The Search for a Rigorous, Generally Applicable Formulation of Inclusive Fitness Theory

Report of a 6 month internship.
Submitted for the Masters course Bioinformatics and Biocomplexity.

Student	:	Jeroen An-Ying Saccheri
Student Number	:	0538824
Supervisor	:	Rutger Hermsen
Second Reviewer	:	Paulien Hogeweg
Date	:	April 24, 2023

Lay Summary

Inclusive fitness theory is an extremely influential idea in the field of evolutionary biology which changed the way we perceive evolutionary success. It examines the evolution of social behaviours such as altruism, whose evolution seemingly challenges Darwin's "survival of the fittest": by evolving altruism, an organism develops behaviours that negatively impact their own reproductive prospects to improve the reproduction of organisms around them, making themselves 'less fit'. A classic, extreme example of this is found in eusocial insects such as honey bees, where female workers do not reproduce at all despite possessing fully functional reproductive organs, and instead invest all their resources into improving the survival of offspring of the colony's queen.

In order to understand how seemingly selfless behaviours can evolve through natural selection, W.D. Hamilton developed inclusive fitness theory, which mathematically demonstrates that behaviours can spread if the genes that cause them are more likely to be passed on to future generations, regardless of the impact on the individual who carries those genes. Inclusive fitness theory explains that behaviours detrimental to an individual's reproductive success can still evolve if they increase 'inclusive fitness', a measure that accounts for the reproductive effects of a trait not only on the individual but also on other organisms with the same trait.

Hamilton used the framework of inclusive fitness theory to develop Hamilton's rule, which outlines the conditions necessary for an altruistic trait to evolve. It defines a relationship between the negative effect of a trait on an individual's reproduction (cost cc), the positive effect a trait gives to others

(benefit b), and how related interacting individuals are to each other (relatedness R):

$$Rb > c,$$

or that the cost of possessing a behaviour should be less than the benefit it provides, weighted by a measure of the degree of relatedness between actor and recipient. The attraction to this rule, as to many scientific equations, lies in its elegance, but its simple exterior belies hidden complexity. Hamilton's original definitions of R, b and c require a restrictive set of assumptions, but efforts to expand definitions of the parameters to more general alternatives result calculations for the three parameters depending on each other, causing difficulties interpreting what they truly represent.

Debate over Hamilton's rule has been an ongoing topic since its inception, but has increased in intensity over the past decade due to a series of critical, high-profile articles that claim the insights it provides are limited. In response, supporters of Hamilton's rule have pointed out the wealth of experimental evidence that demonstrates organisms do evolve to increase their inclusive fitness. However, due to the difficulties applying Hamilton's rule empirically, the methods used to define inclusive fitness and the parameters of Hamilton's rule in these experiments differ from case to case. As a result, critics remain vocally unconvinced, and there is still no current consensus reached for the usefulness of Hamilton's rule.

The goal of this project was to assess whether modern formulations of Hamilton's rule avoid the original's limiting assumptions, and to what extent generality comes at a cost to its descriptive power. We begin by reviewing the parameters of Hamilton's original formulation, the limiting assumptions these require, and how extensions to their definitions can avoid these assumptions. While explaining these extensions, we assess whether the adjustments necessary to improve parameter generality impact the rule's descriptive power. Afterwards, we consider why there are few empirical results that explicitly test Hamilton's rule, both by hypothesising how it could be applied to simple theoretical models, and giving an assessment of previous experimental results looking to apply Hamilton's rule, what previous studies have in common, and where inconsistencies between studies lie.

As a final goal, we aimed to find a new formulation that avoids all assumptions while retaining descriptive power, but this was unsuccessful. However, we provide a novel demonstration of how practical applications that aim to retain causal description while applying Hamilton's rule cannot avoid a set of limiting assumptions. In addition, we show that the conditions for these assumptions are typically not met by experimental applications of Hamilton's rule and that in general the rule's practical applications may be difficult to interpret. We then give some recommendations for directions of further study to be taken.

Abstract

Since its introduction in the 1960s, inclusive fitness theory and its results, especially Hamilton's rule, have been the topic of unresolved and ongoing debate. Over the past decade, arguments have increased in intensity following an outspoken critique by Nowak et al. in 2010. Opponents claim that general applications of the theory result in complex relationships between the quantities it describes, giving no real insights into how social behaviours may be positively selected; but specific applications are limited in scope, making it difficult to compare between cases.

This project looks to assess to what extent generality of Hamilton's rule comes at a detriment to its causal descriptive power. To assess this, I first describe the various forms that the parameters of cost, benefit and relatedness in Hamilton's rule may take, and assess their theoretical limitations in each case. I then consider how Hamilton's rule may be applied to a computational simulation model where all details of the system may be known, provide rigorous definitions of parameters that may be applied to an empirical study, and demonstrate what restrictions apply. Finally, I analyse previous results that utilised Hamilton's rule to interpret experimental data, examining to what extent predictions between experiments using inclusive fitness theory may be unified.

It is shown that while there is a wealth of empirical data indicating that altruism can evolve if it increases an individual's 'inclusive fitness', how this fitness is defined subtly differs from case to case, with no clear path to link these results to each other. Despite the value in these experimental results, further work must be taken in order to unify predictions made by Hamilton's rule before it can be considered a fully general method usable to interpret the conditions for the evolution of altruism. Nevertheless, it remains the most general method known for understanding the evolution of altruism thus far.

Contents

1 Introduction

If the survival of the fittest is an inalienable truth, why would an individual ever pay a cost to help others? This question of how altruism could evolve led to W.D. Hamilton's proposal of inclusive fitness theory, which studies the effect of a trait/gene on fitness while including the effects of social interactions [13, 14]. Since the turn of the last decade, inclusive fitness' value has been called into question by outspoken critics who claim the approach is inherently flawed [26], sparking an intense and ongoing debate [2]. Opponents claim the theory's application is limited by a specific set of assumptions, that the quantities it describes are impractical/impossible to describe empirically, and that it lacks predictive power [1, 27]. However, supporters argue that, with some extensions, inclusive fitness can be fully general, and with adjustments to specific scenarios can also be used to interpret and/or predict selection events [7, 22]. While some experimental studies seemingly indicate positive selection coincides with an increase to inclusive fitness [38, 20], none have yet devised a consistent method for inclusive fitness to be applied in practice without caveats [27].

Issues applying inclusive fitness to empirical data (or indeed, applying any theory in practice) can be divided into two classes. The first is epistemological issues coming from practical limitations of gathering information on the system to be studied. The second are conceptual issues arising from inconsistencies or limitations of the theory itself. Epistemological problems are common when trying to use inclusive fitness theory to analyse evolution of behaviours in real systems, due in part to the complex nature of social networks within populations of interacting organisms. This can make it difficult to investigate conceptual limitations, but problems gathering data can largely be circumvented when using simulation modelling (such as individual based models). As such, for this report problems of the first class are largely ignored, save for some consideration in Section 4 where previous attempts to apply inclusive fitness theory to experimental data found are analysed.

The aim of this project is to investigate the limits of Hamilton's rule as a descriptive tool when applied more generally. To achieve this, we will consider three key questions: How can we define cost, benefit and relatedness in a way that avoids limiting assumptions? Can a general form of Hamilton's rule, with parameters defined in this manner, still give causal descriptions? Why are there seemingly few empirical results that explicitly test Hamilton's rule? By addressing these questions, we hoped to gain insight into the scope and limitations of inclusive fitness theory when applied to individual-based models. However, our search for formulations that can be generally applied proved unsuccessful, due in part to known limitations. Furthermore, we identify other issues that are less commonly discussed in the literature; in the discussion we will analyse to what extent these limitations may restrict the potential scope of inclusive fitness theory, and suggest potential avenues for future research.

2 Formulations of Inclusive Fitness

Hamilton introduced inclusive fitness theory through two perspectives, "inclusive fitness" and "neighbour-modulated fitness" [14]. Both perspectives assume a baseline level of fitness that is not influenced by the social trait being studied, but differ in their focus on the fitness contributions of the trait from the perspective of either the actor or the recipient. To illustrate with a classic example: consider a population of organisms, some of which are altruists. The neighbour-modulated fitness of an altruist in our example population is a sum of the cost it pays for possessing altruism plus the benefits of all altruism it received from social interactions. Its inclusive fitness is instead the cost of possessing altruism (as before) plus the total benefit it has given to other altruists (indirect fitness). Note that within this classic description is the implicit assumption that costs and benefits may be separated and considered separately.

To begin, let us write a mathematical definition of the most basic form of neighbour-modulated fitness - it can be argued that the perspective of inclusive fitness gives more valuable insight into the nature of social behaviours [42], but neighbour-modulated fitness is more convenient for later derivations. Starting at this basic form requires a set of assumptions. First, defining fitness as expected fecundity of an individual, we accept the assumption that this individual's reproductive outcome may be decomposed into a linear sum of fitness effects it experiences, including a baseline fitness independent of the trait being investigated. Second, consider a binary phenotype: altruist or non-altruist. Furthermore, phenotype and genotype are one-to-one related: there is a binary genotype G which may take values of $\{0, 1\}$ which directly corresponds to the phenotype of {non-altruist,altruist}. Each of these assumptions will later be addressed, in Section 3.

With the above assumptions, the neighbour-modulated fitness W of individual may be written as

$$W_i = W_0 - CG + bG', \tag{1}$$

where C is the fitness cost on an individual for possessing the trait, b is the fitness benefit an individual possessing it gives to others, G and G' are the genotypes of oneself and one's interaction partners respectively, and W_0 is a baseline fitness unrelated to the trait. Studying altruism implies the trait is costly to self, hence why the sign before C is negative. One important thing to note is G is the genotype for one individual, but if there are interactions with multiple partners, G' should be the sum of all their genetic/trait values, such b is the benefit received per interaction with an altruist, while C is the overall cost paid (indicated by capitalising). Though equation 1 is built on the example of an altruistic trait that is costly to the bearer and beneficial to interaction partners, it should be clear that by changing the sign of C and b it can be used to investigate the fitness effects of a range of social traits from spite to cooperation.

2.1 Early forms of Hamilton's rule

We will focus on altruism, since that is what inclusive fitness was originally derived to study, as well as where it has arguably garnered the most attention and important results. The most famous of these results is Hamilton's rule, which states the conditions necessary for altruism to evolve.

Its original (non-general) form [14, 13] is derived using the set of assumptions given for equation 1. It also makes an assumption of pairwise interaction, meaning it only considers the interactions between two individuals at a time. This assumption is necessary because it simplifies the calculations required to determine the relatedness coefficient, r, a critical component of the rule that describes the level of genetic similarity between interacting individuals (more on this shortly). Assuming interactions are split into pairs requires considering the cost per-interaction, instead of the overall cost of possessing a trait in equation 1, hence why c in this section is lower-case. With the above assumptions, a single interaction can be written as a payoff matrix, where columns represent genotype of the actor g', while rows represent genotype of the recipient, g:

g\g'	0	1
0	0	b
1	-c	b-c

Note this notation is biased towards the neighbour-modulated perspective, since it considers contributions from the perspective of the receiver. For altruism to be positively selected, the neighbour-modulated fitness of altruists should be higher than the neighbour-modulated fitness of non-altruists (such that future generations contain an increased proportion of altruists). Since we assume a baseline fitness independent of genotype, this corresponds to altruists receiving a higher payoff than non-altruists. We write the payoff received per interaction for a receiving altruist as Φ_1, and for a receiving non-altruist as Φ_0:

$$\Phi_1 = P(g' = 1|g = 1)b - c$$
$$\Phi_0 = P(g' = 1|g = 0)b$$

If interactions are random, altruists cannot receive more payoff than non-altruists, and thus altruism cannot evolve. To demonstrate this, let us write the proportion of altruists in the population as α. If interactions are random, the probability of being the recipient of an altruistic interaction is also α, independent of recipient genotype. Putting this interaction probability into the average payoffs seen above and solving for $\Phi_1 > \Phi_0$ results in $c < 0$, i.e. that the cost of interaction should be negative, which is no longer altruism.

It is clear, then, that for altruism to evolve altruists must preferentially interact with other altruists, while avoiding non-altruists. This preference may be written as the difference in probability of interacting with an individual of the same own altruism status, compared to interacting with one of different status [3]:

$$r = P(g' = 1|g = 1) - P(g' = 1|g = 0), \tag{2}$$

where we have defined a new parameter, 'relatedness' r, to describes this probability difference. Using this new parameter, we can rewrite the probability of being the recipient of an altruistic interaction as follows:

$$P(g' = 1|g = 1) = (1 - r)\alpha + r$$
$$P(g' = 1|g = 0) = (1 - r)\alpha \tag{3}$$

As a probability, r can take values between 0 and 1; $r = 0$ represents random interactions as before, and $r = 1$ represents altruists exclusively interact with other altruists. The overall probability of an interaction having an altruistic actor is still α ($P(g' = 1) = \alpha$ is trivially provable using the law of total probability), only now depending on r value, altruists preferentially interact with other altruists. Using these interaction probabilities to solve the inequality $\Phi_1 > \Phi_0$ gives:

$$rb > c \tag{4}$$

Equation 4 is Hamilton's rule. In essence, it states that altruism is positively selected if the neighbour-modulated fitness of altruists is higher than non-altruists, which can only be true if the fitness costs paid by altruists are less than the fitness benefits they expect to receive.

2.1.1 Hamilton's original derivations require altruism to be rare

While r defined in the probabilistic manner of equation 2 gives an approach to understand how altruism could be favourable, calculating this probability for an empirical setup requires knowledge of all interactions that have happened. Hamilton wished gain more predictive insight into why an individual would preferentially interact with another from the perspective of favouring kin. To do so, he heuristically redefined relatedness in the case of diploid organisms that reproduce sexually, taking the perspective that an altruistic interaction represents a trade-off between one's own reproduction and the reproduction of others (given how likely they are to be altruistic):

> "(relatedness is) the fractional weighting which A (an actor) gives to one unit of B (a recipient)'s fitness compared to one unit of his own." [12]

For an allele to be positively selected, its replicas must form an increasing proportion of the next generation's gene pool. Hamilton reasoned that the relatedness of a single interaction for a focal allele - which A definitely possesses, but B may not - should be the likelihood that recipient B's reproduction produces offspring possessing that allele, compared to the likelihood A's own offspring possess the allele. Hamilton thus considers the 'weighting of one unit of an individual's fitness' as the probability that a gamete of that individual will possess this allele (for sexually reproducing organisms). The relatedness of a single interaction, r, is then the probability a gamete of B possesses the allele, divided by the probability a gamete of A possesses the allele.

However, note that Hamilton himself recognised that his definition had limitations and that a more precise definition was needed in order to make accurate predictions. Alleles for social traits such as altruism can also directly influence the fitness of individuals *not* possessing that allele, and positive selection of an allele requires that it increases not by frequency in a population, but *by proportion* [29]. Hamilton's reasoning that the 'fractional weighing' is the probability of B producing altruistic offspring divided by A producing altruistic offspring corresponds to a frequency increase, not a proportion increase, as the altruistic allele will also increase the frequency of non-altruistic alleles. It neglects the fact that contributions to non-altruist alleles should be included in the calculation and negatively weighted, since increasing their frequency decreases the proportion of altruistic alleles within the population.

While this shows some flaws in Hamilton's 1972 definition of relatedness, it is still a good approximation if altruism is a rare trait [23]. If altruism is rare, contributions to the frequency of non-altruists have a negligible impact on population proportions, and any increase to altruist frequency can be considered as an increase to its population proportion also. From an alternate perspective, if a trait is sufficiently rare then the probability of randomly interacting with an individual carrying that trait is negligible, and relatedness defined for equation 3 covers all interactions involving an altruist (under this simple set up the payoff for non-altruists interacting among each other is zero, so their interactions are ignored).

For now, we will continue with the assumption that altruism is a rare trait (though this will be addressed later in this section). Assume that there is an allele responsible for altruism, and possessing one copy is enough to influence a binary phenotype (i.e. the phenotype of homozygous and heterozygous altruists is identical). Assume also that gametes are produced by a random choice between alleles, i.e. that meiotic drive and other mechanisms that may bias the gametes are negligible (this assumption is made slightly more explicit during derivation involving the Price equation later). For diploid organisms, the probability a gamete of A also possesses altruism, given that A is an altruist, is $\frac{1}{2}(1 - f_A) + f_A$, where f_A is the probability A is homozygous at the altruistic locus. Writing the probability a gamete of B possesses the altruistic allele as f_{AB}, Hamilton denotes the relatedness for a single interaction between a given actor A and recipient B, r_{AB}, as:

$$r_{AB} = \frac{2f_{AB}}{1 + f_A}, \tag{5}$$

an expression found by applying the heuristic definition expressed above to this simple diploid scenario [12]. Averaging over all interactions then gives the relatedness for the population, r. To provide a practical method to calculate this relatedness per interaction, Hamilton interprets probabilities f_A and f_{AB} via Wright's coefficients, which calculate the probability of individuals having the same alleles *by descent* through analysing their lineage [44]. f_A is interpreted as Wright's coefficient of inbreeding for an altruistic actor A at the locus of altruism, the probability that two alleles at that locus are identical by descent. f_{AB}

is interpreted as Wright's coefficient of kinship, the probability that a randomly selected allele at the altruistic locus from A is identical by descent to a randomly selected allele from the altruistic locus of B. If we assume negligible inbreeding, $f_A \to 0$, and equation 5 becomes $2f_{AB}$ [17]. The numerator of equation 5 is Wright's coefficient of relationship, which is often used to describe Hamilton's relatedness by experimental studies that assume negligible inbreeding [25, 16].

At this point, it seems important to note the distinction between 'relatedness' and 'relationship' [6]. Relatedness refers to genetic similarity between individuals, while relationship is calculated from pedigree, and thus requires genes to be identical by descent. It can be seen that Hamilton's relatedness of equation 5 given by r_{AB} is only generally equal to relationship under two conditions:

1. $f_A << 1$.
2. f_{AB} is well-approximated by Wright's coefficient of kinship .

The first of these requires the assumption of negligible inbreeding. The assumptions required for the second are more complex, but can be split into two key considerations: when is genetic similarity approximated by kinship, and how is this kinship approximation affected by selection? Genetic similarity is approximated by kinship if a trait is rare, which implies that alleles are identical by descent. Hence, Hamilton's relatedness can only be approximated by relationship under the assumption of rare trait, but this is required for Hamilton's original relatedness regardless. Key work in the field also claims that weak selection is a necessary assumption [11, 23, 36], claiming that strong selection can cause genetic similarity to deviate from kinship estimates whose accuracy relies on an allele having a 50% likelihood to be passed on to a successful gamete (for diploid organisms). However, the extent this is skewed by selection is not clear, models have indicated that assuming weak selection is robust for approximations of relatedness with linear fitness functions and a simple island structure [24], but the impact of selection on the link between relatedness and relationship for more complex systems is not yet understood.

2.1.2 Relatedness can be adjusted to describe non-rare traits

Issues using relationship to approximate relatedness are hard to avoid, but the mistake in Hamilton's derivation that requires an increase in frequency to be equal to an increase in proportion can easily be corrected. By doing this heuristically, we arrive at a simplified form of the statistical description of relatedness that will later be derived using the Price equation in the next section, which is said to be fully general.

Since this is purely a demonstration that the rare-trait assumption can be avoided, for simplicity let us consider a haploid organism with an allele that determines altruism, with the values 1 or 0. Let us denote the proportion of altruists in the population at an initial time with α. Imagine a probability H that that an altruistic actor interacts with another altruist, H. The probability

an altruistic actor will interact with a non-altruist is then $1 - H$:

$$P(g = 1|g' = 1) = H \qquad P(g = 0|g' = 1) = 1 - H.$$

Since we have assumed interactions are between pairs (this assumption is discussed in section 2.3.2, we can adjust these probabilities to be conditional on the recipient instead of the actor:

$$\begin{aligned} P(g' = 1|g = 1) &= \frac{P(g = 1|g' = 1)P(g' = 1)}{P(g = 1)} &&= \frac{H\cancel{\alpha}}{\cancel{\alpha}} \\ P(g' = 1|g = 0) &= \frac{P(g = 0|g' = 1)P(g' = 1)}{P(g = 0)} &&= \frac{(1 - H)\alpha}{1 - \alpha}. \end{aligned}$$

Let us denote the expected fitness of altruists W_A and expected fitness of altruists W_B. For altruism to be positively selected, the proportion of altruists should be increasing, or:

$$\frac{\alpha}{\alpha + (1 - \alpha)} < \frac{\alpha W_A}{\alpha W_A + (1 - \alpha)W_B}$$

Rearranging the above gives $W_A > W_B$, or that altruists should have higher fitness than non-altruists. Using the assumption of a linear cost function with a baseline fitness, this corresponds to:

$$\cancel{W_0} - c + bP(g' = 1|g = 1) > \cancel{W_0} + bP(g' = 1|g = 0)$$

Substituting our calculated probabilities into this equation and rearranging to the form of Hamilton's rule gives $\frac{H-\alpha}{1-\alpha}b > c$, showing the relatedness:

$$R = \frac{H - \alpha}{1 - \alpha}. \tag{6}$$

This is an example of a relatedness coefficient that doesn't require the assumption of rare altruism, which can also be directly obtained from equation 3 by substituting H. Note that if we do take the assumption of rare altruism, then $\alpha \to 0$ and equation 6 becomes H, which is the relatedness that Hamilton's heuristic reasoning for equation 5 would lead to: for this simple model all haploid individuals have the same probability of giving altruism to their offspring, so from the perspective of the gene the reproduction of all altruistic individuals has the same value. The 'weighing' that A gives to B only needs to consider the probability they are interacting with an altruist. The form of equation 6 will reappear in the next section 2.2, where we realise that relatedness rewritten in terms of a covariance relationship between actors and recipients.

2.2 Hamilton's rule derived using the Price equation

Inspired by Hamilton's results on altruism, George R. Price developed Price's equation [32], a general description of how the average proportion of individuals

in a population possessing a trait/genotype value changes over time. Indeed, the equation is derived so generally that the value investigated need not be trait/genotype, the population can be things other than biological organisms, and changes over dimensions other than time may be considered, resulting in a broad range of applications in fields ranging from economics [19] to cosmology [9], but for our purposes we will be strictly considering its original biological context.

To simplify how to initially consider genetic differences between parent and offspring generations, I will introduce the Price equation from the perspective of haploid organisms reproducing asexually. Note however that this can be easily extended to diploid organisms reproducing sexually by considering offspring as the products of successful gametes and weighing organism genotype values by their ploidy (a derivation of the Price equation for any ploidy can be found in Grafen's *A geometric view of relatedness*, page 33 [11] but no derivation will be given here). Consider a population of reproducing organisms of two generations, parents and offspring, with a trait/genotype value of interest G (difficulties splitting a population over time into discrete generations will be discussed in section 2.4). W represents fitness; in the current context of a population split into parents and offspring it makes most sense to measure this by fecundity. The discrete Price equation describes the change in trait frequency between generations of parents and offspring ($\mathbb{E}(W)\Delta\mathbb{E}(G)$):

$$\mathbb{E}(W)\Delta\mathbb{E}(G) = \mathrm{Cov}(W, G) + \mathbb{E}(W\Delta G). \tag{7}$$

On the left hand side, the change in frequency is split into $\mathbb{E}(W)$ and $\Delta\mathbb{E}(G)$, where $\mathbb{E}(W)$ is the average fitness of the parent population (i.e. the change in population size) and $\Delta\mathbb{E}(G)$ is the difference between average trait values of parents and offspring. On the right hand side $\mathrm{Cov}(W, G)$ is the covariance between parental fitness and trait/genotype value, and $\mathbb{E}(W\Delta G)$ is the average difference between each parent and their offspring, weighted by parent fitness.

Price's equation doesn't aim to predict selection through mechanistic interpretations of biological processes, but rather gives a statistical interpretation of the components of evolutionary change, referred to as selection and transmission [7]. Within the square brackets, $\mathrm{Cov}(W, G)$ represents change due to selection (how much fitness varies with the trait) and $\mathbb{E}(W\Delta G)$ represents change due to imperfect transmission (which can account for factors such as mutational bias, meiotic drive or other non-Mendelian effects). The generality of equation 7 allows for a derivation of Hamilton's rule that does not require assumptions of weak selection or rare mutations [8], considering genetic similarity instead of pedigree relationship. We begin, as before, by showing this derivation in a simplest form, assuming a linear fitness function as seen in equation 1 (and its assumptions).

We will now show that using the Price equation, we can arrive at the simple result shown in equation 6, first by using it to derive Hamilton's rule [33, 22]:

$$
\begin{aligned}
0 &< \mathrm{Cov}(G, W) \\
&\quad \mathrm{Cov}(G, W_0 - CG + bG') \\
&\quad \cancel{\mathrm{Cov}(G, W_0)} - \mathrm{Cov}(G, CG) + \mathrm{Cov}(G, bG') \\
&\quad - C\mathrm{Var}(G) + b\mathrm{Cov}(G, G') \\
C &< b\frac{\mathrm{Cov}(G, G')}{\mathrm{Var}(G)} \qquad (8)
\end{aligned}
$$

For the average trait value to increase between generations, $\Delta\mathbb{E}(G)$ in equation 7 must be positive (and as an expected fecundity, $E(W)$ is also positive). Since we are concerned about the conditions necessary for altruism to be selected, assume an unbiased transmission, $\mathbb{E}(W\Delta G) = 0$. As a result, the selection term must be positive, $\mathrm{Cov}(G, W) > 0$. Substituting equation 1 into the fitness of $\mathrm{Cov}(G, W) > 0$, one arrives back at Hamilton's rule (equation 8). W_0 is assumed constant, hence $\mathrm{Cov}(G, W_0) = 0$. In the same way, if C and b are constants that represent the cost of being an altruist and the benefit received from altruists, and so can be taken out of covariances they are involved in. Note that C is an overall cost, and not cost per interaction. For it to be constant, we thus require considering either single interactions, an expected number of interactions, or a cost unrelated to the number of interactions.

Equation 8 is Hamilton's rule, derived from the Price equation, which provides a new definition of relatedness:

$$R = \frac{\mathrm{Cov}(G, G')}{\mathrm{Var}(G)} \qquad (9)$$

Notice that this can be linked back to the heuristic relatedness we defined earlier to avoid the assumption of altruism as a rare trait: with the probability of altruist-altruist interaction H and proportion of altruists α, equation 9 gives:

$$R = \frac{\mathrm{Cov}(G, G')}{\mathrm{Var}(G)} = \frac{\cancel{\alpha}(H - \alpha)}{\cancel{\alpha}(1 - \alpha)},$$

where we used the covariance identity $\mathrm{Cov}(G, G') = \mathbb{E}(GG') - \mathbb{E}(G^2)$. As such, under the assumptions of negligible inbreeding/weak selection and rare mutants, this definition of relatedness is equal to the pedigree definition given by equation 5 (a more formal proof of this is given by Orlove et al. [29]), but without these assumptions they are not equivalent. This statistical form of relatedness allows b and C to be interpreted as regression coefficients which describe the slope of best fit between fitness and trait value [34] (between W and G for C or W and G' for b), an idea which will be expanded upon in the next section.

2.3 Hamilton's rule's parameters as regression coefficients

2.3.1 Early simple regression techniques are superseded by partial regression

We have, until now, assumed that fitness is linear, with the genotype of self G and interaction partners G' as independent variables and cost and benefit as coefficients, but applying the Price equation to redefine relatedness shows the potential for writing parameters as linear regression coefficients, since equation 9 shows relatedness as the regression coefficient given by regressing interaction partner genotype G' on genotype G. By using linear regression to fit a linear model to data, linearity is not assumed but instead approximately fit, with costs and benefits written as coefficients of a regression of fitness on genotypes. Using a linear model for the fitness substituted into the price equation, residuals are distributed around 0 and so drop out of Hamilton's rule [22], leaving the three parameters of cost, benefit and relatedness as usual.

The possibility of representing cost and benefit as regression slopes was first found found using simple linear regressions: the cost C as the slope of a simple regression of fitness on genotype of self (W on G), and benefit b as the slope when simply regressing fitness on genotype of interaction partners (W on G') [22]. However, simple regression techniques are inaccurate fitting multiple independent variables with correlations between them. As we have seen, for altruism to evolve there must be some form of preferential structure for altruists to interact with other altruists, hence G and G' should be correlated. In the multiple regression of W on G and G', both are predictor variables; if they are correlated, a simple regression method (which assumes independence of predictors) is unsuitable, and partial regression should be used. Queller [33] uses the notation $\beta_{WG|G'}$ to denote the partial regression coefficient of a regression of fitness W on G, while holding G' constant; as an alternative perspective, it predicts variation in W with the variation in G that isn't predicted by G' (A proof of this, as well as the formula of partial regression derived from residual sum of squares, is given in appendix 5.1). Note if G and G' are independent from each other then partial regression gives the same coefficients as simple regression (but otherwise not). Writing $\beta_{G'G}$ as the simple regression of G' on G for relatedness, and using Queller's notation for the partial regressions of cost and benefit, we can rewrite equation 8 in the form:

$$\beta_{WG|G'} + \beta_{G'G}\beta_{WG'|G} > 0, \tag{10}$$

where $\beta_{WG|G'} = -C$, $\beta_{WG'|G} = b$ and $\beta_{G'G} = R$.

Taking partial regression coefficients attempts to isolate the effects of G and G' on fitness, such that C and b can be interpreted as 'cost paid for possessing a trait' and 'benefit received from others'. However, fitting a linear fitness function of the form seen in equation 1 has a problem: fitness effects of social interactions are often non-additive. Taking a linear regression is still always possible, no matter how poor the fit, with differences between observed and predicted values absorbed by the regression's residuals, and it will still be able to

correctly describe the overall direction of selection. That being said, imposing a linear approximation onto a non-linear fitness function can cause complex underlying relationships between the parameters involved, such that $\beta_{WG'|G}$ and $\beta_{WG|G'}$ no longer solely represent the fecundity payoffs from an interaction, but also contain information about interaction structure (which we would prefer to fully contained in R). A demonstration of this property is shown in the next section discussing non-linear fitness functions.

2.3.2 Partial regression can describe non-linear fitness, but makes interpreting parameters confusing

We aim to illustrate that non-linear fitness functions can introduce complex relationships into cost and benefit, by using the example of Queller's synergistic extension, which describes a simple non-additive, pairwise interaction where interacting altruists receive additional benefits for interacting with other altruists [35]. Consider a payoff matrix as in section 2.1 (with the same starting assumptions), but now with a synergistic effect d, a fecundity benefit that is only received if both actor and recipient are altruists:

g\g'	0	1
0	0	b
1	-c	b-c+d

The expected fecundity payoff per interaction for altruists, Φ_1, and non-altruists, Φ_0, now take the following form:

$$\Phi_1 = P(g' = 1|g = 1)(b + d) - c$$
$$\Phi_0 = P(g' = 1|g = 0)b$$

As in section 2.1, we call the proportion of altruists in the population α, and then use probabilities of interacting with an altruist given by equation 3. Solving for $\Phi_1 > \Phi_0$ gives a condition for altruism to evolve:

$$rb - c + d\left((1 - r)\alpha + r\right) > 0. \tag{11}$$

Equation 11 looks different from the original form of Hamilton's rule given in equation 4. However, if we now approach this with linear regression, we see the original form return. Imagine that as defined for equation 3, r gives a hypothetical perfect value of the preferential probability that individuals will pair with partners of the same altruism status, as opposed to random interactions, and α is the proportion of altruists in the population. With r and α, we can write the relative frequencies of interaction pairs, such as:

$$\begin{aligned} P(G = 1 \cap G' = 1) &= P(G = 1)P(G' = 1|G = 1) \\ &= \alpha((1 - r)\alpha + r) \\ &= \alpha^2 + r\alpha(1 - \alpha) \end{aligned}$$

Recipient genetic value G	Actor genetic value G'	Relative frequency $P(G \cap G')$
0	0	$(1-\alpha)^2 + r\alpha(1-\alpha)$
0	1	$(1-r)\alpha(1-\alpha)$
1	0	$(1-r)\alpha(1-\alpha)$
1	1	$\alpha^2 + r\alpha(1-\alpha)$

All relative frequencies of interaction pairs are given by the following table:

Performing partial linear regressions on these interactions using the relative frequencies of interaction and their expected fecundity payoffs (an example of the partial derivative formula is seen in equation 19) we obtain:

$$\beta_{WG|G'} = -c + \frac{1}{1+r}(r + (1-r)\alpha)d$$
$$\beta_{WG'|G} = b + \frac{1}{1+r}(r + (1-r)\alpha)d$$

These partial regression coefficients can be used to describe Queller's synergistic extension in the familiar form of Hamilton's rule given by equation 10. The partial regression coefficients above also give the same predictions as the explicit solution of equation 11, since substituting these regressions into equation 10 simplifies to the same solution. That being said, our regression terms for cost and benefit now include descriptions of interaction structure: they are no longer independent of the relatedness, making them harder to interpret when the underlying fitness function is unknown.

2.3.3 Non-linear extensions only provide additional value for specific cases

Using partial regression to fit an assumed linear fitness function gives a relatively general method to predict the direction of selection. When the underlying fitness is non-linear then regression coefficients for cost and benefit stray from their initial interpretations as fecundity payoffs since they also include information about interaction structure. Non-linear fitness functions can instead be considered in an attempt to minimise the inclusion of interaction structure within the b and c parameters and make them more mechanistically interpretable. Changing the assumed fitness function also changes the form of Hamilton's rule, such as Queller's synergistic extension, which may also be written in the form [35, 22]:

$$\beta_{WG} + \beta_{G'G}\beta_{WG'} + \beta_{(GG')G} > 0. \tag{12}$$

This equation is based on fitting a model for fitness that includes a synergistic extension term; if the real fitness function of an experiment is approximated more accurately by including this term then regressing against it could describe the components responsible for changing fitness more accurately, or make it easier to interpret individual parameters by avoiding the absorbtion of interaction structure into cost and benefit terms (as seem with equation 11). However,

this difficulty of interpretation can persist within these new regression parameters: in the case of this simple synergistic extension, a regression approach gives $\beta_{(GG')G} = d((1-r)\alpha + r)$, now including information about interaction structure within the synergy term.

Van Veelen et al. [40] claim that Hamilton's rule fails for synergy in the case of more than two participants, claiming that Hamilton's rule contains an insufficient number of parameters. They provide a hypothetical example of a three-player game in order to illustrate this. However, Gardner et al. [7] provide a rebuttal: they expand the interaction matrix to include 2 interaction partners, perform a linear regression and show that similarly to the synergistic example given above, complexities are absorbed into the regression terms of cost and benefit (note this is still done for single interactions, a point that will be expanded upon in Section 2.4). In doing so, Gardner generates the same predictions of positive selections using Hamilton's rule that were generated by Van Veelen's explicit solution, despite the initial claim that an approach using Hamilton's rule would be insufficient. To say Hamilton's rule contains insufficient parameters to model social interactions of a dimension higher than 2 is thus incorrect; it ignores functions of complex interactions within the parameters of cost and benefit. While this shows potential to extend to higher dimensional interactions, it also shows that difficulties interpreting parameters continue to increase when considering this: formulae for cost and benefit in Gardner's solution for 3 dimensional interactions are significantly more complex than in the 2 dimensional case. Gardner's method could even be extended to cases of N-dimensional interactions, where the effect of every possible combination of cooperators and non-cooperators as social partners is considered. However, implementing this approach in practice can be challenging due to the complexity of interaction structure. Some researchers have proposed strategies for simplifying these structures, but such tricks are only applicable to specific systems (see experimental example 5 of section 4 [37]).

2.3.4 Phenotype describes costs and benefits more causally than genotype, but mapping is complex

Considerations of fitness functions with non-linear interactions have been shown to still be generally approachable using partial regression, but with the penalty of cost and benefit parameters straying from their causal interpretations. In a similar way, regression calculations for the relatedness seen until now also neglect one causal aspect of interactions: it describes interaction structure using correlations between genotypes, but costs and benefits of altruistic interaction are determined by phenotype. Until now, 'genes' and 'trait values' have been used somewhat synonymously, but the mapping between a discrete genotype and continuous phenotype is often complex and non-linear [22].

Forms of Hamilton's rule that consider impact of phenotype on fitness aim to bring b and c close to their causal interpretation by using performing a linear regression to fit fitness to a function that depends on phenotype, instead of genotype [33, 22]. Consider a fitness function resembling the simple linear form

given in equation 1, where genotype is replaced by phenotype:

$$W_i = W_0 - C_P P + b_P P'.$$

Here, P and P' represent the phenotype of self and interaction partners, respectively, and C_P, b_P are the partial regression coefficients found by regressing fitness against phenotype instead of genotype; because fecundity changes should be directly caused by phenotype, it is hoped the coefficients of this regression may be more causally interpretable as cost and benefit.

Using the Price equation, the condition for positive selection remains as $\mathrm{Cov}(W, G) > 0$. Substituting the phenotypic fitness function into this condition gives the phenotypic extension to Hamilton's rule [33]:

$$\begin{aligned} 0 \quad &< \quad \mathrm{Cov}(G, W_0 - C_P P + b_P P') \\ &\quad - C_P \mathrm{Cov}(G, P) + b_P \mathrm{Cov}(G, P') \\ C_P \quad &< \quad b_P \frac{\mathrm{Cov}(G, P')}{\mathrm{Cov}(G, P)}. \end{aligned} \tag{13}$$

Dividing by $\mathrm{Cov}(G, P)$, we see this extension gives a new form of the relatedness: $R = \frac{\mathrm{Cov}(G,P')}{\mathrm{Cov}(G,P)}$ [35, 22]. Could Hamilton's rule in this form describe altruism in the same way as forms we have already covered?

Attempting to construct a demonstration of this is difficult. Simplifications until now have required Hamilton's rule to consider genotype and phenotype as binary, since there has been the goal of describing individuals participating in multiple interactions, and without binary altruism it is difficult to perform regressions describing where contributions come from. However, in order to give a speculative demonstration that under certain conditions, phenotypic relatedness could resemble Hamilton's rule, consider a system where pairwise interaction partners only interact with one partner for their entire life, to allow for a non-binary genotype/phenotype where fecundity contributions can be known. Assuming phenotype is exclusively dependent on genotype (or that any other variables involved can be written in terms of genotype), we write function f that maps genotype onto phenotype: $P = f(G)$. If f is differentiable at the genotype means $\{\bar{G}, \bar{G}'\}$ and well approximated by the first order Taylor expansion (i.e. sufficiently linear) then linking regressions on phenotype to those on genotype is relatively trivial:

$$\begin{aligned} R_P = \frac{\mathrm{Cov}[G, P']}{\mathrm{Cov}[G, P]} &= \frac{\mathrm{Cov}[G, f(G')]}{\mathrm{Cov}[G, f(G)]} \\ &\approx \frac{\mathrm{Cov}[G, f(\bar{G}') + f'(\bar{G}')(G' - \bar{G}')]}{\mathrm{Cov}[G, f(\bar{G}) + f'(\bar{G})(G - \bar{G})]} \\ &= \frac{f'(\bar{G}')\mathrm{Cov}[G, G']}{f'(\bar{G})\mathrm{Var}[G]} = \frac{f'(\bar{G}')}{f'(\bar{G})} R \end{aligned}$$

If cost and benefits are simple regressions then this set of assumptions imposed on f can then be used to transform equation 13 into equation 8. An example

using the cost parameter:

$$\begin{aligned} f'(\bar{G})C_P &= f'(\bar{G})\frac{\mathrm{Cov}[W,P]}{\mathrm{Var}[P]} \\ &= f'(\bar{G})\frac{\mathrm{Cov}[W,f(G)]}{\mathrm{Var}[f(G)]} \\ &= f'(\bar{G})\frac{\mathrm{Cov}[W,f(\bar{G})+f'(\bar{G})(G-\bar{G})]}{\mathrm{Var}[f(\bar{G})+f'(\bar{G})(G-\bar{G})]} \\ &= \frac{\mathrm{Cov}[W,G]}{\mathrm{Var}[G]} = C \end{aligned}$$

This is, however, a very simple case that cannot be scaled up to multiple interactions, as we have needed to include non-binary genotype/phenotype for f to be differentiable. By assuming f is well approximated by its first order Taylor expansion we have implied a relatively linear relationship between genotype and phenotype, as well as a function determining phenotype that only depends on genotype without other factors such as environmental influence. In addition, this equivalence has been shown for simple regressions, the accuracy of which decreases with correlation between independent variables. For altruism to evolve, P and P' should be correlated, since G and G' must be correlated for altruism to be positively selected, and regardless of environmental factors one can still typically expect some correlation between $\{G, P\}$, and $\{G', P'\}$. However, attempting to prove (or disprove) that these considerations of phenotype can be generally absorbed into a partial regression form of equation 10 is non-trivial, with potential functions for cost and benefit quickly expanding to become intractable.

2.4 The discrete Price equation struggles describing changes over non-discrete generations

Arguments that partial regressions can always absorb non-linearity into costs and benefits have led to the conclusion that Hamilton's rule is as general as Price's equation itself, holding true whenever Price's equation holds [4, 7, 22]. While Price's equation is very general, it can still be difficult to apply depending on properties of the system it attempts to describe, which can lead to inconsistent predictions [28]. Derivations shown until now have used single pairwise interactions, where fitness effects of a social trait can be decomposed into direct and indirect effects more easily, but the time scales these interactions occur at has been ambiguous. Hamilton's rule uses the discrete Price equation (equation 7), which describes the difference in a character of interest between two discrete 'assemblages', whereas social interactions of biological systems take place over continuous time. Such overlapping generations and the concomitant continuous change in a population interaction structure complicates the application of inclusive fitness in ways that, to our best knowledge, have largely been overlooked.

We have previously used the common analogy of parent and offspring generations with the implication that interactions between the parent generation

influence their reproductive ability to affect the offspring generation, but in practise these generations may not be clearly separable, allowing different generations to interact with each other. Current processes of applying inclusive fitness theory typically look to consider the parent generation as the generation at some starting point in time t_1, with an offspring generation as the population at some ending time t_2; fitness is considered as fecundity, i.e. the number of individuals at time t_2 that are direct offspring of time t_1 and a fitness function is made that projects the impact of trait values (be they genotype or phenotype) of individuals and their interaction partners at time t_1 (i.e. a description of the interaction structure at t_1, or the relatedness) onto their reproductive success calculated from the number of offspring they have generated by time t_2. This leads to some inconsistencies: interactions are continuously happening at all times between t_1 and t_2, not only with individuals present at time t_1 in a different interaction neighbourhood but also potentially with offspring given birth to before time t_2; yet these interactions are not considered by a relatedness that only takes a snapshot of interactions at t_1. Despite this, the benefit and cost measure fecundity contributions that may have values dependent on length of timestep, and as such may incorrectly attribute fecundity payoffs to an interaction structure not responsible for the observed change. A relatedness derived from a snapshot at t_1 could potentially be a fair representation of all interactions between t_1 and t_2, such as in a system where interaction structure is mediated by physical distance and distance moved between timesteps is small relative to the distance interactions are given, or if the act of interacting binds you to a partner for a non-negligible duration. However, increasing the timestep between t_1 and t_2 still reduces the correlation between the interaction structure at t_1 and that between t_1 and t_2. As such, applications of inclusive fitness to non-discrete generations that look to use the Price equation perform better with smaller timesteps.

Assessing who is giving and receiving fitness contributions between t_1 and t_2 also becomes more difficult for longer timesteps. For example, consider two altruistic individuals A_1 and A_2 that do not interact with each other, but A_2 gives a fitness benefit to the offspring of A_1. In this case, no interaction has taken place between A_1, but their lineage is fitter as a result of interactions with A_2. To assess fitness contributions over a timestep long enough for interactions between individuals not present at the beginning of the timestep to occur, contributions can be considered at the level of lineage rather than individuals. At level of lineage, altruistic interactions between individuals of the same family represent a lowered cost for lineage carry the trait of altruism.

3 Practical Applications

Difficulties interpreting the results may be why the number of studies that attempt to fully describe an empirical system using Hamilton's rule is limited [5]. Extensions of Hamilton's rule testing its generality can be demonstrated on theoretical models such as the derivations we have seen, but setting up theoretical models to demonstrate the evolution of altruism requires crafting initial conditions such that altruists preferentially interact with other altruists (i.e. an altruist is more likely to receive altruism than a random individual in the population), resulting in hypothetical set ups that, though interesting, can feel somewhat removed from biological context [41].

Beginning this project, we aimed to consider how one would apply Hamilton's rule to a computational model of altruism that can evolve population structure on its own, without needing to be pre-sorted into groups. An example of such a model would be the individual-based model of Hermsen et al. [15], which describes a population of competing actors which can reproduce (asexually), express varying levels of altruism, inherit altruism from their parent (with mutation), and move through random diffusion. Both competition and altruism given by an individual are modelled by a Gaussian distribution with its mean at their position, and a variance that determines the scales at which altruism and competition are experienced. If the scale of competition is larger than that of altruism and motility is limited, colonies of altruists emerge [15], thus self-organising into an interaction structure where altruists preferentially interact with other altruists.

However, applying Hamilton's rule to the model proved difficult. In addition to the problem of how to consider non-discrete generations, the fitness function is also dependent on a competition term that scales with density. Altruists form colonies with greater density than non-altruists, meaning the expected competition experienced by altruists and non-altruists is different. This could potentially be absorbed into regression coefficients as an increased cost, but competition is non-linear and non-altruists also experience it; it was not possible to calculate to what degree cost terms found by partial regression include this competition from the fitness function alone. In addition, generations are non-discrete, and we have discussed no solution to this problem was found. As a result, we tried to focus on even simpler problems, to see where the boundary of difficulty of application lies.

3.1 Applications to simple theoretical models are difficult to extend generally

D.S. Wilson used a simple theoretical model in an attempt to outline the conditions necessary for social traits to evolve, and though he himself did not explicitly state it, it demonstrates Hamilton's rule [43]. Assumptions in its setup allow for a simple description of interactions and non-discrete generations. It begins with mixed, isolated groups of individuals that differ in one trait; consider the simplest case with two groups and a binary trait, type A or type B

Figure 1: The original image of Wilson's group set-up used to demonstrate the evolution of a social trait [43]. Groups are identical in size (N=20), the group on the left has 5 A, 15 B ($\alpha = 0.25$) while the group on the right has 15 A, 5 B ($\alpha = 0.75$).

(Figure 1). The trait affects the fitness of an individual, but also that of an individual in the same group. In between generation steps, the groups are mixed and re-separated, and a new generation begins. Wilson intended to use this contrived setup to demonstrate a form of group selection, but gives it biological justification by comparing the theoretical scenario to eggs laid on a leaf, where caterpillar may travel a small distance and interact with local environment before dispersing.

Individuals carrying trait A affect everyone around them except for themselves. Wilson begins with a linear effect of interactions, giving the per capita fitness change for individuals of type A and B in terms of an expected effect of the trait on self $-c$ and effect given to others b:

$$\text{Average fitness change in A} = -c + N(\alpha - 1/N)b$$
$$\text{Average fitness change in B} = N(\alpha)b$$

where N is number of individuals in a group (groups are of equal size), and α is the proportion of A individuals in a group (β is the percentage of B individuals present). For trait A to be positively selected, altruists must receive more benefits on average than non-altruists. The population is split into equal groups indexed by i. For A individuals to be positively selected overall their fitness change per individual averaged over the groups should be greater than that of B. Rearranging this statement, using the equal sizes of groups to remove a factor of N and pulling out cost term returns us to something resembling Hamilton's rule once again, with a relatedness of $R = N\left(\frac{\sum_i \alpha_i\beta_i}{\sum_i \beta_i} - \frac{\sum_i \alpha_i^2}{\sum_i \alpha_i}\right)$.

$$\frac{\sum_i N\alpha_i[-c + N(\alpha_i - 1/N)b]}{\sum_i N\alpha_i} > \frac{\sum_i N\beta_i[N\alpha_i b]}{\sum_i N\beta_i}$$
$$\frac{\sum_i N\alpha_i c}{\sum_i N\alpha_i} > \frac{\sum_i N\alpha_i\beta_i}{\sum_i \beta_i} - \frac{\sum_i N\alpha_i(\alpha_i - 1/N)b}{\sum_i}$$
$$c < b\left[N\left(\frac{\sum_i \alpha_i\beta_i}{\sum_i \beta_i} - \frac{\sum_i \alpha_i^2}{\sum_i \alpha_i}\right)\right]$$

However, if we try to substitute an arbitrary fitness function that considers one's own genotype G and the proportion of type A individuals in their group such as $W = f(G, \alpha)$, we obtain:

$$\frac{\sum_i N\alpha_i[f(G=1,\alpha_i - 1/N)]}{\sum_i N\alpha_i} > \frac{\sum_i N\beta_i[f(G=0,\alpha_i]}{\sum_i N\beta_i}$$

Further simplification requires the assumption that cost and benefit terms in the function can be separated, even with equal group sizes and discrete generations.

3.2 Counterfactual cost and benefit are causally interpretable, but require similar assumptions

Ideally, Hamilton's rule wishes to describe a causal relationship between genotype and fitness, which regression models based on correlation do not always imply. In order to try to provide causal interpretations, costs and benefits can be described counterfactually [39], and use of counterfactual logic is typically how biologists like to think of these costs and benefits : for example, to understand the cost of carrying a gene, one might ask 'what would my fecundity be if I didn't hold this gene?'.

In this novel work, our goal was to find whether Hamilton's rule can be formally defined using purely causal definitions. We will see this is only possible if we are willing to make specific assumptions (assumptions that have also been necessary for some previously demonstrated forms of Hamilton's rule). We begin by writing formal definitions of how to use counterfactual logic to describe cost and benefit using Judea Pearl's 'do calculus' [30] (as well as a new notation, described in Table 1):

$$c_1 = \mathbb{E}_{\Omega 1}[W_i|\mathrm{do}(G_i = 0)] - \mathbb{E}_{\Omega 1}[W_i] \quad = F_{1++} - F_{1-+} \tag{14}$$

$$b_k = \mathbb{E}_{\Omega k}[W_i] - \mathbb{E}_{\Omega k}[W_i|\forall j \neq i, \mathrm{do}(G_j = 0)] \quad = F_{k\bar{x}+} - F_{k\bar{x}-}, \qquad k \in \{0, 1\} \tag{15}$$

'Do calculus' aims to show causal relationships by comparing the real world to a counterfactual one where the 'Do' operation has been performed. For example, equation 14 shows the fitness cost an altruist pays over a unit of time as the difference between its actual fitness, and the fitness that it would have if it wasn't an altruist, but still received all the benefits it currently receives . In the same way, equation 15 describes the fitness benefit an altruist (or non-altruist) receives as the difference between its actual fitness, and the fitness that it would have if all its interaction partners were non-altruists .

Equations 14 and 15 describe a general method to mechanistically derive the cost and benefit experienced for a simulation, without requiring knowledge of social interaction structure. Do notation allows us to write all possible combinations of experienced cost and benefit for both altruists and non-altruists, which we can simplify to the form $F_{k\mp\pm}$ (Table 1). Note that we have chosen to

F notation	Do notation	Cost	Benefit
F_{1++}	$\mathbb{E}_{\Omega 1}[W_i \vert \mathrm{do}(G_i = 0)]$	no	yes
F_{1-+}	$\mathbb{E}_{\Omega 1}[W_i]$	yes	yes
F_{1+-}	$\mathbb{E}_{\Omega 1}[W_i \vert \forall j, \mathrm{do}(G_j = 0)]$	no	no
F_{1--}	$\mathbb{E}_{\Omega 1}[W_i \vert \forall j \neq i, \mathrm{do}(G_j = 0)]$	yes	no
F_{0x+}	$\mathbb{E}_{\Omega 0}[W_i]$	n/a	yes
F_{0x-}	$\mathbb{E}_{\Omega 0}[W_i \vert \forall j \neq i, \mathrm{do}(G_j = 0)]$	n/a	no

Table 1: F notation, its corresponding Do notation, and the experienced costs and benefits they represent

invert the signs of cost and benefit, as they have opposite fitness effects, and the 'do fitness' of non-altruists is written in the form $F_{0x\pm}$ as we are not interested in the potential costs experienced by turning non-altruists into altruists.

Figure 2 shows an illustration of how counterfactual costs and benefits are calculated through differences between the fecundities found from different counterfactual scenarios and the corresponding F notation and cost/benefit labels involved. Importantly, under this construction $F_{1+-} \not\equiv F_{0x-}$, i.e. the fitness of an altruist after switching all altruists to non-altruists is not necessarily equivalent to the fitness of a non-altruist after switching all altruists to non-altruists (shown in Figure 2 as 'x?'). This can be described as the assumption of equivalent environments. Non-equivalence is reasonable: in viscous populations, for example, altruism may provide a means to alleviate competition between kin and attain higher population density [18], and the removal of altruism in such a scenario would have a greater impact on denser colonies, resulting in a lower fitness for altruists than non-altruists. In addition, $F_{1++} - F_{1-+} \equiv c_1 \not\equiv F_{1+-} - F_{1--}$ (shown as 'c1?' in Figure 2). Assuming these to be equal is the assumption of weak additivity, that costs and benefits may be separately applied without loss of generality [4].

We now look to re-examine Hamilton's rule using these counterfactual costs and benefits. Starting from the most general statement possible: for altruism to be positively selected, the expected fitness of altruists should be higher than the expected fitness of non-altruists:

$$F_{1-+} > F_{0x+} \quad , \qquad \text{or} \qquad F_{1-+} - F_{0x+} > 0.$$

Using equation 15, we can always separate benefits from both terms:

$$(F_{1--} + b_1) - (F_{0x-} + b_0) > 0.$$

However, to separate our counterfactual cost, c_1, we must assume weak additivity, so $c_1 = F_{1+-} - F_{1--}$:

$$-c_1 + b_1 + F_{1+-} - F_{0x-} - b_0 > 0.$$

Finally, assuming equal environments between altruists and non-altruists gives $F_{1+-} = F_{0x-}$, resulting in:

$$b_1 - b_0 > c_1 \tag{16}$$

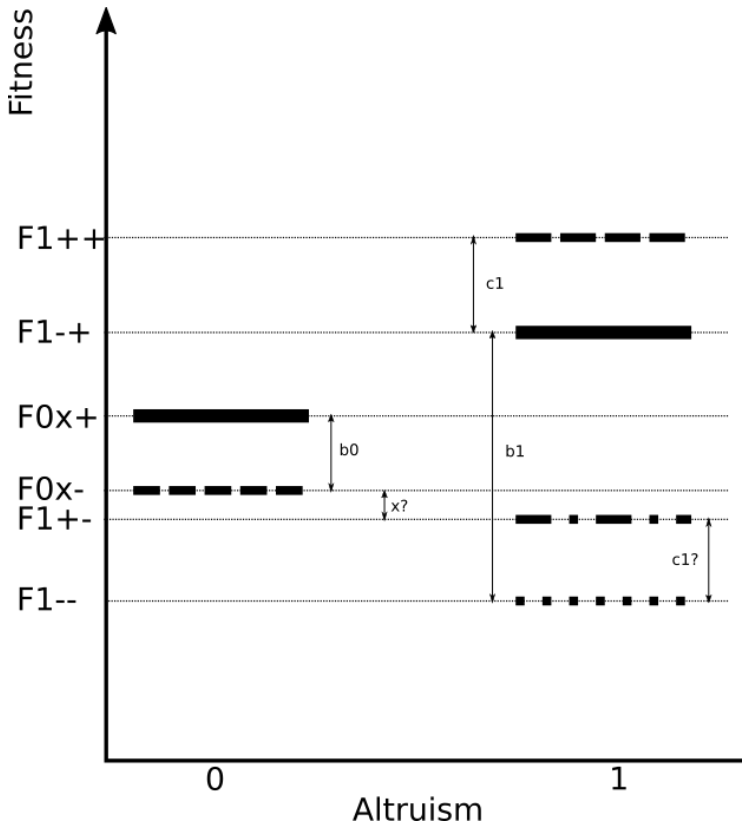

Figure 2: A graph of fitness against genotype for counterfactual cases to illustrate counterfactual costs and benefits.

Equation 16 states the difference between benefit experienced by altruists and benefit experienced by non-altruists must be greater than the fitness cost paid. This somewhat resembles Hamilton's rule, describing a benefit, some interaction structure, and a limiting cost. If benefits received are a constant moderated by some interaction structure $f(G, G')$ and relatedness is the difference in interaction structure between altruists and non-altruists, $R = f(1, G'_1) - f(0, G'_0)$, then equation 16 becomes Hamilton's rule, but a method to define R in terms of counterfactuals to allow moving from $b_1 - b_0$ to Rb was not found.

4 Analysis of Experimental Results

Difficulties in application to natural systems result in limited explicit tests of Hamilton's rule [5]. Nevertheless, there are still some experimental studies that indicate traits/behaviours are selected for to increase inclusive fitness; these studies do so by applying Hamilton's rule with assumptions that each argues can be relevantly made for the specific system they study. In this section, we look to analyse some of these experiments to consider the difficulties they experience in attempting to apply Hamilton's rule, and whether these limit their conclusions. Bourke et al. [5] performed a literature survey to find a non-exhaustive list of 12 studies that explicitly estimate all parameters of Hamilton's rule, with search criteria to only include single species systems that commonly exhibit 'altruism'. Several of these studies look at similar systems, for example 3 articles on the list observe carpenter bees; difficulties applying Hamilton's rule experienced in studies of similar systems are typically the same, as well as the studied behaviours, so we have considered a shortened list (table 2). One addition not present in Bourke's survey has also been given (5.).

	Species	Social Behaviour	Source
1.	Allodapine bee *X.pubescens*	Usurped females remain to guard nest	[16]
2.	Lace bug *G.solani*	Females dump eggs to be guarded	[21]
3.	Tiger salamander *A.tigrinum*	Young cannibalise non-kin	[31]
4.	Wild turkey *M.galloparvo*	Males participate in group mating dances	[20]
5.*	Myxobacteria *M.xanthus*	Cells aggregate to form fruiting bodies	[37]

Table 2: A list of the experimental studies of Hamilton's rule discussed in this section. Study 5 was not present in Bourke's original list.

Model systems are chosen as examples to attempt to demonstrate Hamilton's rule usually because of perceived properties that make them easier to be described. However, for each of these, different definitions of relatedness, cost and benefit are used, with seemingly little protocol of how to be able to compare between experiments. For example, several [16, 20, 31] use pedigree relatedness, but only one [20] provides evidence that genetic similarity is reasonably approximated by pedigree kinship, and even in this case available data was limited, and while the approximation is shown empirically, it is not demonstrated whether this accuracy is because they can reasonably assume altruism is a rare trait; in fact for all cases the altruistic behaviour described is relatively common. In addition to inconsistent descriptions of relatedness, cost and benefit are also defined in different ways. Some studies [31, 16] attempt to use causal methods to approximate the cost or benefit of actions, but do not demonstrate that fitness effects can be separated, or that fitness is approximately linear. On the other

hand, for studies that use correlation to make a linear model for fitness [37, 20, 21] discuss their parameters in a causal manner, but do not address the effects of whether behaviour is rare on the statistical properties of fecundity differences between altruists and non-altruists.

Studies argue the specific adjustments they make to Hamilton's rule can be justified for their use cases, and that Hamilton's rule is being upheld, but the lack of consistency applying it limits how much information this gives. Furthermore, biological complexities of any real social behaviour allow for several counterarguments on why assumptions made for specific systems may neglect describing important behaviour.

1: The first study by Hogendoorn et al. [16] looks to describe the observed behaviour of nest-guarding for carpenter bees. *X.pubescens* forms primitive social nests with dominant reproducing females and non-reproductive guards. Subordinate guards are mostly composed of young pre-reproductive females, which stay in the nest for up to two weeks before leaving to start their own/attempt to dominate another nest. Dominance is maintained not through any chemical repression, but by the dominant female physically wrestling competitors from the egg laying chambers; if a new dominant female takes over, they destroy most of the previous brood. Interestingly, formerly reproductive females that lose a battle of dominance sometimes also stay to guard the new brood, despite a high likelihood that none of their own offspring are within it.

As a model system, it exhibits some traits that allow for a relatively smooth application of Hamilton's rule. Social interaction typically only happens between two interacting individuals at a time, the new dominant female and the former dominant female, and this interaction marks the start of a new generation, since upon taking over the nest the new female destroys current brood. In addition, broods take around 45 days to incubate, a significant length of time with respect to the expected remaining lifespan of a recently deposed dominant female. As such, effects of interactions between individuals of different generations are minimised, since takeovers are typically done by daughters of the former dominant female, and even if not will likely be of that generation. Number of guards in a colony is low, typically around 1-8 at one time, so individual contributions to expected brood survival can be calculated as a non-negligible amount. Since females that lost hierarchy battles always have the chance to leave to start/usurp a new nest, cost can be calculated as the expected loss of reproductive function between these two cases. Relatedness is calculated as pedigree, which is easy to follow since most interaction happen between daughters or sisters, and outsiders are considered to have a relatedness of 0. The study finds this form of social interaction is seemingly selected through indirect fitness, and that formerly dominant females preferentially stay as guards if their relatives take over.

Even in this excellent use case, there are still potential points of contention. There are important details regarding how dominance is behaviour is selected the authors exclude: while formerly dominant females did preferentially stay as guards depending on relatedness to new females, the biggest deciding factor in whether a queen stayed behind or left was their age at the time of takeover.

Young females have higher chances of successfully establishing a new nest, meaning the potential cost to pay for staying as a guard is much higher than that for old females. Dominance hierarchy is established by fighting, so physical status may also have a role to play in deciding whether to stay/leave. Bees were also shown to prioritise staying as a guard if their young offspring are currently present, which will not always be successfully removed from the nest and can instead win the struggle for dominance, representing a reduced cost to the guard, meaning the cost term likely has several relationships within it. In addition, dominant females are also primary foragers of the nest, meaning they are the only ones that leave, how this interacts with expected lifespan was a topic of discussion in the article with speculative conclusions. Aside from any of this, the choice of leaving or staying is not only taken by the former dominant female, but also the new dominant female, who decides whether or not to allow the previous dominant female to stay in the nest, which may be mediated by the threat they still represent.

2: Lace bugs of the species *G.solani* exhibit maternal care, with females guarding and caring for clusters of their eggs [21]. They also participate in egg dumping, where some females will lay their eggs on clusters of eggs already present, in order for them to also receive the guarding and maternal care from other females. Unlike in the case of parasitism, however, Loeb et al. observe that lace bugs preferentially dump eggs with those of kin. They find experimentally that by dumping eggs, lace bugs improve the survival percentage of the original brood, since when they hatch juveniles will experience less predation. They calculate relatedness using genetic markers and cost and benefit through expected survival probabilities to argue that this example of egg dumping is kin-selected.

The argument that egg dumping is altruistic requires the dumpers to pay a cost. On the contrary, dumpers are more likely to survive to produce a second generation of eggs, since they do not participate in maternal care and so do not need to guard. Experimental evidence did indicate that differences between the two behaviours give negligible fitness benefits for the dumpers long-term, but there were also no observed genotypic differences in fertility, which would imply that opportunity to survive should at least represent a slightly positive effect. A negligible cost imposes minimal restrictions on relatedness, as benefits can be given randomly and still have positive effect on the population overall. Furthermore, the positive relationship between number of individuals in a herd and experienced predation estimated as a benefit is extremely environmentally dependent, as under conditions limited by resource availability instead of predation, a smaller brood benefits survival.

3: Pfennig et al. [31] argue that tiger salamanders avoid cannibalistic interactions because it increases inclusive fitness. Tiger salamanders will travel long distances to return to the place of their birth to spawn, meaning populations of young with high average relatedness are formed . The larvae are cannibalistic, but have been shown to preferentially consume non-related individuals. Pfennig et al. argue that this is through kin recognition by showing that Hamilton's rule is met as a condition for such behaviour to evolve. After running through and

refuting a list of other potential alternative hypotheses such as disease avoidance, they come to the conclusion of kin selection by deductive reasoning. It is then stated that for kin selection to evolve, Hamilton's rule will be satisfied if $c/b < 1/2$, i.e. if for an interaction between siblings, the survival cost of not eating an available food source is outweighed by the benefit experienced by the sibling not being eaten multiplied by the relationship between the two siblings. However, when applied to a system of salamanders interacting, this again becomes more difficult. The cost of not cannibalising is dependent on the availability of prey including the density of salamander populations themselves, potential cannibalistic interactions are also dependent on size differences between larvae that may represent other fitness advantages. A kin-recognising salamander surrounded by non-kin will have fitness benefits if they are smaller than it, due to more available food, but risks being cannibalised itself if they are larger, making it difficult to break down cost and benefit without additional variables such as size or aggressiveness in feeding. In addition, average relationship between individuals will only maximally be 1/2 (if interactions only take place with siblings); since salamanders meet with kin that are not siblings, or may also seek non-kin to eat, the genetic similarity between interacting salamanders should be lower, but the study does not make this number clear. We see again that when trying to apply causal modelling, population structure and difficulties interpreting cost parameters can limit conclusions given by Hamilton's rule.

4: Krakauer et al. [20] describe 'lekking' observed in wild turkeys, a behaviour in which groups of males will participate in mating dances together. The males in the group are divided into a dominant male which is allowed to mate with the female, and a group of subordinates who cooperate in the lek without chance to reproduce. Males also have the option to dance alone, hence becoming a subordinate removes the chance of mating for the season, corresponding to a fitness cost. Dominant males in a cooperative lek are far more sexually successful than solitary males, representing a fitness benefit. While multiple males can sometimes be present in a cooperative lek, they are typically in pairs, so interactions are relatively simple and can be separated into discrete generations. In addition to this, the groups can be formed while males are still young, and Krakauer et al. observed that males can leave a cooperative relationship, but appear to never join new cooperative relationships, i.e. males are only subordinate to one individual throughout their lifetime. In other observed cooperative lekking systems, choices that appear to be costs for a single mating season give reproductive benefits over one's lifespan: subordinates of a previous mating season receive help in subsequent mating seasons from the new generation of young males. In wild turkeys this appears to not be the case; Krakauer et al. observe that for subordinate males that left coalitions to attempt solitary mating, none were successful. If these observed properties of the system are accurate, this appears to be an excellent system for applying Hamilton's rule. Krakauer et al. combine pedigree relationship with genetic relatedness by using sequencing for genealogical comparisons between individuals on representative loci to show consistency between relationship and genetic relatedness, and demonstrate that males in cooperative leks are significantly more related

than background relatedness.

However, dominance in such cooperative mating rituals are usually determined by some phenotypic difference between individuals. Since Krakauer et al. observe no subordinate males leaving a coalition were successful in mating, it is difficult to know what the expected cost of becoming a subordinate actually is. It is estimated using the expected fitness of all subordinate males, but this may be an overestimate. In addition, statistical tests are taken for a relatively small number of turkeys (7 dominant, 8 subordinate and 14 solitary), meaning some of the observed conclusions about the system may be inaccurate, especially since they do not seem to resemble properties of other cooperative lekking behaviour. Lekking appears to also take place in groups of multiple subordinates, which by the article's own admission would require further assumptions to incorporate, but were not investigated due to limited sample size. Still, this is this may be the best example of Hamilton's rule demonstrated in this list, though small sample size limit the strength of its conclusions.

5: Smith et al. [37] provide a methodology for considering interactions with many partners for the purpose of investigating microbes. *M.xanthus* form fruiting bodies under stress conditions, a small percentage of the individuals in this fruiting body disperse spores and the rest die. Strains of cheaters exist that have a highly increased chance sporulating when among cooperators, but a system of only cheaters sporulates less than cooperators. Since interactions between microbes are often strong and non-linear, Smith et al. look to develop a generalised methodology to model microbial systems.

Colonies of *M.xanthus* formed of cooperator and cheater strains are grown in a lab and put under stress to form fruiting bodies. These fruiting bodies are harvested, and spores analysed. Fitness is measured as sporulation efficiency. Smith et al. reason that any smooth fitness function involving G and G' can be expanded around the point of a non-cooperator in a non-cooperative environment ($G = 0, G' = 0$) and use the assumption genotype of self takes binary values to reduce this expansion to the form:

$$w = \sum_{j=1}^{\infty} b_j G'^j + \sum_{k=0}^{\infty} d_k G G'^j \tag{17}$$

where b_0 is the baseline fitness and d_0 is the cost of cooperation with all interaction partners as non-cooperators.

Smith et al. use ancova to estimate a fitness function for their dataset that also included non-linear terms GG' and G'^2, and then performed a Taylor expansion to represent this fitness function with a 30th order polynomial of the form given by equation 17, which provided accurate predictions of selection on altruistic behaviour. However, though authors claim this is a generalized model, the assumption required for this system to work is that the interaction partners of one's interaction partners are the same as one's own. In *M.xanthus* this is reasonable, since bacteria that form a fruiting body all participate in the act of clustering and may receive similar benefits, but cannot be said of systems in general.

The assumption that the interaction partners of one's interaction partners are the same as the interaction partners of oneself could hypothetically be extended to systems where altruists preferentially interact with other altruists by spatial structure instead of more deterministic mechanism such as kin selection (with the implication that social interactions are strongly constrained by distance and time) and populations are sufficiently dense that the contributions of individuals to the total fitness benefits are negligible. Here it is reasonable to suggest that the benefits received from altruism for two individuals that interact with each other is the same, as they sit in the same interaction neighbourhood and thus should have roughly the same interaction partners G'. In such a system, coefficients of fitness function equation 17 might be interpretable as payoffs of interactions with j individuals. However, for interactions that do not allow for $(G')' = G'$, how to approach the relatedness is less clear.

Smith et al. claim that their extension provides better estimates of empirical results than Hamilton's rule of the form $rb > c$. However, it can be argued that the coefficients of Smith's model that represent higher dimensions of interaction can be absorbed into cost and benefit when changing their interpretation, as seen for interactions with 2 individuals in the additive case seen earlier, and thus again that perceived differences in predictive quality are due to incorrect application. The described system is also additive: *M.xanthus* cheaters have lower fitness than altruists when reproducing normally, so participating in a fruiting event gives an additive effect. In addition, since fruiting only occurs in times of stress, the cost parameter within Smith et al.'s regression model may be higher than the actual experienced cost; if one is likely to die anyway, the cost of forming a fruiting body with the chance to sporulate becomes more attractive for both direct and indirect fitness.

5 Discussion

In summary, attempts to formally apply inclusive fitness to empirical data in an clearly explicable manner remain elusive. We have shown that when working from causal interpretations of cost and benefit, assumptions of weak additivity and equal environments allow approaching something resembling Hamilton's rule. Regression approaches allow for greater generality, but writing c and b as regression coefficients limits the ability to interpret them as benefits and costs or even being able to interpret them as distinct from each other dependant on the system studied. While equation 16 requires some assumptions that may be avoided by regression approaches, it does not require knowledge of all social interactions in order to be applied, whereas regression approaches do; gathering such data may only be possible for simulation models, where one has the opportunity to calculate costs and benefits in a counterfactual way by repeating scenarios while removing altruism with various permutations of altruists and observing fitness differences. Looking at differences between regression and causal interpretations of Hamilton's rule requires further discussion, Okasha and Martens [28] suggest Fisher's 'average effect of a gene substitution' gives potential for causal interpretation, but also conclude this is limited by how Fisher defines 'environmental constancy'; what systems this can be done for is a potential line of inquiry.

In its most general form, Hamilton's rule is true for any system the Price equation holds for. However, attempting to harness this generality to build an approach allowing for clear interpretation of any system of social fitness is proved difficult, even in the case of simulation models. Parameters b, c and R may have complicated relationships between each other, meaning that observing changes to one or two of the three quantities may not be enough to predict the behaviour of the system as a whole [27]. Nevertheless, proponents of Hamilton's rule believe this to be a worthy price to pay for generality [10], and furthermore that extensions can always be made to attempt to mitigate relationships between c, b and R. However, these extensions often entail breaking down the interaction structure into a more realistic interpretation of what may be happening, and experimental examples indicate such extensions should be performed on a case by case basis; under application, generality is lost. Differences in perspective can result in different rules even for simple theoretical models: an extension by Grafen [11] considers phenotype similarly to Queller's derivation (equation 13), but takes events from the perspective of actor rather than recipient (i.e. inclusive rather than neighbour-modulated), resulting in subtle differences to Queller's rule that are equivalent only under assumptions of symmetric interactions [36].

Experimental examples indicate that in many systems, individuals act altruistically with preference towards their own kin, but general methods attempting to quantify fecundity payoffs and interaction structure of these systems in a causal way are still not fully satisfactory. Note that all experimental examples attempt to describe relatively 'ideal' systems where altruism takes the value of 1 or 0 (altruistic or non-altruistic) instead of discrete values or even a continuous scale, and the impact of time or environment are not considered on phenotype.

It may always be possible for the partial regression coefficients of cost and benefit to absorb additional independent variables, such as environmental factors, through correlations to genotype. However, if costs and benefits are represented with functions that include intricate descriptions of interaction structure and experienced environment, how does Hamilton's rule provide more information about a system than Robertson's rule ($\beta_{WG} > 0$)? Birch et al. [4] argue that the parameters of Hamilton's rule are "causally interpretable in a wide range of cases", while Robertson's are not. However, as we have repeatedly seen through derivations, the causal interpretations of Hamilton's rule are still limited by how effectively the relatedness parameter describes interactions taking place. If the regression parameter $\beta_{WG|G'}$ has negligibly more causal interpretation than β_{WG}, Hamilton's rule seemingly provides little additional information to Robertson's rule: investigating causality for complex cases requires a deeper understanding of interaction strength and structure.

References

[1] Benjamin Allen, Martin A. Nowak, and Edward O. Wilson. "Limitations of inclusive fitness". In: *Proceedings of the National Academy of Sciences* 110.50 (Dec. 2013), pp. 20135–20139.

[2] J. Birch and S. Okasha. "Kin selection and its critics". In: *BioScience* 65.1 (2015), pp. 22–32.

[3] Jonathan Birch. "Hamilton's Rule and Its Discontents". In: *The British Journal for the Philosophy of Science* 65.2 (June 2014), pp. 381–411.

[4] Jonathan Birch. "Hamilton's Two Conceptions of Social Fitness". In: *Philosophy of Science* 83.5 (Dec. 2016), pp. 848–860.

[5] Andrew F. G. Bourke. "Hamilton's rule and the causes of social evolution". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 369.1642 (May 2014), p. 20130362.

[6] R. H. Crozier. "Coefficients of Relationship and the Identity of Genes by Descent in the Hymenoptera". In: *The American Naturalist* 104.936 (Mar. 1970), pp. 216–217.

[7] A. Gardner, S. A. West, and G. Wild. "The genetical theory of kin selection: Hamilton's rule still OK". In: *Journal of Evolutionary Biology* 24.5 (May 2011), pp. 1020–1043.

[8] Andy Gardner. "Price's equation made clear". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 375.1797 (Apr. 2020), p. 20190361.

[9] Andy Gardner and Joseph P. Conlon. "Cosmological natural selection and the purpose of the universe". In: *Complexity* 18.5 (May 2013), pp. 48–56.

[10] Andy Gardner, Stuart A. West, and Nicholas H. Barton. "The Relation between Multilocus Population Genetics and Social Evolution Theory." In: *The American Naturalist* 169.2 (Feb. 2007), pp. 207–226.

[11] Alan Grafen. “A geometric view of relatedness”. In: *Oxford Surveys in Evolutionary Biology* 2 (1985), pp. 28–89.

[12] W. D. Hamilton. “Altruism and Related Phenomena, Mainly in Social Insects”. In: *Annual Review of Ecology and Systematics* 3.1 (Nov. 1972), pp. 193–232.

[13] W. D. Hamilton. “The Evolution of Altruistic Behavior”. en. In: *The American Naturalist* 97.896 (Sept. 1963), pp. 354–356.

[14] W.D. Hamilton. “The genetical evolution of social behaviour, parts I and II”. In: *Journal of Theoretical Biology* 7.1 (1964), pp. 1–52.

[15] Rutger Hermsen. *Emergent multilevel selection in a simple spatial model of the evolution of altruism*. preprint. Evolutionary Biology, Dec. 2021.

[16] Katja Hogendoorn and Remko Leys. “The superseded female’s dilemma: ultimate and proximate factors that influence guarding behaviour of the carpenter bee Xylocopa pubescens”. In: *Behavioral Ecology and Sociobiology* 33.6 (Dec. 1993).

[17] Albert Jacquard. *The Genetic Structure of Populations*. OCLC: 858931451. Berlin, Heidelberg: Springer Berlin Heidelberg, 1974.

[18] Jasmeen Kanwal and Andy Gardner. “Population viscosity promotes altruism under density-dependent dispersal”. In: *Proceedings of the Royal Society B: Biological Sciences* 289.1970 (Mar. 2022), p. 20212668.

[19] Thorbjørn Knudsen. “General selection theory and economic evolution: The Price equation and the replicator/interactor distinction”. In: *Journal of Economic Methodology* 11.2 (June 2004), pp. 147–173.

[20] Alan H. Krakauer. “Kin selection and cooperative courtship in wild turkeys”. In: *Nature* 434.7029 (Mar. 2005), pp. 69–72.

[21] Michael L. G. Loeb. “Evolution of Egg Dumping in a Subsocial Insect”. In: *The American Naturalist* 161.1 (Jan. 2003), pp. 129–142.

[22] James A. R. Marshall. *Social Evolution and Inclusive Fitness Theory : An Introduction*. bibtex: marshall2015. Princeton: Princeton University Press, 2015.

[23] Richard E. Michod and W. D. Hamilton. “Coefficients of relatedness in sociobiology”. In: *Nature* 288.5792 (Dec. 1980), pp. 694–697.

[24] C. Mullon and L. Lehmann. “The robustness of the weak selection approximation for the evolution of altruism against strong selection”. In: *Journal of Evolutionary Biology* 27.10 (Oct. 2014), pp. 2272–2282.

[25] Peter Nonacs and Hudson K. Reeve. “The Ecology of Cooperation in Wasps: Causes and Consequences of Alternative Reproductive Decisions”. In: *Ecology* 76.3 (Apr. 1995), pp. 953–967.

[26] Martin A. Nowak, Corina E. Tarnita, and Edward O. Wilson. “The evolution of eusociality”. In: *Nature* 466.7310 (Aug. 2010), pp. 1057–1062.

[27] Martin A. Nowak et al. “The general form of Hamilton’s rule makes no predictions and cannot be tested empirically”. In: *Proceedings of the National Academy of Sciences* 114.22 (May 2017), pp. 5665–5670.

[28] Samir Okasha and Johannes Martens. “The causal meaning of Hamilton’s rule”. In: *Royal Society Open Science* 3.3 (Mar. 2016), p. 160037.

[29] M.J. Orlove and Constance L. Wood. “Coefficients of relationship and coefficients of relatedness in kin selection: A covariance form for the RHO formula”. In: *Journal of Theoretical Biology* 73.4 (Aug. 1978), pp. 679–686.

[30] Judea Pearl. “The Do-Calculus Revisited”. In: (2012).

[31] D. W. Pfennig. “A test of alternative hypotheses for kin recognition in cannibalistic tiger salamanders”. In: *Behavioral Ecology* 10.4 (July 1999), pp. 436–443.

[32] George R. Price. “Selection and Covariance”. In: *Nature* 227.5257 (Aug. 1970), pp. 520–521.

[33] David C. Queller. “A GENERAL MODEL FOR KIN SELECTION”. In: *Evolution; International Journal of Organic Evolution* 46.2 (Apr. 1992), pp. 376–380.

[34] David C. Queller. “Does population viscosity promote kin selection?” In: *Trends in Ecology & Evolution* 7.10 (Oct. 1992), pp. 322–324.

[35] David C. Queller. “Kinship, reciprocity and synergism in the evolution of social behaviour”. In: *Nature* 318.6044 (Nov. 1985), pp. 366–367.

[36] David C. Queller and Keith F. Goodnight. “ESTIMATING RELATEDNESS USING GENETIC MARKERS”. In: *Evolution* 43.2 (Mar. 1989), pp. 258–275.

[37] Jeff Smith, J. David Van Dyken, and Peter C. Zee. “A Generalization of Hamilton’s Rule for the Evolution of Microbial Cooperation”. In: *Science* 328.5986 (June 2010), pp. 1700–1703.

[38] Roland E. Stark. “Cooperative Nesting in the Multivoltine Large Carpenter Bee Xylocopa sulcatipes Maa (Apoidea: Anthophoridae): Do Helpers Gain or Lose to Solitary Females?” In: *Ethology* 91.4 (Apr. 2010), pp. 301–310.

[39] Matthijs van Veelen. “Can Hamilton’s rule be violated?” In: *eLife* 7 (2018), e41901.

[40] Matthijs van Veelen. “Group selection, kin selection, altruism and cooperation: When inclusive fitness is right and when it can be wrong”. In: *Journal of Theoretical Biology* 259.3 (Aug. 2009), pp. 589–600.

[41] Markus Waibel, Dario Floreano, and Laurent Keller. “A Quantitative Test of Hamilton’s Rule for the Evolution of Altruism”. In: *PLoS Biology* 9.5 (May 2011). Ed. by Nick H. Barton, e1000615.

[42] Stuart A. West and Andy Gardner. “Adaptation and Inclusive Fitness”. In: *Current Biology* 23.13 (July 2013), R577–R584.

[43] D S Wilson. “A theory of group selection.” In: *Proceedings of the National Academy of Sciences* 72.1 (Jan. 1975), pp. 143–146.

[44] Sewall Wright. “Coefficients of inbreeding and relationship”. In: *The American Naturalist* 56.645 (July 1922), pp. 330–338.

Appendix

5.1 Forms of Partial Regression

For regressions of a dependent variable y on multiple predictor variables x_j, the sum of squares takes the form:

$$S = \sum_i \left(y_i - E(y) - \sum_j m_j (x_{ij} - E(x_j)) \right)^2$$

where x_{ij} is the ith value of the predictor variable x_j, and m_j are the jth partial regression coefficients of y on x_j with other $x_k \neq x_j$ fixed. Finding the regression coefficients m_k that minimise this sum of squares is done by differentiating with respect to m_k, giving solutions for the partial regression coefficients in the form:

$$m_k = \frac{\mathrm{Cov}(y, x_k)}{\mathrm{Var}(x_k)} - \sum_{j \neq k} m_j \frac{\mathrm{Cov}(x_j, x_k)}{\mathrm{Var}(x_k)}. \tag{18}$$

For the field of inclusive fitness, partial regressions are typically done as regressions of fitness on two predictor variables, genotype of self G and interaction partners G'. To consider regressions with two predictor variables, we rewrite the two regression coefficients using Queller's notation, $m_1 = \beta_{yx_1|x_2}$ and $m_2 = \beta_{yx_2|x_1}$. Simultaneously solving equation 18 for m_1 and m_2, we can write the partial regression $\beta_{yx_1|x_2}$ of y on x_1 while holding x_2 constant:

$$\beta_{yx_1|x_2} = \frac{\beta_{yx_1} - \beta_{yx_2}\beta_{x_2x_1}}{1 - \rho_{12}^2}, \tag{19}$$

where ρ_{12} is the 'correlation coefficient' for x_1, x_2 [7, 22], and β_{ab} is the simple regression of a on b (proof of this can trivially be seen from equation 18 with only one predictor variable):

$$\rho_{12} = \frac{\mathrm{Cov}(x_1, x_2)}{\sqrt{\mathrm{Var}(x_1)\mathrm{Var}(x_2)}}; \qquad \beta_{ab} = \frac{\mathrm{Cov}(a, b)}{\mathrm{Var}(b)}.$$

Least squares regression calculates the partial regression coefficient for one predictor variable while holding all others constant. However, as mentioned in section 2.3.1, another perspective of partial regression coefficient $\beta_{yx_1|x_2}$ is that it predicts y using the parts of x_1 that are not predicted by x_2. We here show the equivalence of these two perspectives.

Consider a regression of x_1 on x_2 with residuals ϵ, and a new variable $\hat{x}_1$ which represents x_1 predicted by x_2:

$$x_1 = a + \beta_{x_1x_2}x_2 + \epsilon$$
$$\hat{x}_1 = a + \beta_{x_1x_2}x_2.$$

We write $x_{1.2} = x_1 - \hat{x}_1$ as x_1 not predicted by x_2. Our alternative perspective of the partial regression coefficient $\beta_{yx_1|x_2}$ suggests that it should take the form of a simple regression between y and our new variable $x_{1.2}$:

$$\beta_{yx_1|x_2} = \frac{\text{Cov}(y, x_{1.2})}{\text{Var}(x_{1.2})} \tag{20}$$

We now look to demonstrate this is equivalent to the form shown in equation 19. First, we use definition of $x_{1.2}$, and properties of variance and covariance:

$$\beta_{yx_1|x_2} = \frac{\text{Cov}(y, x_1 - \hat{x}_1)}{\text{Var}(x_1 - \hat{x}_1)} = \frac{\text{Cov}(y, x_1) - \text{Cov}(y, \hat{x}_1)}{\text{Var}(x_1) + \text{Var}(\hat{x}_1) - 2\text{Cov}(x_1, \hat{x}_1)}$$

Next, we substitute $\hat{x}_1 = a + \beta_{x_1x_2}$, and use the fact a and $\beta_{x_1x_2}$ are constants to simplify:

$$\begin{aligned}
\beta_{yx_1|x_2} &= \frac{\text{Cov}(y, x_1) - \text{Cov}(y, a + \beta_{x_1x_2}x_2)}{\text{Var}(x_1) + \text{Var}(a + \beta_{x_1x_2}x_2) - 2\text{Cov}(x_1, a + \beta_{x_1x_2}x_2)} \\
&= \frac{\text{Cov}(y, x_1) - \beta_{x_1x_2}\text{Cov}(y, x_2)}{\text{Var}(x_1) + \beta_{x_1x_2}^2\text{Var}(x_2) - 2\beta_{x_1x_2}\text{Cov}(x_1, x_2)}.
\end{aligned}$$

Finally, we substitute the function for a simple regression: $\beta_{x_1x_2} = \frac{\text{Cov}(x_1,x_2)}{\text{Var}(x_2)}$ and rearrange:

$$\begin{aligned}
\beta_{yx_1|x_2} &= \frac{\text{Cov}(y, x_1) - \frac{\text{Cov}(x_1,x_2)}{\text{Var}(x_2)}\text{Cov}(y, x_2)}{\text{Var}(x_1) + \left(\frac{\text{Cov}(x_1,x_2)}{\text{Var}(x_2)}\right)^2\text{Var}(x_2) - 2\frac{\text{Cov}(x_1,x_2)}{\text{Var}(x_2)}\text{Cov}(x_1, x_2)} \\
&= \frac{\text{Cov}(y, x_1) - \frac{\text{Cov}(y,x_2)}{\text{Var}(x_2)}\text{Cov}(x_2, x_1)}{\text{Var}(x_1) - \frac{\text{Cov}(x_1,x_2)^2}{\text{Var}(x_2)}} \\
&= \frac{\frac{\text{Cov}(y,x_1)}{\text{Var}(x_1)} - \frac{\text{Cov}(y,x_2)}{\text{Var}(x_2)}\frac{\text{Cov}(x_2,x_1)}{\text{Var}(x_1)}}{1 - \frac{\text{Cov}(x_1,x_2)^2}{\text{Var}(x_1)\text{Var}(x_2)}}
\end{aligned}$$

Substituting back the function for a simple regression gives the same form as equation 19, showing the equivalence between perspectives of the partial regression.