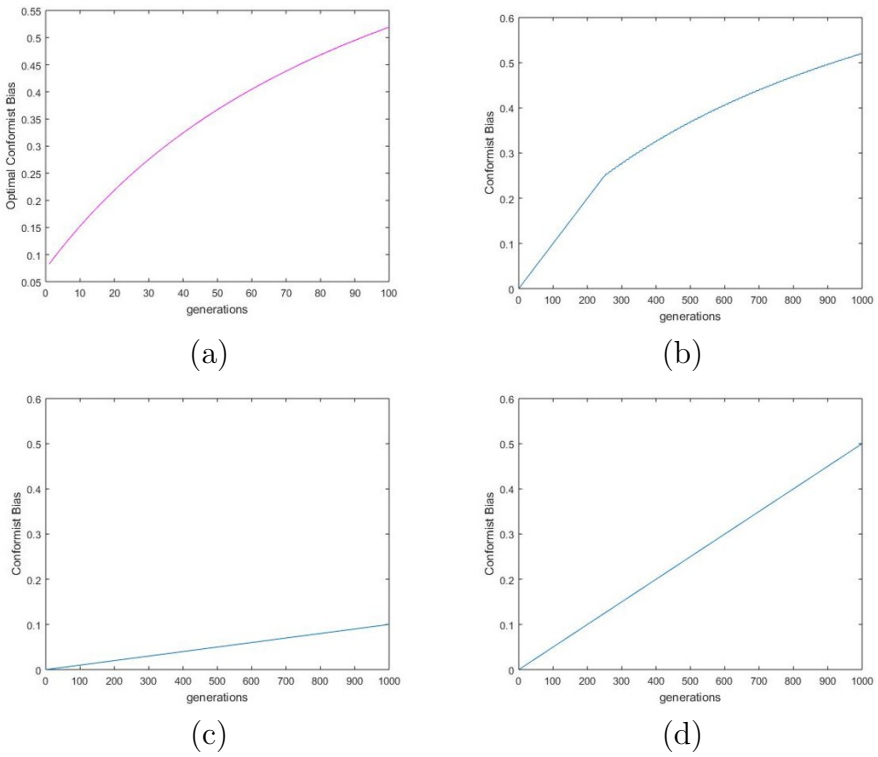


Figure 4.2: Optimal and evolved levels of conformist bias as group size increases.

optimal conformist bias would be would require knowing a lot of particular details about the population, but here we are talking in very general terms about the evolution of traits influenced by conformist bias. The equation just needs to capture a couple of features. First, it needs to start out small to represent the fact we described earlier that conformist bias is not that beneficial in smaller groups. Second, because of how it is incorporated into the equations describing cultural evolution, the level of conformist bias needs to stay between 0 and 1. These features are captured by equation 4.4, as shown in figure 4.2(a).

So, over the generations, the optimal level of conformist bias increases and slowly levels off. However, the actual level of conformist bias depends on the selective pressures acting on the population, and, as we will see, the actual level of conformist bias is key to how much altruism can be sustained. We will talk about three cases: one where actual conformist bias evolves so quickly that the actual level is nearly optimal throughout the generations, figure 4.2(b), another whether the conformist bias evolves very slowly, figure 4.2(c), and an

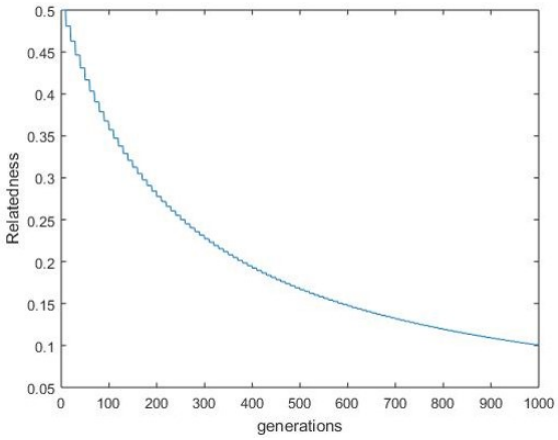


Figure 4.3: Relatedness as group size increases.

intermediate case where the evolution is not too slow but conformist bias does not evolve quickly enough to keep up with what the optimal level is, figure 4.2(d).

As group size increases, conformist bias increases, but also relatedness decreases (fact 3). We capture this fact by having relatedness change according to the following equation:

$$R = .5 - \frac{N-9}{2(N-9)+50} \tag{4.5}$$

Like the equation for $\mathcal{C}_{optimal}$, knowing the exact equation for how relatedness evolves requires knowing specifics about reproduction and group structure, so this equation only captures some general features of the change in relatedness as group size increases: it starts out around $\frac{1}{2}$ as most members of a small kin group are your parents or siblings, and it decreases over time to some level above 0, as shown in figure 4.3.

Note that the way conformist bias and relatedness change are built in as assumptions of the model in such a way they match the facts of the evolution of human groups described earlier. Note also that the model provided here incorporates kinship in a way that differs from the big mistake hypothesis in that the altruism we end up with is always meant to be directed to members of the group, and is not just a misfiring of an attempt to help genetic relatives.

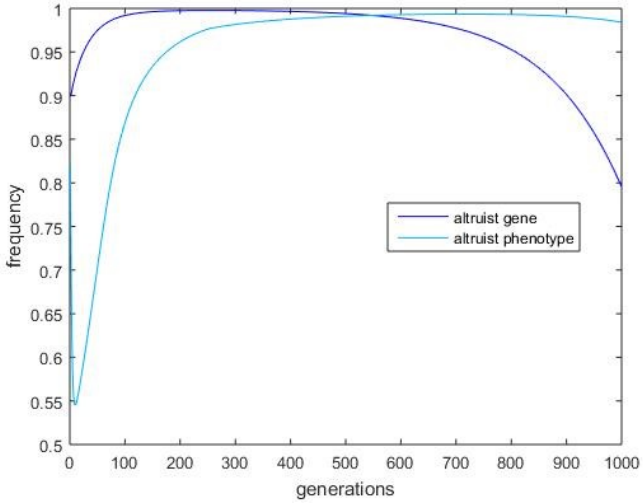
4.4.3 Groups With Various Levels of Altruism

As mentioned, how much altruism can be sustained depends on the level of conformist bias in the population. Here, I present the results from the model described in the previous section in terms of the speed of evolution of conformist bias, corresponding to the three cases pictured in figure 4.2(b)-(d).

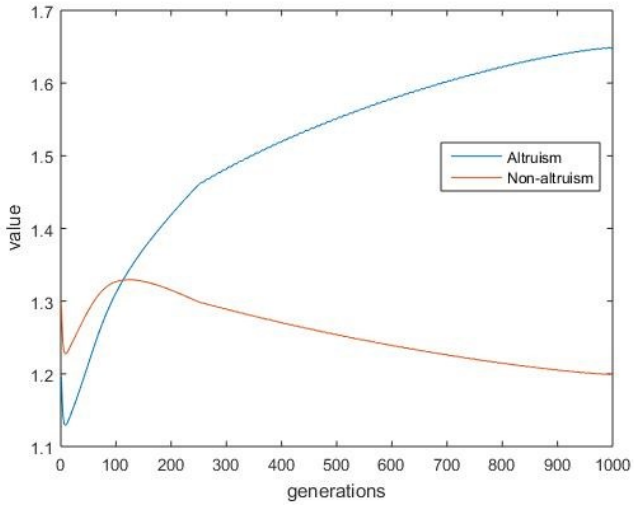
For the results discussed here, the starting group size is 9 which captures that interactions are within a small kin group. Additionally, $h^2 = .8$, which represents a trait that is highly heritable, and the overlap between generations is $o = .45$, which means that in any generation children make up a little over half the population. Similar results can be obtained for a variety of parameter values.

Figure 4.4 shows the evolution of the population when conformist bias increases quickly. We track how the frequency of altruist genes and altruist phenotypes (or, cultural altruism) change over the generations in figure 4.4(a). Initially, the altruistic gene is favored biologically because relatedness is high, but at around 250 generations, relatedness decreases to a point where these genes are no longer favored and they begin to decrease in frequency. For the altruistic phenotype, there is an initial decrease since conformist bias is low: people are born altruistic but fairly quickly learn that altruists have lower material payoffs than non-altruists.

However, as conformist bias increases, the *value* of altruism increases, as shown in figure 4.4(b). Remember that the perceived values of these behaviors is what affects cultural evolution. Initially, when conformist bias is low, the value of altruism is lower than the value of non-altruism. This is because the value of these strategies mostly depends on the material payoffs (and the material payoffs of altruists are lower than the material payoffs of non-altruists). However, as conformist bias increases, people care more about how prevalent a trait is, and the value of altruism increases to the point where it is higher than the value of non-altruism. People are still born altruistic (although it is decreasing, relatedness is still



(a)



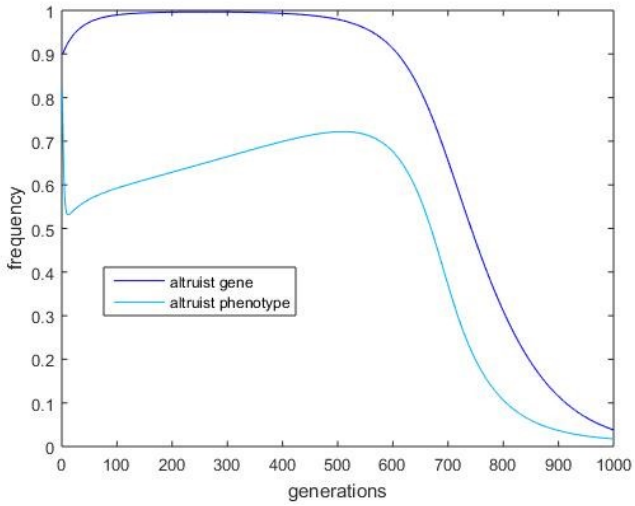
(b)

Figure 4.4: Evolution when conformist bias increases quickly.

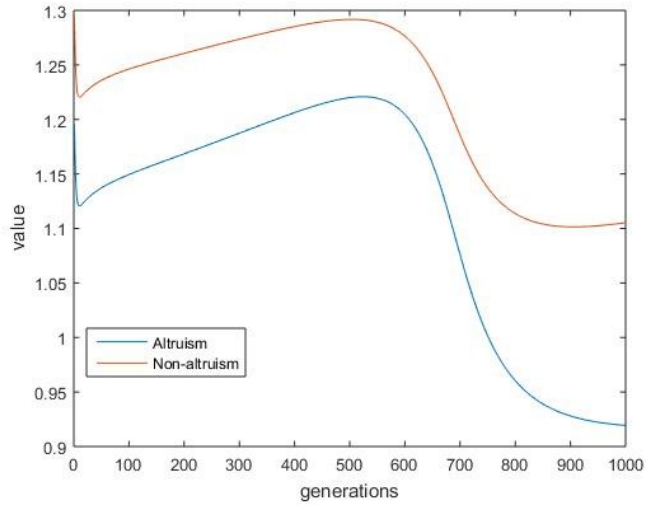
high enough that the the inclusive fitness of altruists is greater than the inclusive fitness of non-altruists) so there is a high frequency of altruists. As conformist bias increases, this high frequency matters more for the calculation of value. Cultural altruism increases to a point where it can be maintained even when the genes for altruism begin to disappear (when relatedness decreases to a point where altruism is no longer biologically favored). So, when conformist bias increases quickly, cultural altruism can be maintained for long periods of time.

What happens when conformist bias increases slowly? In this case, again, the altruistic gene is initially favored biologically because relatedness is high, but at around 250 generations, relatedness decreases to a point where these genes are no longer favored and begin to decrease in frequency. However, in this case, since conformist bias increases so slowly, the altruistic phenotype is not sustained culturally, as shown in figure 4.5(a). So while we start out with a high frequency of altruism, it eventually collapses as the genes for altruism disappear. This is because, as figure 4.5(b) shows, the value of altruism never increases to the point where its greater than the value of non altruism. The value of altruism increases slightly as conformist bias increases, but once the altruistic genes disappear, the frequency of altruists drops and so does the value of altruism.

Finally, we can consider the case where conformist bias increases at an intermediate speed. In this case, cultural altruism can be sustained for a short period of time. Again, the altruistic gene is initially frequent, but drops out of the population as relatedness decreases. The altruistic phenotype first dips then increases, just as in the first situation, but it does not increase to such high levels so it eventually drops off toward the end of the 1000 generations, as shown in figure 4.6(a) In this case, the value of altruism increases to where it is higher than the value of non-altruism but it does not increase quite as quickly as in the first case. This is show in figure 4.6(b). This means that altruism never reaches as high of frequency as in the first case, so it cannot be sustained for as long of a period of time.

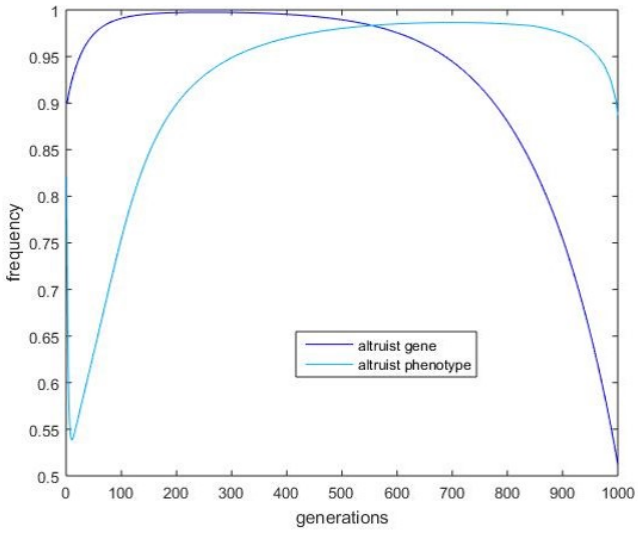


(a)

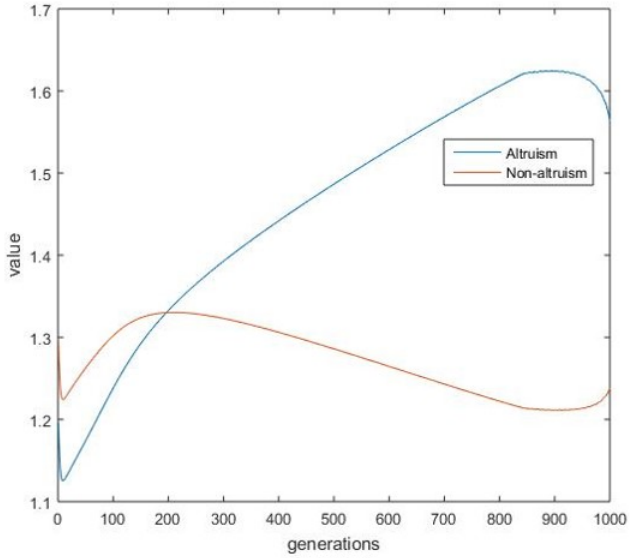


(b)

Figure 4.5: Evolution when conformist bias increases slowly.



(a)



(b)

Figure 4.6: Evolution when conformist bias increases at an intermediate speed.

So, when we vary how conformist bias evolves, this force of cultural evolution can interact with biological evolution in order to generate groups with different levels of altruism in order to start the group selection argument.

4.5 Conclusion

We started out with this puzzle: why do humans behave altruistically, and why do they direct their altruistic actions towards non-kin so frequently? Since in humans, behavior is due to both our biological makeup and our culture, we looked to gene-culture co-evolution to help us explain the phenomenon. We saw why kin selection, where the benefits of altruistic actions fall on kin, was dismissed as an explanation of the widespread existence of altruism in humans. We then saw that group selection, where altruistic groups out-compete non-altruistic groups, provided an incomplete explanation.

I then provided a model showing that kin selection could play an important part in explaining human altruism in generating the groups with various levels of altruism that group selection can act on. The model provided here allows us to have a complete picture of the evolution of human altruism, and reinstates kin selection as an important evolutionary force. That is, it resolves some of the issues with the leading explanation of the evolution of altruism in human societies, and made important progress toward solving the puzzle of altruism that we started with.

In this model, we used information about how relatedness changed systematically over time, which made inclusive fitness an incredibly useful framework for conceptualizing the evolutionary process. The inclusive fitness calculations organized around a parameter which changed over time (relatedness) and allowed us to provide explanations in terms of that parameter.

Chapter 5

Conclusion

Inclusive fitness has recently been under intense debate, with critics claiming it is unusable for studying evolution and proponents claiming it is an indispensable part of evolutionary theory. We have taken a closer look at this debate to determine the explanatory value of inclusive fitness for evolutionary theory.

Chapter 2 examined the argument that inclusive fitness is indispensable for evolutionary theory and showed that this argument rests on a confusion between correlation and causation in a non-obvious way. Chapter 3 then defended the inclusive fitness framework against several criticisms, and argued that they are not really criticisms of inclusive fitness but are actually about the orthogonal issue of using abstract versus idealized models. Then, chapter 4 provided a case study to show how inclusive fitness can be an extremely useful way of conceptualizing the evolutionary process, using the evolution of altruistic behaviors in humans as the demonstrative example.

So, first I argued against inclusive fitness proponents: I claimed that inclusive fitness is not necessary for evolutionary explanations. Then, I argued against inclusive fitness critics: I claimed that inclusive fitness is indeed adequate for providing evolutionary explanations and

can in fact be a very useful tool when relatedness is a key evolutionary factor. In all, we have seen that although inclusive fitness is not necessary for evolutionary explanations, it can nonetheless provide a useful way of conceptualizing the evolutionary process.

Bibliography

Abbot, P., J. Abe, J. Alcock, S. Alizon, J. A. C. Alpedrinha, and et al. (2011). Inclusive fitness theory and eusociality. *Nature 471*, E1–E4.

Allen, B. and M. A. Nowak (2015). Games among relatives revisited. *Journal of theoretical biology 378*, 103–116.

Allen, B., M. A. Nowak, and E. O. Wilson (2013). Limitations of inclusive fitness. *Proceedings of the National Academy of Sciences 110*, 20135–20139.

Aoki, K. (2001). Theoretical and empirical aspects of gene–culture coevolution. *Theoretical population biology 59*(4), 253–261.

Bergstrom, T. (1995). On the evolution of altruistic ethical rules for siblings. *The American Economic Review*, 58–81.

Bergstrom, T. (2000). Evolution of behavior in family games. *Department of Economics, UCSB*.

Birch, J. (2014a). Gene mobility and the concept of relatedness. *Biology & Philosophy 29*(4), 445–476.

Birch, J. (2014b). Hamilton's rule and its discontents. *British Journal for the Philosophy of Science 65*(2), 381–411.

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

Birch, J. and S. Okasha (2015). Kin selection and its critics. *BioScience 65*(6), 22–32.

Bowles, S. and H. Gintis (2002). Behavioural science: homo reciprocans. *Nature 415*(6868), 125–128.

Bowles, S. and H. Gintis (2011). *A cooperative species: Human reciprocity and its evolution*. Princeton University Press.

Boyd, R., H. Gintis, S. Bowles, and P. J. Richerson (2003). The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences 100*(6), 3531–3535.

Boyd, R. and P. J. Richerson (1985). *Culture and the evolutionary process*. University of Chicago press.

Cavalli-Sforza, L. L. and M. W. Feldman (1978). Darwinian selection and altruism. *Theoretical population biology 14*(2), 268–280.

Chakraborty, A. and R. Harbaugh (2007). Comparative cheap talk. *Journal of Economic Theory 132*(1), 70–94.

Darwin, C. (1871). *The Descent of Man*. Number v. 2. D. Appleton and Company.

Darwin, C. (1959). *On the origin of species by means of natural selection. 1859*. J. Murray, London.

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

Eliashberg, J. and R. L. Winkler (1981). Risk sharing and group decision making. *Management Science 27*(11), 1221–1235.

Fehr, E. and J. Henrich (2003). Is strong reciprocity a maladaptation? on the evolutionary foundations of human altruism. In P. Hammerstein (Ed.), *Mathematical Evolutionary Theory*, pp. 55–82. MIT Press.

Feldman, M. W. and L. L. Cavalli-Sforza (1989). On the theory of evolution under genetic and cultural transmission with application to the lactose absorption problem. In M. W. Feldman (Ed.), *Mathematical Evolutionary Theory*, pp. 145–173. Princeton University Press.

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

Frank, S. (2013). Natural selection. vii. history and interpretation of kin selection theory. *Journal of Evolutionary Biology 26*, 1151–1184.

Gardner, A., S. West, and G. Wild (2011). The genetical theory of kin selection. *Journal of Evolutionary Biology 24*(5), 1020–1043.

Godfrey-Smith, P. (2009). *Abstractions, idealizations, and evolutionary biology*. Springer.

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

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. (1985). A geometric view of relatedness. *Oxford surveys in evolutionary biology 2*(2).

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

Grafen, A. (2007a). Detecting kin selection at work using inclusive fitness. *Proceedings of the Royal Society of London B: Biological Sciences 274*(1610), 713–719.

Grafen, A. (2007b). The formal darwinism project: a mid-term report. *Journal of evolutionary Biology 20*, 1243–1254.

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. (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.

Hamilton, W. D. (1975). Innate social aptitudes of man: an approach from evolutionary genetics. *Biosocial anthropology 133*, 155.

Hammerstein, P. and R. Selten (1994). Game theory and evolutionary biology. In S. Hart (Ed.), *Handbook of Game Theory with Economic Applications*, Volume 2. Amsterdam: Elsevier Science.

Henrich, N. and J. P. Henrich (2007). *Why humans cooperate: A cultural and evolutionary explanation*. Oxford University Press.

Karlin, S. and C. Matessi (1983). The eleventh ra fisher memorial lecture: Kin selection and altruism. *Proceedings of the Royal Society of London B: Biological Sciences 219*(1216), 327–353.

Lehmann, L. and F. Rousset (2014). The genetical theory of social behaviour. *Philosophical Transactions of the Royal Society B 369*(1642).

Maciejewski, W., F. Fu, and C. Hauert (2014). Evolutionary game dynamics in populations with heterogenous structures. *PLoS Comput Biol 10*(4), e1003567.

Marshall, J. A. (2011). Queller's rule ok: Comment on van veelen 'when inclusive fitness if right and when it can be wrong'. *Journal of Theoretical Biology 270*, 185–188.

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

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

Nowak, M. A., C. E. Tarnita, and E. O. Wilson (2011). Nowak el al. reply. *Nature 471*, E9–E10.

Nozick, R. (1969). Newcombs problem and two principles of choice. In *Essays in honor of Carl G. Hempel*, pp. 114–146. Springer.

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.

Orlove, M. and C. L. Wood (1978). Coefficients of relationship and coefficients of relatedness in kin selection: a covariance form for the rho formula. *Journal of Theoretical Biology 73*(4), 679–686.

Page, K. M. and M. A. Nowak (2002). Unifying evolutionary dynamics. *Journal of Theoretical Biology 270*, 93–98.

Pinker, S. (2003). *The blank slate: The modern denial of human nature*. Penguin.

Queller, D. C. (1992). Quantitative genetics, inclusive fitness, and group selection. *The American Naturalist 139*(3), 540–58.

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.

Rousset, F. (2002). Inbreeding and relatedness coefficients: what do they measure? *Heredity 88*, 371–380.

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

Skyrms, B. (1996). *Evolution of the Social Contract*. Cambridge University Press.

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.

Taylor, P. D. Wild, G. and A. Gardner (2007). Direct fitness or inclusive fitness: how shall we model kin selection? *Journal of Evolutionary Biology 20*, 301–309.

Taylor, P. and S. Frank (1996). How to make a kin selection model. *Journal of Theoretical Biology 180*, 26–37.

Taylor, P. and W. Maciejewski (2012). An inclusive fitness analysis of synergistic interactions in structured populations. *Proceedings of the Royal Society London B 279*(1747).

Tomasello, M., A. P. Melis, C. Tennie, E. Wyman, and E. Herrmann (2012). Two key steps in the evolution of human cooperation. *Current Anthropology 53*(6), 673–692.

Traulsen, A. (2010). Mathematics of kin- and group-selection: formally equivalent? *Evolution 64*(2), 316–323.

van Veelen, M. (2005). On the use of the price equation. *Journal of Theoretical Biology 237*, 412–426.

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*, 589–600.

van Veelen, M. (2011). The replicator dynamics with n players and population structure. *Journal of theoretical biology 276*(1), 78–85.

van Veelen, M., J. García, M. W. Sabelis, and M. Egas (2012). Group selection and inclusive fitness are not equivalent; the price equation vs. models and statistics. *Journal of theoretical biology 299*, 64–80.

Weirich, P. (2016). Causal decision theory. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2016 ed.).

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.

Wild, G. and A. Traulsen (2007). The different limits of weak selection and the evolutionary dynamics of finite populations. *Journal of Theoretical Biology 247*(2), 382–390.

Wilson, E. O. (2012). *The Social Conquest of Earth*. W. W. Norton.

Appendix A

Appendix

A.1 Equivalence with Neighbor-Modulated Fitness

A.1.1 The Price equation describes the replicator dynamics

Following the definition provided in section 2.2.2, we can calculate the neighbor-modulated fitness of a pairwise interaction as follows:

$$f_i = s_{ii} + s_{-ii} \tag{A.1}$$

Keeping track of probabilities of receiving payoffs was necessary in section 3.4.1 in order to show the connection between neighbor-modulated fitness and inclusive fitness, but since we are only dealing with neighbor-modulated fitness we can use this less complicated expression. In these calculations, we will track the change in g, genetic value.

By definition, $\dot{E}(g) = \sum_i g_i \dot{x}_i + \sum_i \dot{g}_i x_i$. As mentioned, for simplicity we will assume there is no transmission bias and set $\sum_i \dot{g}_i x_i = 0$. Then, since the replicator dynamics provides us

an equation for $\dot{x}_i$, we can plug the replicator dynamics into the Price equation:

$$\begin{aligned}
\dot{E}(g) &= \sum_i g_i \dot{x}_i \\
&= \sum_i g_i x_i [s_{ii} - \frac{1}{n}\sum_j s_{jj} + s_{-ii} - \frac{1}{n}\sum_j s_{-jj}] \\
&= \sum_i g_i x_i s_{ii} - \sum_i g_i x_i \frac{1}{n}\sum_j s_{jj} + \sum_i g_i x_i s_{-ii} - \sum_i g_i x_i \frac{1}{n}\sum_j s_{-jj} \\
&= E(s_{ii}g) - E(s_{ii})E(g) + E(s_{-ii}g) - E(s_{-ii})E(g) \\
&= Cov(s_{ii}, g) + Cov(s_{-ii}, g)
\end{aligned} \tag{A.2}$$

This is the Price equation with fitness partitioned into two components, the effect the focal organism has on its own fitness and the effect the social partner has on the focal organism's fitness. Theorists often derive this from the original Price equation in order to use neighbor-modulated fitness calculations and introduce relatedness calculations (see Queller, 1992, for example).

Hamilton's Rule can easily be derived from this equation. Since the way an organism affects the fitness of itself and others is (to a certain degree) predicted by its phenotype, we can write both fitness terms as the following regressions:

$$s_{ii} = \alpha_{s_{ii}p} + \beta_{s_{ii}p} \cdot p + \epsilon_{s_{ii}p} \tag{A.3}$$

$$s_{-ii} = \alpha_{s_{-ii}p'} + \beta_{s_{-ii}p'} \cdot p' + \epsilon_{s_{-ii}p'} \tag{A.4}$$

Since the α's are the intercepts of the regression, they are constants and cannot covary with g. The ϵ's are error terms, which do not covary with g when payoffs are additive (Queller, 1992). So, plugging (A.3) and (A.4) into (A.2), we're left with:

$$\dot{E}(g) = \beta_{s_{ii}p} Cov(p, g) + \beta_{s_{-ii}p'} Cov(p', g) \tag{A.5}$$

When we can interpret $\beta_{s_{ii}p}$ as a cost c and $\beta_{s_{-ii}p'}$ as a benefit b this gives us:

$$\dot{E}(g) > 0 \text{ when } b \cdot \frac{Cov(p', g)}{Cov(p, g)} - c > 0 \tag{A.6}$$

where $\frac{Cov(p',g)}{Cov(p,g)}$ is the direct fitness version of relatedness.

A.1.2 The replicator dynamics describes the Price equation

When there are a finite number of types, g_i can be written as an indicator function:

$$g_j^{<i>} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

For Page and Nowak (2002), who were considering phenotypes rather than genotypes, assuming a finite number of types was a restriction. Here, in considering genotypes, it is a natural assumption to make.

We can then use this indicator function in the Price equation with two fitness components derived above, and simplify:

$$\begin{aligned} \dot{E}(g) &= Cov(s_{ii}, g^{<i>}) + Cov(s_{-ii}, g^{<i>}) \\ &= E(s_{ii}g^{<i>}) - E(s_{ii})E(g^{<i>}) + E(s_{-ii}g^{<i>}) - E(s_{-ii})E(g^{<i>}) \\ &= \sum_i g_j^{<i>} x_i s_{ii} - \sum_i g_j^{<i>} x_i \frac{1}{n} \sum_j s_{jj} + \sum_i g_j^{<i>} x_i s_{-ii} - \sum_i g_j^{<i>} x_i \frac{1}{n} \sum_j s_{-jj} \\ &= x_i s_{ii} - x_i \frac{1}{n} \sum_j s_{jj} + x_i s_{-ii} - x_i \frac{1}{n} \sum_j s_{-jj} \\ &= x_i[f_i(x) - \bar{f}] \end{aligned} \tag{A.7}$$

Since $g_j^{<i>} = 1$ when $i = j$ and 0 otherwise, $\sum_i g_j^{<i>} x_i = x_i$, and this simplifies to yield the replicator dynamics.

A.2 Equivalence with Inclusive Fitness

A.2.1 The Price equation describes the replicator dynamics

This is done in the same way as appendix A.1, except we take into account that the genetic value of the focal organism times its relatedness to its social partner is a measure of the social partner's genetic value:

$$\begin{aligned}\dot{E}(g) &= \sum_i g_i(x_i[s_{ii} - \frac{1}{n}\sum_j s_{jj} + x_i r_{i-i}[s_{i-i} - \frac{1}{n}\sum_j s_{j-j}]) \\ &= \sum_i g_i x_i s_{ii} - \sum_i g_i x_i \frac{1}{n}\sum_j s_{jj} + \sum_i g_i r_{i-i} x_i s_{-ii} - \sum_i g_i r_{i-i} x_i \frac{1}{n}\sum_j s_{-jj} \\ &= Cov(s_{ii}, g) + Cov(s_{i-i}, g')\end{aligned} \tag{A.8}$$

This is again a version of the Price equation where the fitness effect is split into two components. Here, though, fitness is split into the effect the focal organism has on its own fitness and the effect the focal organism has on its social partner's fitness.

In order to relate this to Hamilton's Rule, we can again notice that the fitness components are predicted by phenotype. Since in this case the focal organism causes the fitness effects, both for itself and its social partner, the phenotype of the focal organism predicts both fitness effects. So, we use the phenotype of the focal organism in both regressions:

$$s_{ii} = \alpha_{s_{ii}p} + \beta_{s_{ii}p} \cdot p + \epsilon_{s_{ii}p} \tag{A.9}$$

$$s_{-ii} = \alpha_{s_{i-i}p} + \beta_{s_{i-i}p} \cdot p + \epsilon_{s_{i-i}p} \tag{A.10}$$

We can then plug (A.9) and (A.10) into (A.8) and rearrange to obtain the inclusive fitness version of Hamilton's Rule:

$$\dot{E}(g) > 0 \text{ when } b \cdot \frac{Cov(p, g')}{Cov(p, g)} - c > 0 \tag{A.11}$$

A.2.2 The replicator dynamics describes the Price equation

We can again let g_i be an indicator function and write:

$$\begin{aligned}
\dot{E}(g) &= Cov(s_{ii}, g^{<i>}) + Cov(s_{-ii}, g'^{<i>}) \\
&= \sum_i g_j^{<i>} x_i s_{ii} - \sum_i g_j^{<i>} x_i \frac{1}{n} \sum_j s_{jj} + \sum_i g_j^{<i>} r_{i-i} x_i s_{i-i} - \sum_i g_j^{<i>} r_{i-i} x_i \frac{1}{n} \sum_j s_{j-j} \\
&= x_i s_{ii} - x_i \frac{1}{n} \sum_j s_{jj} + x_i r_{i-i} s_{i-i} - x_i r_{i-i} \frac{1}{n} \sum_j s_{j-j} \\
&= x_i [f_i(x) - \bar{f}]
\end{aligned} \tag{A.12}$$

Again, this simplifies to yield the replicator dynamics.