Interpreting selection when individuals interact

Interpreting Selection when Individuals Interact

Jarrod D. Hadfield
Institute of Evolutionary Biology, University of Edinburgh, UK
j.hadfield@ed.ac.uk

Caroline E. Thomson
Department of Zoology, Edward Grey Institute, University of Oxford, Oxford, OX1 3PS, United Kingdom
caroline.thomson@zoo.ox.ac.uk

Summary

1. A useful interpretation of quantitative genetic models of evolutionary change is that they a) define a set of phenotypes that have a causal effect on fitness and on which selection acts, and b) define a set of breeding values that change as a correlated response to that selection because they covary with the phenotypes.

2. When the expression of one trait causes variation in other traits then there are multiple paths by which a trait can cause fitness variation. Because of this there are multiple ways in which selection can be defined, and still be consistent with a causal effect of traits on fitness.

3. We use this result to show that genetical theories of natural/kin selection ignore causation and because of this we suggest they shed little light on the nature of selection.

4. When traits expressed by an individual are affected by traits of their social partners (indirect genetic effects), we suggest a causal partitioning that allows selection to be cast in terms of Hamilton's costs and benefits.

5. We show that previous attempts to understand Hamilton's rule in the context of indirect genetic effects either lack generality, or do not adequately describe all the ways in which an individual's actions constitute a cost to the individual or a benefit to its social partner(s).

6. Our results allow us to explore Hamilton's rule in a multitrait setting. We show that evolution always increases inclusive fitness, and when the traits are measured in units of generalised genetic distance evolutionary change in the traits is in the direction in which inclusive fitness increases the fastest.

7. However, we show that Hamilton's rule only holds in a multitrait context when the suite of traits are at equilibrium. When they are out of equilibrium, the conditions for altruism to evolve may be more or less stringent depending on genetic architecture and how costs and benefits are defined.

Introduction

Kin selection models and the concept of inclusive fitness are important tools for studying the evolution of traits involved in social interactions (Hamilton, 1964a,b). Indirect genetic effect (IGE) models were developed in animal and plant breeding to meet the same need, but prior to, and in isolation from, Hamilton's work (Griffing, 1967; Willham, 1963, 1972). Their key feature is that the trait values of an actor can determine the trait values of a recipient, and therefore affect the recipient's fitness in two ways: directly, or indirectly via their effect on the recipient's own traits (Moore *et al.*, 1997; Wolf *et al.*, 1999). Since these models have been introduced into evolutionary biology there have been

several attempts to relate the parameters of IGE models to the components of inclusive fitness, and therefore Hamilton's rule (Cheverud, 1984; Bijma & Wade, 2008; McGlothlin *et al.*, 2010; Gardner *et al.*, 2011; Hadfield, 2012; McGlothlin *et al.*, 2014).

McGlothlin *et al.* (2014) acknowledged (and added to) the profusion of IGE Hamilton's rules, and concluded that because they are all decompositions of the same evolutionary equation they all offer equally valid perspectives; any differences are merely a matter of semantics. Similar conclusions have been reached by other authors regarding the alternative statistical partitions of total selection that give rise to group-selection and kin-selection approaches (Frank, 1998; Marshall, 2011). However, Okasha (2016) has recently argued that from a causal perspective kin and group selection are not equivalent processes, and that the correct partition separates the causal effects of phenotypes on individual fitness from those on group fitness.

Here we try to understand natural selection in IGE, and other quantitative genetic models, from a causal perspective. Much ground work has already been done in this respect using path-analytic techniques (Arnold, 1983; Conner, 1996; Scheiner *et al.*, 2000; Morrissey, 2014), but to our knowledge it has not been done explicitly in the context of IGEs. From a causal perspective we believe that there is one type of partition that is consistent with Hamilton's idea, or at least most biologist's understanding of it (Okasha & Martens, 2016); the partition should allow the benefit to be the causal effect of the actor's actions on the recipient's fitness and the cost to be the causal effect of the actor's actions on the actor's own fitness (Grafen, 1982). We derive a general method for obtaining such a partition in IGE models and show that the resulting partition will generally differ from those developed earlier (McGlothlin *et al.*, 2014). Maternal effect models are one of the most commonly employed IGE models and several authors have previously sought to understand them in the context of Hamilton's rule (e.g. Cheverud, 1984; Hadfield, 2012). However, the cross-generational nature of

maternal effects greatly complicates their interpretation, leading some to exclude them from the class of models to which their results apply (McGlothlin *et al.*, 2014). We show when and why maternal effect models are hard to understand in terms of cost and benefit, and show that the causal partition we present holds in all instances.

Kin selection and IGE models have usually been constructed for single traits. When thinking about multiple traits, the Lande (1979) equation fundamentally changed the way evolutionary quantitative geneticists think about phenotypic selection and the response to that selection. Characterising selection in terms of partial derivatives placed selection more firmly in the realm of cause and effect (Grafen, 1988; Frank, 1997), and paved the way for the use of multiple regression as an empirical tool that facilitates a greater understanding about the causes of fitness variation from correlational data (Lande & Arnold, 1983). In addition, expressing how the response to this selection is warped by genetic correlations between traits using a compact matrix notation, provided a clear way of understanding and visualising the evolution of multiple traits (Phillips & Arnold, 1989; Schluter, 1996). Although IGE models have often been developed using multivariate notation, when interpreted in the context of Hamilton's rule only single trait (McGlothlin *et al.*, 2010, 2014), or special case two-trait models (Cheverud, 1984; Hadfield, 2012), have been analysed. Here we explore the conditions under which altruism evolves when multiple traits are involved in social interactions, and the consequences this has for inclusive fitness. We find that Hamilton's single trait rule breaks down when there are multiple traits (Cheverud, 1984), much as the breeder's equation does in standard quantitative genetic models (Lande, 1979).

Methods and Results

In this section we present the methods and results together, along with how

they connect with previous theory. Our main result is a multitrait version of Hamilton's rule that incorporates IGEs (Sections 5 & 6), but to obtain it various intermediate results need to be derived and clarified. Given the length and complexity of the section we start with a road map. In section 1) we re-express the Lande Equation in a way that emphasises its causal and correlational components, and in a way that reveals the generality of its basic logic. In section 2) we show that the suite of traits under selection can be transformed without altering the predicted evolutionary response, but only some transforms retain the notion that the traits causally affect fitness. There is not a unique transform that satisfies this because their are multiple ways of defining causality in a multivariate system. In section 3) we introduce the coefficient matrix of a causal system ($\boldsymbol{\Psi}$) and explore its structure in a range of social and non-social settings. In section 4) we show that path analytic approaches to natural selection (Arnold, 1983) use $\boldsymbol{\Psi}$ to define how traits cause fitness variation that is distinct from how it is usually defined (Morrissey, 2014) and use it to reveal the causal logic of genetical theories of selection (Robertson, 1966; Queller, 1992). In section 5) we apply these results to social systems in order to define the cost as the causal effect of the actors behaviour on its own fitness and the benefit as the causal effect of the actors behaviour on the fitness of the recipient. We then compare this to previous IGE definitions of cost and benefit. In section 6) we explore the consequences of moving from a single trait to a multitrait system for Hamilton's rule and the evolution of inclusive fitness. The derivation of the less intuitive results are provided in the Appendix.

1) Evolution as a correlated response in breeding value to selection on phenotype

The Lande (1979) Equation is usually expressed as

$$\Delta\mathbf{a} = \mathbf{G}\boldsymbol{\beta}_{\mathbf{z}} \tag{1}$$

where the between-generation change in breeding values ($\Delta\mathbf{a}$) for a suite of traits ($\mathbf{z}$) is the product of the variance-covariance matrix of breeding values ($\mathbf{G}$) and the selection gradient ($\boldsymbol{\beta}_\mathbf{z}$). Expressing selection through the selection gradient was a major innovation, and connects the theory of selection with the fact that Darwinian explanations are causal (Okasha, 2006): $\boldsymbol{\beta}_\mathbf{z}$ is defined as $E[\partial w/\partial \mathbf{z}]$, the average effect of perturbing a trait on relative fitness (w) whilst holding all other traits constant.

An alternative, and more general, way of expressing this equation is:

$$\Delta\mathbf{a} = \mathrm{COV}(\mathbf{a}, \mathbf{z}^\top)\boldsymbol{\beta}_\mathbf{z} \tag{2}$$

where $\mathrm{COV}(\mathbf{a}, \mathbf{z}^\top)$ is the covariance between the trait breeding values and phenotypes (Kirkpatrick & Lande, 1989; Moore *et al.*, 1997), where $^\top$ denotes matrix transpose. This formulation has three benefits. First, it shows that we can usefully think of the change in breeding value as a correlated response to selection on phenotype. Second, it also makes it clear that the Lande Equation is a special case. Only when inheritance patterns are simple does $\mathrm{COV}(\mathbf{a}, \mathbf{z}^\top) = \mathbf{G}$, and different expressions must be sought when there are additional complications, such as maternal effects (Kirkpatrick & Lande, 1989) or IGEs generally (Moore *et al.*, 1997). Finally, it makes clear that the traits in which we are trying to predict evolutionary change don't necessarily have to be the same traits that define selection: the vector of breeding values ($\mathbf{a}$) don't have to be for the same traits ($\mathbf{z}$) that selection acts upon. For example, Kirkpatrick & Lande (1989) derived a very general maternal effect model (henceforth the K-L model) where $\mathbf{z}$ are the traits of the individual *and* also the individual's mother such that $\mathrm{COV}(\mathbf{a}, \mathbf{z}^\top)$ is not a square matrix (as in neighbourhood models (Nunney, 1985) and the closely related contextual analysis (Heisler & Damuth, 1987; Goodnight *et al.*, 1992)). To emphasise this we will use $\mathbf{a}^{(I)}$ to denote the vector of breeding values for the focal individual for which fitness is defined:

$$\Delta \mathbf{a}^{(I)} = \mathrm{COV}(\mathbf{a}^{(I)}, \mathbf{z}^{\top})\boldsymbol{\beta}_{\mathbf{z}} \tag{3}$$

2) Evolution and selection when trait values are subject to a linear transform

Traits that cause fitness variation are often transformed prior to analysis, and so we first note the rather abstract result that any full-rank linear transformation of the traits that cause fitness variation $\widetilde{\mathbf{z}} = \boldsymbol{\Lambda}\mathbf{z}$ gives identical evolutionary dynamics:

$$\mathrm{COV}(\mathbf{a}^{(I)}, \widetilde{\mathbf{z}}^{\top})\boldsymbol{\beta}_{\widetilde{\mathbf{z}}} = \mathrm{COV}(\mathbf{a}^{(I)}, \mathbf{z}^{\top})\boldsymbol{\Lambda}^{\top}\boldsymbol{\Lambda}^{-\top}\boldsymbol{\beta}_{\mathbf{z}} = \mathrm{COV}(\mathbf{a}^{(I)}, \mathbf{z}^{\top})\boldsymbol{\beta}_{\mathbf{z}} = \Delta \mathbf{a}^{(I)} \tag{4}$$

In the Lande Equation and the K-L model the identity transform is used: $\boldsymbol{\Lambda} = \mathbf{I}$. Other transforms have been used, but then the selection vector ($\boldsymbol{\beta}_{\widetilde{\mathbf{z}}}$) is often hard to interpret in terms of the original traits causing fitness variation. Notable examples of such 'non-causal' transforms are the eigenvectors of $\mathbf{G}$ (Blows *et al.*, 2004) and the non-negative square root of $\mathbf{G}$ (Lande, 1979). Some transforms retain the interpretation of causality and merely change the scale on which the traits are measured: for example when $\boldsymbol{\Lambda}$ is diagonal and contains the reciprocal of the trait means or trait standard deviations (Hansen & Houle, 2008).

However, there are a set of non-diagonal transforms (i.e. those that don't merely change the scale on which the traits are measured) that do retain the interpretation that the traits causally effect fitness, and different transforms reflect different choices about how we partition the causal graph. To understand this, imagine the scenario where trait k affects trait l which affects fitness, so we have the causal graph $k \rightarrow l \rightarrow w^{(I)}$. We could imagine two experiments, one in which we simply perturb k and look at the effect on fitness, and one in which we

look at the effects on fitness if we perturb k but somehow ensure that l remains unperturbed. In the first case we would see an effect on fitness, in the second we would not: k does not affect fitness conditional on l. The question then is which experiment should we envisage when we want to understand selection? In many cases the choice is entirely dependent on the interests of the researcher: both experiments are revealing and interesting. However, in the case of social evolution - for example when trait k in a social partner affects trait l in a focal individual which then affects the focal individual's fitness ($k^{(S)} \rightarrow l^{(I)} \rightarrow w^{(I)}$) - we believe that the first experiment is the one that best captures the notion of benefit in Hamilton's rule: the second experiment would lead to the conclusion that the actions of the social partner can have no benefit for the focal individual. Below, we show how transforms can be constructed which allow us to state which traits should remain constant, and which should be allowed to vary, when we perturb a single trait in an (hypothetical) experiment. These results allow us to generalise our intuition about the simple example introduced above to more complicated situations where there are more traits, and more complex causal relationships between them.

3) Trait determination as an intra- and inter-individual linear system

In what follows we will assume that the set of phenotypes that could have a causal effect on an individual's fitness are an individual's own traits ($\mathbf{z}^{(I)}$) and the traits of its social partners ($\mathbf{z}^{(S)}$) such that:

$$\mathbf{z} = \begin{bmatrix} \mathbf{z}^{(I)} \\ \mathbf{z}^{(S)} \end{bmatrix} \tag{5}$$

We will use the matrix $\mathbf{\Psi}$ to capture the effects of the phenotypes on each other such that $\psi_{i,j}$ is the effect of phenotype j on phenotype i. To allow the notation to accommodate social situations we can partition $\mathbf{\Psi}$ into quadrants representing the effects of the focal individual's traits on its own traits (top left)

the effects of the focal individual's traits on the social partners' traits (bottom left) the effects of the social partners' traits on the focal individual's traits (top right) and the effects of the social partners' traits on social partners' traits (bottom right):

$$\boldsymbol{\Psi} = \begin{bmatrix} \boldsymbol{\Psi}^{(I)} & \boldsymbol{\Psi}^{(I,S)} \\ \boldsymbol{\Psi}^{(S,I)} & \boldsymbol{\Psi}^{(S)} \end{bmatrix} \tag{6}$$

In the first example given above where trait k affects trait l and both are measured in the same individual, there are no social partners so:

$$\boldsymbol{\Psi} = \boldsymbol{\Psi}^{(I)} = \begin{bmatrix} 0 & 0 \\ \psi_{l,k} & 0 \end{bmatrix} \tag{7}$$

Morrissey (2014) considers this scenario and denotes $\boldsymbol{\Psi}^{(I)}$ as $\mathbf{b}$. In the context of a 2-player game where individuals interact symmetrically then:

$$\boldsymbol{\Psi} = \begin{bmatrix} \mathbf{0} & \boldsymbol{\Psi}^{(I,S)} \\ \boldsymbol{\Psi}^{(S,I)} & \mathbf{0} \end{bmatrix} \tag{8}$$

where $\boldsymbol{\Psi}^{(I,S)} = \boldsymbol{\Psi}^{(S,I)}$. Here, $\psi_{l,k}^{(I,S)}$ represents the effect of trait k in the social partner on trait l in the focal individual and $\psi_{l,k}^{(S,I)}$ reflects the effect of trait k in the focal individual on trait l in the social partner. In the indirect genetic effect literature, $\boldsymbol{\Psi}^{(I,S)}$ is often simply denoted as $\boldsymbol{\Psi}$ (Moore *et al.*, 1997).

In the above examples there is either no social partner or one social partner. It might be imagined that in maternal effect models there is only one social partner (the mother) but because the individual's trait values and/or fitness are affected by maternal traits, which in turn may be affected by grandmaternal traits, and so on, there may in fact be an infinite number of social partners. In this instance we will, with some abuse of the notation, use $\boldsymbol{\Psi}^{(I,S)}$ to denote the effect of the mother's traits on her offspring's traits. This matrix is denoted $\mathbf{M}$

in Kirkpatrick & Lande (1989) and $\psi_{l,k}^{(I,S)}$ is the effect of the k^{th} trait in the mother on the l^{th} trait in the offspring. If trait values are ordered by generation, with the individual's (offspring) generation first then the maternal generation, grand-maternal generation, and so on, the K-L model can be represented by the infinite matrix:

$$\boldsymbol{\Psi} = \begin{bmatrix} \mathbf{0} & \boldsymbol{\Psi}^{(I,S)} & \mathbf{0} & \mathbf{0} & \dots \\ \mathbf{0} & \mathbf{0} & \boldsymbol{\Psi}^{(I,S)} & \mathbf{0} & \dots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{\Psi}^{(I,S)} & \dots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{9}$$

Kölliker *et al.* (2005) allow offspring traits to effect maternal traits and denote the matrix $\boldsymbol{\Psi}^{(S,I)}$ as $\mathbf{O}$. This would add a subdiagonal to $\boldsymbol{\Psi}$.

If we denote the vector of trait values $\mathbf{z}$, breeding values $\mathbf{a}$ and environmental values $\mathbf{e}$ for the focal partner followed by its social partners then:

$$\mathbf{z} = \mathbf{a} + \boldsymbol{\Psi}\mathbf{z} + \mathbf{e} \tag{10}$$

This equation can be rearranged (Gianola & Sorensen (2004); see Hadfield *et al.* (2011) for an application to IGE models):

$$(\mathbf{I} - \boldsymbol{\Psi})\mathbf{z} = \mathbf{a} + \mathbf{e} \tag{11}$$

such that we can have $\boldsymbol{\Lambda} = \mathbf{I} - \boldsymbol{\Psi}$ and $\widetilde{\mathbf{z}} = \mathbf{a} + \mathbf{e}$. The matrix $\boldsymbol{\Lambda}$ is sometimes referred to as the Jacobian and can be interpreted in terms of partial derivatives:

$$\boldsymbol{\Lambda} = \frac{\partial \widetilde{\mathbf{z}}}{\partial \mathbf{z}} \tag{12}$$

Consequently,

$$\boldsymbol{\beta}_{\widetilde{\mathbf{z}}} = \boldsymbol{\Lambda}^{-\top}\boldsymbol{\beta}_{\mathbf{z}} = \frac{\partial \mathbf{z}}{\partial \widetilde{\mathbf{z}}}\frac{\partial w}{\partial \mathbf{z}} = \frac{\partial w}{\partial \widetilde{\mathbf{z}}} \tag{13}$$

and we can view the selection gradient $\boldsymbol{\beta}_{\widetilde{\mathbf{z}}}$ as measuring the effect on fitness if we perturb the inputs ($\widetilde{\mathbf{z}}$) into the system. To make the distinction between $\beta_{\mathbf{z}}$ and $\beta_{\widetilde{\mathbf{z}}}$ clear, a hypothetical two-trait system with a single social partner is illustrated in Figure 1. The causal paths by which $\mathbf{z}$ and $\widetilde{\mathbf{z}}$ respectively affect the fitness of the focal individual are highlighted.

Figure 1 here

Deriving the equation for evolutionary change gets a little complicated when the trait values of the individual are correlated with the number of individuals for which they are the social partner. In what follows we will assume that a) the covariance between trait value and group size is constant across generations and b) that if the covariance is non-zero then variation in group size is small. Assuming them to be met, two key relationships emerge:

$$\begin{aligned} \mathrm{COV}(\mathbf{a}^{(I)}, \widetilde{\mathbf{z}}^{\top}) = &\quad \mathrm{COV}(\mathbf{a}^{(I)}, \mathbf{z}^{\top}\boldsymbol{\Lambda}^{\top}) \\ = &\quad (\mathbf{G} \;\; r_1\mathbf{G} \;\; \ldots r_n\mathbf{G}) \end{aligned} \tag{14}$$

where r_m is the relatedness between the individual and the m^{th} of n social partners. This equation tells us that the covariance between breeding values of one individual and the transformed traits of another are equal to $r\mathbf{G}$. The change in phenotype is:

$$\Delta\mathbf{z}^{(I)} = \quad \left(\boldsymbol{\Lambda}^{-1}\Delta\mathbf{a}\right)^{(I)} \tag{15}$$

When focal and social partners belong to the same generation then $\Delta\mathbf{a}^{(I)} = \Delta\mathbf{a}^{(S)}$ and in the examples given above Equation 15 reduces to:

$$\Delta\mathbf{z}^{(I)} = \quad \left(\mathbf{I} - \boldsymbol{\Psi}^{(I)} - \boldsymbol{\Psi}^{(I,S)}\right)^{-1}\Delta\mathbf{a}^{(I)} \tag{16}$$

In maternal effect models $\Delta\mathbf{a}^{(I)}$ and $\Delta\mathbf{a}^{(S)}$ may differ because they refer to different generations. In deriving Equation 16 when social partners belong to different generations we therefore have to also assume c) that there has been a constant force of selection, and as a consequence a constant response to that selection. Note that in maternal effect models assumption a) implies assumption c) because in these models group size (the number of offspring) and fitness are equivalent (Hadfield, 2012).

4) Non-social selection and evolution

In the non-social example - where only the individual's own traits affect each other - the transform $\mathbf{\Lambda} = \mathbf{I} - \mathbf{\Psi}$ results in selection gradients that are equivalent to the path-analytic selection gradients obtained by Arnold (1983). By combining Equations 4, 14 and 16 the change in mean phenotype is:

$$\Delta\mathbf{z}^{(I)} = \quad \mathbf{\Lambda}^{-1}\mathbf{G}\boldsymbol{\beta}_{\tilde{\mathbf{z}}} \tag{17}$$

which was obtained by Morrissey (2014) (where $\mathbf{G}$ was denoted as $\mathbf{G}_\epsilon$).

If we include fitness in the traits under selection such that $\mathbf{z} = \left[w^{(I)}, \mathbf{z}^{\top(I)}, \mathbf{z}^{\top(S)}\right]^\top$ then clearly the first element of $\boldsymbol{\beta}_\mathbf{z}$ is one and the rest are zero. If we explicitly state that there is then no direct path between the original traits and fitness (i.e. the first row of $\mathbf{\Psi}$ is all zeros), then:

$$\begin{aligned} \boldsymbol{\beta}_{\tilde{\mathbf{z}}} = & \quad \mathbf{\Lambda}^{-\top}\boldsymbol{\beta}_\mathbf{z} \\ = & \quad \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{I} - \mathbf{\Psi}_{/w} \end{bmatrix}^{-\top} \begin{bmatrix} 1 \\ \mathbf{0} \end{bmatrix} \\ = & \quad \begin{bmatrix} 1 \\ \mathbf{0} \end{bmatrix} \end{aligned} \tag{18}$$

where $\boldsymbol{\Psi}_{/w}$ is the coefficient matrix for the original traits (i.e. $\mathbf{z}$ excluding fitness). This gives

$$\begin{aligned} \Delta\mathbf{a}^{(I)} &= COV(\mathbf{a}^{(I)}, \widetilde{\mathbf{z}}^{\top})\boldsymbol{\beta}_{\widetilde{\mathbf{z}}} \\ &= COV(\mathbf{a}^{(I)}, w^{(I)}) \end{aligned} \tag{19}$$

which is Robertson's (1966) covariance (the Price (1970) equation applied to breeding values and without transmission bias (Frank, 1997)). This covariance forms the basis of genetical theories of selection (Gardner *et al.*, 2011) but since it can be derived by explicitly stating that the traits have no causal effect on fitness, such theories are perhaps better described as genetical-correlational theories because the breeding values of traits may just happen to be correlated with fitness. Although ugly, we retain the term genetical-correlational so that in the discussion we can distinguish such theories from genetic approaches to measuring selection that are based on the idea of a causal effect.

5) Social selection and evolution

In the presence of social partners we can partition the non-transformed selection gradient into elements associated with the individual's own traits (non-social selection) and elements associated with the social partners' traits (social selection) (Wolf *et al.* (1999); these two types of selection have also been called direct and parental selection respectively; Kirkpatrick & Lande (1989); Hadfield (2012)):

$$\boldsymbol{\beta}_{\mathbf{z}} = \begin{bmatrix} \boldsymbol{\beta}^{(I)} \\ \boldsymbol{\beta}^{(S)} \end{bmatrix} \tag{20}$$

By applying the $\boldsymbol{\Lambda} = \mathbf{I} - \boldsymbol{\Psi}$ transform we get:

$$\boldsymbol{\beta}_{\widetilde{\mathbf{z}}} = \begin{bmatrix} \left(\mathbf{I} - \boldsymbol{\Psi}^{\top(I)} - \boldsymbol{\Psi}^{\top(S,I)}(\mathbf{I} - \boldsymbol{\Psi}^{\top(S)})^{-1}\boldsymbol{\Psi}^{\top(I,S)}\right)^{-1} \left(\boldsymbol{\beta}^{(I)} + \boldsymbol{\Psi}^{\top(S,I)}(\mathbf{I} - \boldsymbol{\Psi}^{\top(S)})^{-1}\boldsymbol{\beta}^{(S)}\right) \\ \left(\mathbf{I} - \boldsymbol{\Psi}^{\top(S)} - \boldsymbol{\Psi}^{\top(I,S)}(\mathbf{I} - \boldsymbol{\Psi}^{\top(I)})^{-1}\boldsymbol{\Psi}^{\top(S,I)}\right)^{-1} \left(\boldsymbol{\beta}^{(S)} + \boldsymbol{\Psi}^{\top(I,S)}(\mathbf{I} - \boldsymbol{\Psi}^{\top(I)})^{-1}\boldsymbol{\beta}^{(I)}\right) \end{bmatrix} \tag{21}$$

The first subvector of $\boldsymbol{\beta}_{\widetilde{z}}$ is the causal effect of an individual's own $\widetilde{\mathbf{z}}$ traits on fitness. To be consistent with a cost (a negative effect on fitness) we will denote this vector as $-\boldsymbol{\beta}_C$. The second subvector is the causal effect of the social partners' $\widetilde{\mathbf{z}}$ traits on the focal individual's fitness and we will denote this $\boldsymbol{\beta}_B$. We propose that $\boldsymbol{\beta}_C$ and $\boldsymbol{\beta}_B$ represent vector-valued costs and benefits according to Hamilton's definition, and the definition by which they are most widely understood.

With one social partner the change in trait breeding values is then:

$$\begin{aligned} \boldsymbol{\Delta}\mathbf{a}^{(I)} = &\ \mathrm{COV}(\mathbf{a}^{(I)}, \widetilde{\mathbf{z}}^{\top})\boldsymbol{\beta}_{\widetilde{\mathbf{z}}} \\ = &\ [\mathbf{G} \quad r\mathbf{G}] \begin{bmatrix} -\boldsymbol{\beta}_C \\ \boldsymbol{\beta}_B \end{bmatrix} \\ = &\ \mathbf{G}\left[r\boldsymbol{\beta}_B - \boldsymbol{\beta}_C\right] \end{aligned} \tag{22}$$

or more generally:

$$\boldsymbol{\Delta}\mathbf{a}^{(I)} = \ \mathbf{G}\left[r_1\boldsymbol{\beta}_{B1} + r_2\boldsymbol{\beta}_{B2} + \ldots r_n\boldsymbol{\beta}_{Bn} - \boldsymbol{\beta}_C\right] \tag{23}$$

where $\boldsymbol{\beta}_{Bm}$ is the subvector of $\boldsymbol{\beta}_B$ relating to the m^{th} (of n) social partners.

Cost/Benefits in maternal effect models.

In the context of maternal effects $r_m = 1/2^m$, because relatedness drops geometrically with lineal ancestry, and Equation 23 has a form similar to that derived in other cross-generational models (Lehmann, 2007). If we assume $\boldsymbol{\beta}_{\mathbf{z}}$ is only non-zero for the traits of the individual and the mother (i.e. there are no *direct* effects of more distant ancestors, such as grandmothers, on the

individual's fitness) and there are no within individual effects of traits on each other (i.e. $\boldsymbol{\Psi}^{(I)} = \mathbf{0}$) then:

$$\boldsymbol{\beta}_C = -\boldsymbol{\beta}_I \tag{24}$$

and

$$\boldsymbol{\beta}_{Bm} = \boldsymbol{\Psi}^{(m-1)\top(I,S)} \left(\boldsymbol{\beta}^{(S)} + \boldsymbol{\Psi}^{\top(I,S)} \boldsymbol{\beta}^{(I)} \right) \tag{25}$$

which represents the m^{th} maternal ancestors effect on the individual's fitness. Depending on the presence of 'cascading' maternal effects (McGlothlin & Galloway (2014); see Figure 2 for a definition) and the pattern of social selection this equation can be simplified (see Figure 2).

Figure 2 here

Under the maternal performance model envisaged in Cheverud (1984) there are two traits; trait 1 is maternal performance and positively affects trait 2 in the offspring, which increases the offspring's fitness. There is no social selection, $\boldsymbol{\beta}^{(S)} = \mathbf{0}$, and there are no cascading maternal effects since $\boldsymbol{\Psi}^{\top(I,S)}$ is a 2-by-2 null matrix except for the entry $\psi_{2,1}$. In the absence of cascading maternal effects $\boldsymbol{\Psi}^{m\top(I,S)}$=$\mathbf{0}$ when $m > 1$, so that

$$\begin{aligned} \boldsymbol{\beta}_{B1} = & \quad \boldsymbol{\Psi}^{\top(I,S)} \boldsymbol{\beta}^{(I)} \\ = & \quad \begin{bmatrix} \psi_{2,1}\beta_{I,2} \\ 0 \end{bmatrix} \end{aligned} \tag{26}$$

and there are no benefits beyond the mother. Cheverud (1984) equated $\beta_{I,2}$ with the benefit but Hadfield (2012) suggested that $\psi_{2,1}\beta_{I,2}$, as given here, is more appropriate as it represents the effect trait 1 in the mother has on her offspring's fitness.

Cheverud (1984) noted that genetic correlations between traits would alter the expected direction of evolutionary change than that implied by Hamilton's rule, and that maternal performance would only increase if:

$$\frac{\beta_{C,1}}{\beta_{B,1}} > \left(\frac{g_{2,1}}{g_{1,1}} + \frac{1}{2}\right) \tag{27}$$

where the benefit and cost of maternal performance are $\beta_{B,1} = \psi_{2,1}\beta_{I,2}$ and $\beta_{C,1} = -\beta_{I,1}$ respectively. Hadfield (2012) incorrectly interpreted the RHS of Equation 27 as a form of relatedness, not realising it was a function of the non-selection terms in Equations 14 and 16 (i.e. $(\mathbf{I} - \boldsymbol{\Psi}^{(I,S)})^{-1}[\mathbf{G} \quad \mathrm{r}\mathbf{G}]$ where $r = 1/2$).

Cost/benefits in a symmetric 2-player game.

McGlothlin *et al.* (2010) simply equated $\boldsymbol{\beta}_I$ with the cost and $\boldsymbol{\beta}_S$ with the benefit of Hamilton's rule. This was criticised by Hadfield (2012) because it fails to include in the benefit the effect a social partner might have on the recipients fitness via their effect on the recipients phenotype. For example, in the context of the Cheverud (1984) model, $\beta_{S,1} = 0$ because there is no *direct* link between parental performance and offspring fitness and so no benefits would be identified. This contrasts with the benefit as given above, which is a function of the non-social selection gradient $\beta_{I,2}$.

More recently McGlothlin *et al.* (2014) derived several alternative definitions of cost and benefit in IGE models, and made the distinction between their original cost and benefit (which they refer to as 'phenotypic'; McGlothlin *et al.*, 2010) and an alternative definition of cost and benefit which they refer to as 'genetic' after Queller (1992). McGlothlin *et al.* (2014) only consider single trait models, but the multitrait equivalent of their two-player symmetric model has $\boldsymbol{\Psi}^{(S,I)} = \boldsymbol{\Psi}^{(I,S)}$ and $\boldsymbol{\Psi}^{(I)} = \boldsymbol{\Psi}^{(S)} = \mathbf{0}$, which gives:

$$\boldsymbol{\beta}_C = -\left(\mathbf{I} - \boldsymbol{\Psi}^{\top(I,S)}\boldsymbol{\Psi}^{\top(I,S)}\right)^{-1}(\boldsymbol{\beta}^{(I)} + \boldsymbol{\Psi}^{\top(I,S)}\boldsymbol{\beta}^{(S)}) \tag{28}$$

$$\boldsymbol{\beta}_B = \left(\mathbf{I} - \boldsymbol{\Psi}^{\top(I,S)}\boldsymbol{\Psi}^{\top(I,S)}\right)^{-1}(\boldsymbol{\beta}^{(S)} + \boldsymbol{\Psi}^{\top(I,S)}\boldsymbol{\beta}^{(I)}) \tag{29}$$

McGlothlin *et al.*'s (2014) 'genetic' selection gradients have the form:

$$\boldsymbol{\beta}_{C_M} = \quad -\left(\mathbf{I} - \boldsymbol{\Psi}^{\top(I,S)}\right)\boldsymbol{\beta}_C \tag{30}$$

and

$$\boldsymbol{\beta}_{B_M} = \quad \left(\mathbf{I} - \boldsymbol{\Psi}^{\top(I,S)}\right)\boldsymbol{\beta}_B \tag{31}$$

We can view our cost and benefit as the change in the actors and recipients fitness if we perturb an individual's $\widetilde{\mathbf{z}}$ trait (or breeding value) by one unit, whereas McGlothlin's (2014) cost and benefit is the change in the actors and recipients fitness if we perturb an individual's total breeding value ($(\mathbf{I} - \boldsymbol{\Psi}^{\top(I,S)})^{-1}\mathbf{a}^{(I)}$; Moore *et al.* (1997)) by one unit.

In Figure 3 we summarise sections 4) and 5) by showing the different assumptions that various models make about the causal effect of traits on fitness.

Figure 3 here

6) Hamilton's rule and the evolution of inclusive fitness

In the single trait case an altruistic trait will increase if (in the single social partner case):

$$\begin{aligned} 0 < &\quad g\left[r\beta_B - \beta_C\right] \\ \beta_C < &\quad r\beta_B \end{aligned} \tag{32}$$

and this is well understood. However, it should be noted that in the general multivariate case this does *not* imply that if the benefit times relatedness exceeds

the cost for a particular trait, the trait will evolve more altruistic values. For example, imagine a trait that has no effect on the bearer's direct fitness but reduces the fitness of its related social partners a little. In the univariate case such a trait would not evolve. However, if this trait was positively genetically correlated with another trait that had no effect on the social partners fitness but increased its bearer's fitness tremendously, then the first trait would increase because of the correlated response to selection the second trait exerts. In Figure 4 we illustrate this idea with another example.

Figure 4 here

Although it is clear that in a multivariate context the evolution of individual traits cannot be understood in terms of Hamilton's rule, it is unclear whether the evolution of the system as a whole can be understood in such terms. Will a more costly *system* evolve if the relative increase in the benefit is greater than relatedness? To obtain an answer, note that the elements of the selection vectors represent the decrease in the fitness of the actor ($\boldsymbol{\beta}_C$) and the increase in the fitness of the recipient ($\boldsymbol{\beta}_B$) if each $\tilde{z}$ trait is increased by one unit. The elements of $\boldsymbol{\Delta}\mathbf{a}$ represent the amount of evolutionary change for each $\tilde{z}$ trait, and so $\boldsymbol{\beta}_C^\top \boldsymbol{\Delta}\mathbf{a}$ is the total decrease in the actors fitness caused by evolutionary change in all traits and $\boldsymbol{\beta}_B^\top \boldsymbol{\Delta}\mathbf{a}$ is the total increase in the recipients fitness. Consequently, to find the conditions for altruism to evolve we need to find the conditions under which both these quantities increase. This can be achieved by having $\boldsymbol{\beta}_B^* = \boldsymbol{\beta}_B^\top \mathbf{G}^{1/2}$ and $\boldsymbol{\beta}_C^* = \boldsymbol{\beta}_C^\top \mathbf{G}^{1/2}$ where $\mathbf{G}^{1/2}$ is the unique non-negative square-root of $\mathbf{G}$ and the new selection vectors are in units of generalized genetic distance (Lande, 1979). The traits will then evolve so that the recipients fitness increases when:

$$||\boldsymbol{\beta}_C^*|| \cdot \cos(\theta) < \quad r \cdot ||\boldsymbol{\beta}_B^*|| \tag{33}$$

where θ is the angle between $\boldsymbol{\beta}_B^*$ and $\boldsymbol{\beta}_C^*$. The LHS is the scalar projection

of the cost vector onto the benefit vector, with both evaluated in units of generalised genetic distances. Likewise, we can obtain the conditions under which the system will evolve to be more costly to actors:

$$||\boldsymbol{\beta}_C^*|| < \quad r \cdot ||\boldsymbol{\beta}_B^*|| \cdot \cos(\theta) \tag{34}$$

where the RHS is the scalar projection of the benefit vector onto the cost vector multiplied by r. If we only consider situations where r is positive, this latter inequality cannot hold if $\cos(\theta) < 0$ and so θ must lie between 270° and 90°. When $\theta = 0$, $\cos(\theta) = 1$ and both inequalities have Hamilton's form:

$$||\boldsymbol{\beta}_C^*|| < \quad r \cdot ||\boldsymbol{\beta}_B^*|| \tag{35}$$

Moreover, when $\theta = 0$ the relative lengths of the two vectors remain the same under a linear transformation, so that the inequality holds even when the vectors are in their original units:

$$||\boldsymbol{\beta}_C|| < \quad r \cdot ||\boldsymbol{\beta}_B|| \tag{36}$$

This makes intuitive sense because when the two vectors are pointing in the same direction ($\theta = 0$) the problem can be recast as a single trait problem, albeit a trait that is some linear combination of the original traits. Although this scenario may seem unlikely, it is worth noting that when there is system-level equilibrium (i.e. $r\boldsymbol{\beta}_B - \boldsymbol{\beta}_C = \mathbf{0}$) the two vectors must point in the same direction.

More generally, $\cos(\theta)$ will lie between 0 and 1 and so for altruism to evolve relatedness must exceed the cost:benefit ratio by more than that in Hamilton's rule if $||\boldsymbol{\beta}_C^*||$ and $||\boldsymbol{\beta}_B^*||$ are equated with the cost and benefit. In Figure 5 and the discussion we explain why this is the case.

Figure 5 here

In Figure 6 we also provide a graphical depiction of the results in terms of the vector projections.

Figure 6 here

The results can be understood by noting that inclusive fitness is always increasing when $\mathbf{G}$ is non-singular and the system is not at equilibrium:

$$\begin{aligned} \Delta \mathrm{IF} &= \left[r\boldsymbol{\beta}_B - \boldsymbol{\beta}_C\right]^\top \Delta \mathbf{a} \\ &= \left[r\boldsymbol{\beta}_B - \boldsymbol{\beta}_C\right]^\top \mathbf{G} \left[r\boldsymbol{\beta}_B - \boldsymbol{\beta}_C\right] \\ &> 0 \end{aligned} \tag{37}$$

Also, a transformation of the traits into units of genetic distance gives:

$$\Delta \mathrm{IF} = \left[r\boldsymbol{\beta}^*_B - \boldsymbol{\beta}^*_C\right]^\top \left[r\boldsymbol{\beta}^*_B - \boldsymbol{\beta}^*_C\right] \tag{38}$$

such that evolution maximises the increase in inclusive fitness per unit of generalised genetic distance. These two results are analogous to results for multivariate evolution in the absence of social interactions (Lande, 1979), although there fitness, rather than inclusive fitness, is maximised. When the system is not at equilibrium inclusive fitness will increase and so the traits evolve in a way in which both the fitness of the actor and recipient may increase.

Discussion

In this paper we give the conditions under which altruism evolves when social interactions involve multiple traits. We show that the evolution of a single trait within a multitrait system cannot be understood in terms of Hamilton's rule (Hamilton, 1964b), but the evolution of the system can be understood in terms of two Hamilton-like inequalities (Inequalities 33 and 34). The derivation involves transforming the selection gradients of quantitative genetics into Hamilton's costs and benefits, and unlike previous transforms (McGlothlin *et al.*, 2014) the

transform we develop also holds in the context of indirect genetic effect models. We acknowledge that a simpler genetical Hamilton's rule (Gardner *et al.*, 2011) can also be used to determine *if* altruism will evolve in such systems, but we suggest that its simplicity means that it cannot be used to understand *why* altruism evolves.

When predicting whether altruism will evolve, the primary differences between our results and those of Hamilton (1964b) are a) two inequalities have to be satisfied rather than one, b) relatedness may have to exceed the cost:benefit ratio by a substantial amount, depending on how vector-valued costs and benefits are summarised as scalars, and c) genetic architecture plays a non-trivial role in determining whether the inequalities are satisfied. We discuss each of these in turn.

Point a) can be dealt with simply as Hamilton's rule actually consists of two rules: the familiar inequality, $rb > c$, but also the implicit condition that b and c are the same sign. Otherwise, $rb > c$ would be satisfied if mutualism rather than altruism evolved: if b was positive but c negative (a benefit to the actor). Our second inequality (Inequality 34) plays the role of ensuring c has the same sign as b, but in a multivariate context. In a single trait analysis the angle between b and c would be 180° if they had different signs, and so inequality 34 could never be satisfied (because $\cos(\theta) = -1$).

In Hamilton's work only a single trait is considered and so the cost and benefit can be represented by scalars. When multiple traits are involved it is most natural to consider the costs and benefits as vector-valued, with a cost and benefit associated with each trait. However, we show that scalar properties of the cost and benefit vectors (their lengths) or scalar comparisons of the cost and benefit vectors (scalar projections) can be used to obtain inequalities similar in form to those derived by Hamilton. For simplicity, we first consider these inequalities in the absence of genetic constraints (the genetic variance is the same in all directions) in order to address point b). Using scalar comparisons

comes closest to Hamilton's simple rule, where $r > c : b$ (from Inequality 33) is the condition under which the traits will evolve to be beneficial to the recipients. Here, $c : b$ designates the cost projected onto the benefit vector divided by the length of the benefit vector, which for single traits is simply c/b. However, we find our results easier to interpret when we associate the absolute costs and benefits with their respective vector lengths. When the the cost and benefit vectors are in the same direction, the combination of traits that increases the recipients fitness is the same combination that decreases the actors fitness. In this situation we can think about this combination as a new composite trait which obeys Hamilton's single trait rule. If the cost and r-weighted benefit vectors have the same length these two forces cancel and the traits will not evolve (Figure 5A), but if the length of the r-weighted benefit vector is increased the traits will evolve in a direction that increases the recipients fitness (Figure 5B). If the two vectors are not in the same direction then the two vectors can never cancel each other out in all directions, and so (some) traits are guaranteed to evolve. Just as Lande's (1979) multitrait generalisation of the breeder's equation showed that trait values will always change in a way that increases mean fitness, we show that, in a social context, traits will always change in a way that increases inclusive fitness. This implies that if the vectors are in different directions inclusive fitness will increase, and if the vectors have the same length then this increase in inclusive fitness will be shared between the actor and the recipient in the ratio $1 : r$ (Figure 5D). Such a situation is not altruistic but mutualistic, because both parties fitness will increase. To shift the ratio so that all of the increase in inclusive fitness falls to the recipients would require the length of the r-weighted benefit vector to exceed that of the cost vector (Figure 5F), potentially by an amount much larger than Hamilton's single trait rule suggests. As the angle between the two vectors increases the potential for evolution to benefit both parties increases, and so the relatedness required for altruism, rather than mutualism, to evolve becomes larger. Once

the angle becomes obtuse, the traits will always evolve to benefit both parties, and altruism cannot evolve (Figure 5C).

Regarding point c) our results also have a close affinity with the Lande (1979) Equation which demonstrated that the evolution of a single trait cannot be understood without understanding the selection that operates on genetically correlated characters. In this sense Hamilton's single trait rule is also known to fail (Cheverud, 1984) in an easily understood way: a character may evolve to harm relatives even when it has no impact on the actor's fitness if the trait is genetically correlated with a character that increases the actor's fitness. However, a possible way to salvage Hamilton's rule in this situation is to argue that the evolution of the second character constitutes a negative cost (a benefit to the actor) and it is this that allows the first character to evolve in a way that constitutes a negative benefit (a cost to the recipient). This argument is identical to that described above where we need to think about the cost and benefit provided by a suite of traits and show that the system as a whole evolves to be more altruistic when rb exceeds c. Above we showed that this argument does not hold even in the absence of genetic constraints when the cost and benefit are associated with vector lengths. However, if we think about the cost:benefit ratio in terms of scalar projections then in the absence of genetic constraints the condition for altruism does appear to be $r > c : b$. However, in the presence of genetic constraints the inequality is actually $r > c^* : b^*$ where the vector elements do not correspond to the original traits, but weighted combinations of traits for which genetic constraints have been removed. Although working in generalised genetic distances allows for a nice compact formula, it should be understood that this compactness comes at the cost of hiding the genetic constraints. In reality, genetic constraints will disrupt the simple relationship $r > c : b$ and for altruism to evolve r may have to be much larger than $c : b$ if there is much less genetic variance in the direction of the benefit vector than the cost vector. Alternatively, the conditions for altruism to evolve may be less

restrictive if the genetic variance in the direction of the benefit vector is greater than that in the direction of the cost vector. The amount of genetic variance in each direction will depend on the exact patterns of genetic (co)variance between the traits. A notable exception to this is when the cost and benefit vector are in the same direction. Then, the genetic variance along each vector has to be the same (they can be thought of as the same composite trait) and $r > c : b$ will hold. At equilibrium the two vectors must be in the same direction and so at equilibrium the inequalities we present collapse to those of Hamilton's rule, irrespective of genetic architecture, and irrespective of how we choose to define or compare costs and benefits.

We obtained the results outlined above by finding a relationship between the selection gradients from quantitative genetics and the costs and benefits in Hamilton's rule. Previous attempts at finding a correspondence have mainly been done in the context of indirect genetic effect (IGE) models (Cheverud, 1984; McGlothlin *et al.*, 2010; Hadfield, 2012; McGlothlin *et al.*, 2014) whereby an individual may affect both the phenotype and the fitness of its social partner (Moore *et al.*, 1997; Wolf *et al.*, 1999). Although several general transforms have been suggested (McGlothlin *et al.*, 2014) our transform differs from those proposed earlier. Our transform is based on a causal description of how a change in an individual's trait value affects the individuals own fitness (cost) and the fitness of its social partners (benefit). In indirect genetic effect models, where multiple individuals affect each others' trait values and fitness, there are multiple ways we can assign cause. Imagine the causal graph $\left\{k^{(S)} \rightarrow w^{(I)};\ \ k^{(S)} \rightarrow l^{(I)} \rightarrow w^{(I)}\right\}$ where trait k in the social partner affects the fitness of the focal individual by two routes; directly, but also indirectly through its affect on trait l of the focal individual. In a non-social context, the multiple regression approach (Lande, 1979; Lande & Arnold, 1983) captures selection on k through its direct effect, whereas the path-analytic approach

(Arnold, 1983; Conner, 1996; Morrissey, 2014) captures selection on k through both paths. In the context of IGE models, both types of causal assignment have been implicitly used, and attempts have been made to relate the resulting selection parameters to Hamilton's cost and benefit (McGlothlin *et al.*, 2014). In this paper we suggest that all of the ways, direct and indirect, in which a social partner can can affect the fitness of a focal individual should be considered as the benefit in Hamilton's rule (Hadfield, 2012). We believe this to be consistent with how Hamilton's costs and benefits are typically interpreted, and also leads us to an inequality that is pleasingly similar to that of Hamilton's. However, we should stress that we are not criticising the utility of previous transforms (McGlothlin *et al.*, 2014) only that they are hard to reconcile with Hamilton's costs and benefits. Indeed, from an empirical perspective the transform presented in McGlothlin *et al.* (2010) is a more tractable way of measuring selection because the fitness of an individual can be regressed on observable traits (z). The $\tilde{z}$ traits we introduced for mathematical convenience are not directly observable and quantifying selection on them not only involves measuring fitness and the observable traits, but how the observable traits influence each others expression.

Although our inequality is similar to Hamilton's it is not identical and this appears to deny the claim that Hamilton's rule has general validity (Gardner *et al.*, 2011). However, it is a genetical Hamilton's rule for which claims of generality have been made rather than a phenotype-based approach we take here. From a causal perspective we show that taking a genetical view is tantamount to assuming that the cause of fitness variation is fitness itself, and that selection is simply viewed as an association between breeding value and fitness irrespective of whether that association is correlational or causal. The genetical view hides complications such as selection on genetically correlated traits and indirect genetic effects (Gardner *et al.*, 2011) and although this results in generality and simplicity it does so, we believe, at the cost of obscuring the underlying biology that is of interest to many biologists, particularly empiri-

cists. Consequently, we echo Okasha's (2016) statement made in the context of kin and multi-level selection that '*ideally we want a description of evolution to provide insight into the causal factors responsible for the evolutionary change in question, in addition to computing the correct answer*'. However, we must stress that the genetical approach that we criticise is one in which the breeding values of a single trait are treated as the object under selection, without consideration of the other traits that may determine the fitness of the actor and/or the recipient. We have called such an approach genetical-correlational to distinguish it from genetic approaches to measuring selection that do attempt to identify causal relationships. For example, when the breeding values of all traits are considered, then the partial derivative/regression coefficients of fitness on the breeding values are identical to those on phenotypes (Rausher, 1992; Queller, 1992) and this multitrait genetic approach (Stinchcombe *et al.*, 2014) can result in the same decomposition as our causal approach. This equivalence is compatible with the idea that genotypes have a causal effect on fitness via phenotypes. The genetical-correlational approach is not compatible with this idea because the genes that determine the focal trait may simply be in linkage disequilbiria with genes that determine another fitness related trait that has been ignored.

In the context of the multitrait genetic approach we invoked the causal relationship, genotype to phenotype to fitness. However, the theory developed here is not in terms of genotypes but breeding values - the genetic aspect of the phenotype which is at the center of most quantitative genetic theories of evolution. The breeding value is not only a function of an individual's genotype, but also the allele frequencies and linkage disequilibria in the population and the other genotypic values that might exist there (Falconer, 1983). It is then hard to imagine that such a function has a causal effect on fitness in any common sense way: the difference in fitness caused by two different genotypes would change depending on the genotypic composition of the population they were in, even in the absence of any intraspecific interactions. However, it should be

remembered that the effect of a perturbation in a non-linear system will depend on its current state, and so when we describe a causal effect in such systems it makes sense to to talk about the average effect of a perturbation. This idea is central to the general definition of a selection gradients as $E[\partial w/\partial z]$ where the average is taken over individuals. In the linear systems discussed in this paper the effect on fitness of perturbing traits is constant over individuals so we simply use the shorthand $\partial w/\partial z$. Fisher (1958) attached a causal meaning to the average effect (of a gene substitution) and although the validity of this interpretation has been questioned (Falconer, 1985), Lee & Chow (2013) show that if the causal effect is averaged in a specific way then we can retain the idea that breeding values represent the average causal effect of alleles on phenotypes (Okasha & Martens, 2016).

This work is theoretical and we have imposed a causal relationship between traits, and between traits and fitness. Inferring causality from correlational data is fraught with well known problems, and we suggest that to understand selection from a causal perspective, more experiments are required (Grafen, 1988; Morrissey, 2014). Although the type of traits that can be experimentally manipulated is limited, there has been a long history of such experiments (e.g. Andersson, 1982) that have not been well integrated into the general literature on natural selection (Kingsolver *et al.*, 2001). In a social context this is exacerbated by the use of incorrect fitness measures which further confound the causal notion of selection with the correlational aspect of inheritance (Grafen, 1982; Wolf & Wade, 2001; Thomson & Hadfield, 2017). We hope this work encourages people to focus on natural and kin selection as causes of fitness variation, and the consequences this has for understanding the evolutionary process.

Acknowledgements

We thank Joel McGlothlin, Bill Hill, Luke McNally, Jacob Moorad, Barbora

Trubenová and Ian White for useful discussions regarding this work. JDH was supported by a Royal Society Fellowship and CET by EPSRC, The Clarendon Fund and Magdalen College, Oxford.

Data Accessibility

There are no data associated with this manuscript.

Author's contributions

JDH conceived the idea and led the theoretical work. JDH and CET wrote the manuscript.

References

Andersson, M. (1982) Female choice selects for extreme tail length in a widowbird. *Nature*, **299**, 818–820.

Arnold, S.J. (1983) Morphology, performance and fitness. *American Zoologist*, **23**, 347–361.

Bijma, P. & Wade, M. (2008) The joint effects of kin, multilevel selection and indirect genetic effects on response to genetic selection. *Journal of evolutionary biology*, **21**, 1175–1188.

Blows, M.W., Chenoweth, S.F. & Hine, E. (2004) Orientation of the genetic variance-covariance matrix and the fitness surface for multiple male sexually selected traits. *American Naturalist*, **163**, 329–340.

Cheverud, J.M. (1984) Evolution by kin selection - a quantitative genetic model illustrated by maternal performance in mice. *Evolution*, **38**, 766–777.

Conner, J.K. (1996) Understanding natural selection: an approach integrat-

ing selection gradients, multiplicative fitness components, and path analysis. *Ethology Ecology & Evolution*, **8**, 387–397.

Falconer, D.S. (1983) *Introduction to Quantitative genetics*. Longman Group.

Falconer, D.S. (1985) A note on Fisher's average effect and average excess. *Genetical Research*, **46**, 337–347.

Fisher, R.A. (1958) *The Genetical Theory of Natural Selection*. Oxford, UK, Oxford University Press, 2nd edition.

Frank, S.A. (1997) The Price Equation, Fisher's fundamental theorem, kin selection, and causal analysis. *Evolution*, **51**, 1712–1729.

Frank, S.A. (1998) *Foundations of Social Evolution*. Princeton University Press, Princeton.

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

Gianola, D. & Sorensen, D. (2004) Quantitative genetic models for describing simultaneous and recursive relationships between phenotypes. *Genetics*, **167**, 1407–1424.

Goodnight, C.J., Schwartz, J.M. & Stevens, L. (1992) Contextual analysis of models of group selection, soft selection, hard selection, and the evolution of altruism. *American Naturalist*, **140**, 743–761.

Grafen, A. (1988) On the uses of data on lifetime reproductive success. T.H. Clutton-Brock, ed., *Reproductive success*, pp. 454–471. Univeristy of Chicago Press, Chicago.

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

Griffing, B. (1967) Selection in reference to biological groups. I. Individual and group selection applied to populations of unordered groups. *Australian Journal of Biological Sciences*, **20**, 127–139.

Hadfield, J.D., Wilson, A.J. & Kruuk, L.E.B. (2011) Cryptic evolution: does environmental deterioration have a genetic basis? *Genetics*, **187**, 1099–1113.

Hadfield, J. (2012) The quantitative genetic theory of parental effects. N.J. Royle, P.T. Smiseth & M. Kölliker, eds., *The Evolution of Parental Care*, pp. 267–284. Oxford University Press, Oxford, UK.

Hamilton, W.D. (1964a) The genetical evolution of social behaviour. I. *Journal of Theoretical Biology*, **7**, 1–16.

Hamilton, W.D. (1964b) The genetical evolution of social behaviour. II. *Journal of Theoretical Biology*, **7**, 17–52.

Hansen, T.F. & Houle, D. (2008) Measuring and comparing evolvability and constraint in multivariate characters. *Journal of Evolutionary Biology*, **21**, 1201–1219.

Heisler, I.L. & Damuth, J. (1987) A method for analyzing selection in hierarchically structured populations. *American Naturalist*, **130**, 582–602.

Kingsolver, J.G., Hoekstra, H.E., Hoekstra, J.M., Berrigan, D., Vignieri, S.N., Hill, C.E., Hoang, A., Gibert, P. & Beerli, P. (2001) The strength of phenotypic selection in natural populations. *American Naturalist*, **157**, 245–261.

Kirkpatrick, M. & Lande, R. (1989) The evolution of maternal characters. *Evolution*, **43**, 485–503.

Kölliker, M., Brodie III, E.D. & Moore, A.J. (2005) The coadaptation of parental supply and offspring demand. *The American Naturalist*, **166**, 506–516.

Lande, R. (1979) Quantitative genetic analysis of multivariate evolution, applied to the brain:body size allometry. *Evolution*, **33**, 402–416.

Lande, R. & Arnold, S.J. (1983) The measurement of selection on correlated characters. *Evolution*, **37**, 1210–1226.

Lee, J.J. & Chow, C.C. (2013) The causal meaning of Fisher's average effect. *Genetics research*, **95**, 89–109.

Lehmann, L. (2007) The evolution of trans-generational altruism: kin selection meets niche construction. *Journal of evolutionary biology*, **20**, 181–189.

Lütkepohl, H. (2005) *New introduction to multiple time series analysis*. Springer Science & Business Media.

Marshall, J.A. (2011) Group selection and kin selection: formally equivalent approaches. *Trends in Ecology & Evolution*, **26**, 325–332.

McGlothlin, J.W., Moore, A.J., Wolf, J.B. & Brodie, E.D. (2010) Interacting phenotypes and the evolutionary process: III. social evolution. *Evolution*, **64**, 2558–2574.

McGlothlin, J.W. & Galloway, L.F. (2014) The contribution of maternal effects to selection response: an empirical test of competing models. *Evolution*, **68**, 549–558.

McGlothlin, J.W., Wolf, J.B., Brodie, E.D. & Moore, A.J. (2014) Quantitative genetic versions of Hamilton's rule with empirical applications. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, **369**, 20130358.

Moore, A.J., Brodie III, E.D. & Wolf, J.B. (1997) Interacting phenotypes and the evolutionary process: I. direct and indirect genetic effects of social interactions. *Evolution*, **51**, 1352–1362.

744 Morrissey, M.B. (2014) Selection and evolution of causally covarying traits. *Evolution*, **68**, 1748–1761.

746 Nunney, L. (1985) Group selection, altruism, and structured-deme models. *The American Naturalist*, **126**, 212–230.

748 Okasha, S. (2006) *Evolution and the levels of selection*. Oxford University Press.

Okasha, S. (2016) The relation between kin and multilevel selection: an ap-
750 proach using causal graphs. *The British Journal for the Philosophy of Science*, **67**, 435–470.

752 Okasha, S. & Martens, J. (2016) The causal meaning of Hamilton's rule. *Open Science*, **3**, 160037.

754 Phillips, P.C. & Arnold, S.J. (1989) Visualizing multivariate selection. *Evolution*, **43**, 1209–1222.

756 Price, G.R. (1970) Selection and covariance. *Nature*, **227**, 520–521.

Queller, D.C. (1992) A general model for kin selection. *Evolution*, **46**, 376–380.

758 Rausher, M.D. (1992) The measurement of selection on quantitative traits - biases due to environmental covariances between traits and fitness. *Evolution*,
760 **46**, 616–626.

Robertson, A. (1966) A mathematical model of culling process in dairy cattle.
762 *Animal Production*, **8**, 95–108.

Scheiner, S., Mitchell, R., Callahan, H. *et al.* (2000) Using path analysis to
764 measure natural selection. *Journal of Evolutionary Biology*, **13**, 423–433.

Schluter, D. (1996) Adaptive radiation along genetic lines of least resistance.
766 *Evolution*, **50**, 1766–1774.

Stinchcombe, J.R., Simonsen, A.K., Blows, M. *et al.* (2014) Estimating uncertainty in multivariate responses to selection. *Evolution*, **68**, 1188–1196.

Thomson, C.E. & Hadfield, J.D. (2017) Measuring selection when parents and offspring interact. *Methods in Ecology & Evolution*, ***submitted***.

Willham, R.L. (1963) The covariance between relatives for characters composed of components contributed by related individuals. *Biometrics*, **19**, 18–27.

Willham, R.L. (1972) The role of maternal effects in animal breeding: III. Biometrical aspects of maternal effects in animals. *Journal of Animal Science*, **35**, 1288–1293.

Wolf, J.B. & Wade, M.J. (2001) On the assignment of fitness to parents and offspring: whose fitness is it and when does it matter? *Journal of Evolutionary Biology*, **14**, 347–356.

Wolf, J.B., Brodie III, E.D. & Moore, A.J. (1999) Interacting phenotypes and the evolutionary process. II. selection resulting from social interactions. *The American Naturalist*, **153**, 254–266.

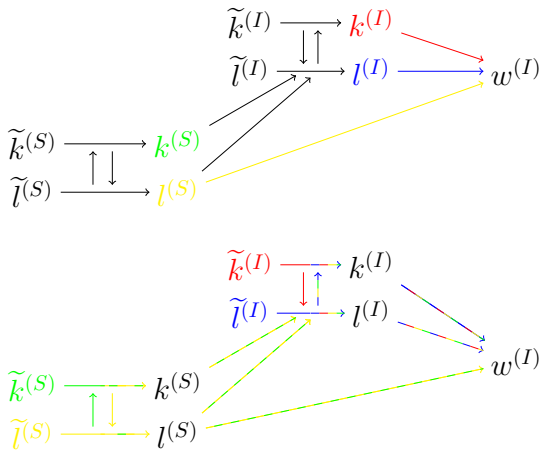

Figure 1: Schematic showing how the fitness and the values of two traits (k and l) in individual I are determined by its own trait values and that of its social partner S. A) Models of evolutionary change such as the Lande Equation and the K-L model define selection as $\partial w/\partial \mathbf{z}$ where the hypothetical experiment would involve perturbing one element of $\mathbf{z}$ holding all other elements constant. The different arrow colours represent the different paths by which each trait affects fitness. Under this scenario $k^{(S)}$ has no causal effect on the focal individual's fitness because there is no *direct* link between $k^{(S)}$ and $w^{(I)}$. B) Alternatively we can think of selection as $\partial w/\partial \widetilde{\mathbf{z}}$. Here $\widetilde{k}^{(S)}$ affects the focal individual's fitness because it affects the expression of $l^{(S)}$ and $l^{(I)}$ (directly) and $k^{(I)}$ (indirectly) all of which affect the focal individual's fitness. The multi-coloured lines represent the fact that multiple traits can have an affect through the same path.

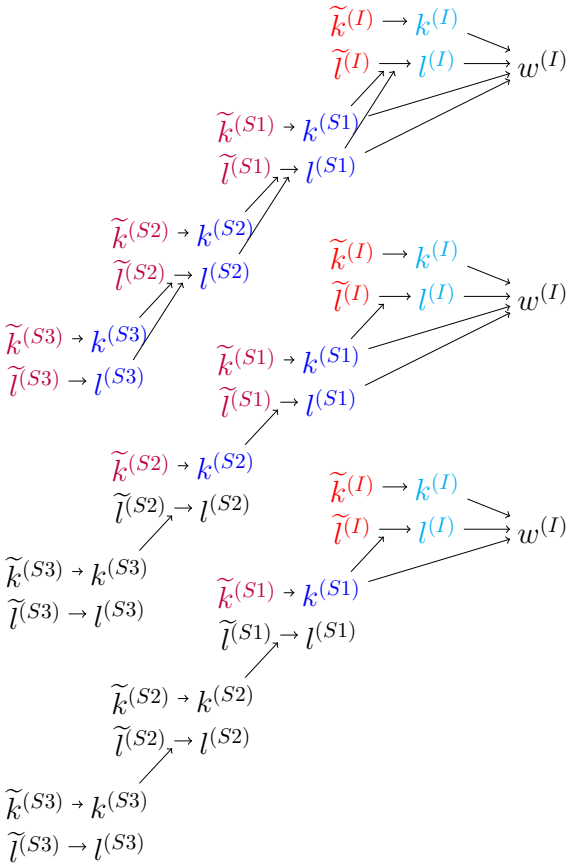

Figure 2: Schematic showing how the fitness and the values of two traits (k and l) in individual I are determined by its own trait values and that of its social partners (its mother $S1$, its grandmother $S2$ and its great grandmother $S3$). In the upper figure, trait l maternally affects itself and so the maternal effects are 'cascading'. With cascading maternal effects, the phenotypes of all maternal ancestors (dark blue+red) affect the traits of the individual (light blue+red) and this can also occur when a trait indirectly affects itself maternally (for example if l maternally affects k and k maternally affects l). In the middle figure there are no cascading maternal effects ($\psi_{l,l}^{(I,S)} = 0$) and only maternal and grandmaternal traits have an impact on the offspring trait values and fitness. The grandmaternal trait has an impact because trait k in the grandmother affects trait l in the mother which affects offspring fitness. In the lower figure there is no direct link between the maternally affected trait (l) and offspring fitness (i.e. no social selection on trait l) and there are no cascading maternal effects. These are the assumptions of Cheverud's (1984) extension of the Willham (1972) model, and there is no causal impact of traits expressed in relatives more distant than the mother on offspring trait values or fitness.

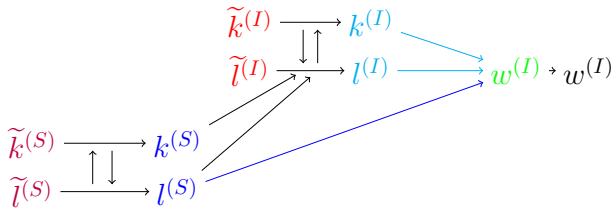

Figure 3: Schematic showing how the fitness and the values of two traits (k and l) in individual I are determined by its own trait values and that of its social partner S. Models of evolutionary change partition the causal graph into a part that causes fitness variation and a part that generates covariances between traits. Different models make different partitions, which are equally valid and merely reflect the researchers interests. The different colours reflect the traits at which different partitions are made under different models; green: Robertson (1966); Price (1970), light blue: Lande (1979), light red: Arnold (1983); Morrissey (2014), light+dark red: Equation 21. Paths downstream of the partition determine selection, and paths upstream determine the trait (co)variances. The partition used by Kirkpatrick & Lande (1989) and McGlothlin *et al.* (2010) differs in that the partition is not defined by a set of traits and is represented by light+dark blue *arrows*; the partition separates the two arrows downstream of $l^{(S)}$.

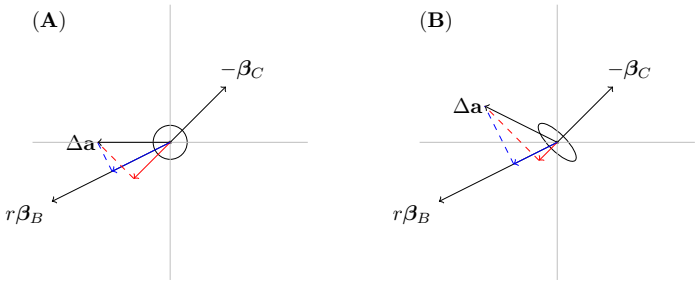

Figure 4: Diagrams depicting the benefit vector to the actor (the negative cost -$\boldsymbol{\beta}_C$) and relatedness-weighted benefit to the recipient ($r\boldsymbol{\beta}_B$), together with the response to selection ($\Delta\mathbf{a}$). The two selection vectors are equal for the trait on the y-axis, but the relatedness-weighted benefit exceeds the cost for the trait on the x-axis. In both cases the system of traits evolves so that the recipients fitness increases at a cost to the actor. This is represented by the projections of the response to selection vector on the r-weighted benefit vector (blue) and the cost vector (red). The blue vector is in the same direction as the benefit vector but the red vector is in the opposite direction to the cost. A) The genetic variances for each trait are equal and there is no genetic correlation ($\mathbf{G}$ is represented by the circle). There is no response to selection on the y-axis because Hamilton's inequality is satisfied ($r\beta_{B,y} = \beta_{C,y}$). B) The genetic variances for each trait are equal but there is a genetic correlation of -0.5 between the traits ($\mathbf{G}$ is represented by the ellipse). The response to selection is deflected towards the direction in trait space with the greatest genetic variance (the major axis of the ellipse) and the trait on the y-axis evolves so that it harms recipients and benefits actors despite $r\beta_{B,y} = \beta_{C,y}$.

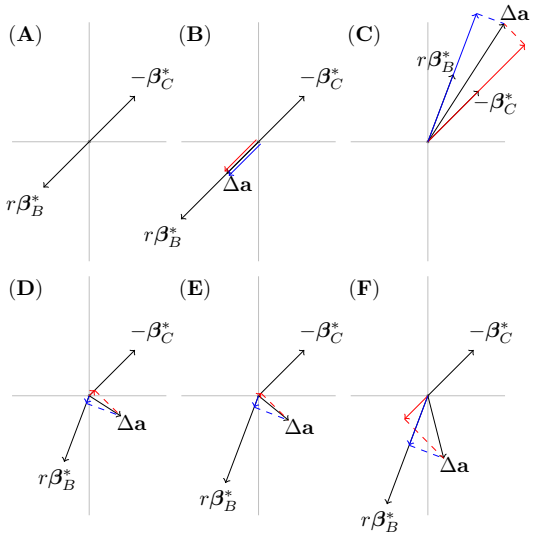

Figure 5: Diagrams depicting the benefit vector to the actor (the negative cost -$\boldsymbol{\beta}_C^*$) and relatedness-weighted benefit to the recipient ($r\boldsymbol{\beta}_B^*$), together with the response to selection ($\Delta\mathbf{a}$). The axes are in generalised genetic distances, or alternatively $\mathbf{G}$ is an identity matrix. The response to selection projected on the benefit vector to the actor and the r-weighted benefit vector to the recipient are in red and blue respectively. When the projections are in the same direction as the selection vectors, evolutionary change increases the fitness of the recipient and the actor respectively. A) The angle between and $\boldsymbol{\beta}_C^*$ and $\boldsymbol{\beta}_B^*$ is $\theta = 0$ and they have the same length $||\boldsymbol{\beta}_C^*|| = ||r\boldsymbol{\beta}_B^*||$. As in Hamilton's rule there is no evolutionary change. B) increasing the benefit and/or relatedness causes evolutionary change in the traits that increases the recipients fitness at a cost to the actor. C) the angle between $\boldsymbol{\beta}_C^*$ and $\boldsymbol{\beta}_B^*$ is 160°. In this case evolutionary change caused by one component of inclusive fitness always moves the traits in a direction that increases inclusive fitness through the other component. Under this scenario it is not possible for the system to evolve so that it benefits recipients at a cost to actors. D) the selection vectors are of the same length but the angle is 25° and lies between 270° and 90°. The two components of inclusive fitness increase equally as the traits evolve such that no party bears a cost. E) Increasing the length of $r\beta_B$ beyond that which is required for Hamilton's *univariate* inequality to be satisfied causes the traits to evolve in a way that preferentially benefits the recipients. However, in this case both parties still benefit although the recipients benefit more than the actors. F) Increasing the length of $r\beta_B$ even more, the traits evolve in a way that further benefits recipients and actually causes a cost to the actors. The Hamilton inequalities for a multivariate system are satisfied: $||\boldsymbol{\beta}_C^*|| \cdot \cos(\theta) < r \cdot ||\boldsymbol{\beta}_B^*||$ and $||\boldsymbol{\beta}_C^*|| < r \cdot ||\boldsymbol{\beta}_B^*|| \cdot \cos(\theta)$.

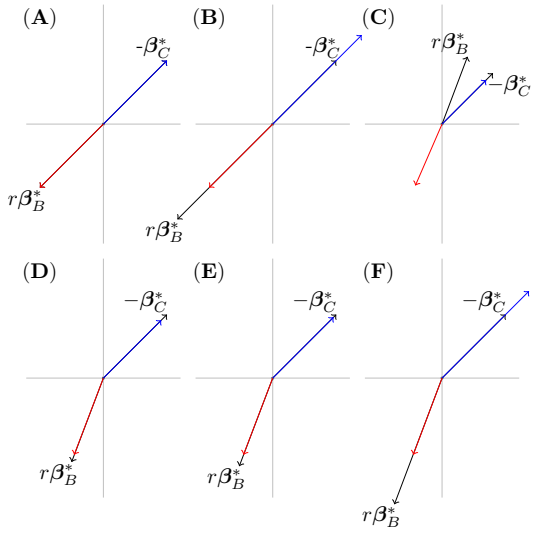

Figure 6: Diagrams depicting the benefit vector to the actor (the negative cost -$\boldsymbol{\beta}^*_C$) and relatedness-weighted benefit to the recipient ($r\boldsymbol{\beta}^*_B$). The axes are in generalised genetic distances, or alternatively $\mathbf{G}$ is an identity matrix. The cost vector projected onto the r-weighted benefit vector ($||\boldsymbol{\beta}^*_C|| \cdot \cos(\theta)$) is in red and the r-weighted benefit vector projected onto the cost vector ($r \cdot ||\boldsymbol{\beta}^*_B|| \cdot \cos(\theta)$) is in blue. When the projected cost is less than the r-weighted benefit the red arrow falls short of $r\boldsymbol{\beta}^*_B$ and inequality 33 is satisfied. When the projected r-weighted benefit is greater than the cost, the blue arrow falls beyond $-\boldsymbol{\beta}^*_C$ and inequality 34 is satisfied. The cost and benefit vectors are those in Figure 5 and panels B) and F) depict a scenario where trait values evolve to be more altruisitic and both inequalities are satisfied: $||\boldsymbol{\beta}^*_C|| \cdot \cos(\theta) < r \cdot ||\boldsymbol{\beta}^*_B||$ and $||\boldsymbol{\beta}^*_C|| < r \cdot ||\boldsymbol{\beta}^*_B|| \cdot \cos(\theta)$.

Appendix

Here we provide the derivation for the less intuitive results. First it will be useful to show that the inverse of $\mathbf{\Lambda}$ can be expressed in three ways. Two are general, with

$$\begin{aligned}\mathbf{\Lambda}^{-1} &= \begin{bmatrix} \mathbf{I}-\mathbf{\Psi}^{(I)} & -\mathbf{\Psi}^{(I,S)} \\ -\mathbf{\Psi}^{(S,I)} & \mathbf{I}-\mathbf{\Psi}^{(S)} \end{bmatrix}^{-1} \\ &= \begin{bmatrix} \mathbf{S}^{-(S)} & \mathbf{S}^{-(S)}\mathbf{\Psi}^{(I,S)}(\mathbf{I}-\mathbf{\Psi}^{(S,S)})^{-1} \\ \mathbf{S}^{-(I)}\mathbf{\Psi}^{(S,I)}(\mathbf{I}-\mathbf{\Psi}^{(I,I)})^{-1} & \mathbf{S}^{-(I)} \end{bmatrix}\end{aligned} \tag{39}$$

and

$$\mathbf{\Lambda}^{-1} = \begin{bmatrix} \mathbf{S}^{-(S)} & (\mathbf{I}-\mathbf{\Psi}^{(I,I)})^{-1}\mathbf{\Psi}^{(I,S)}\mathbf{S}^{-(I)} \\ (\mathbf{I}-\mathbf{\Psi}^{(S,S)})^{-1}\mathbf{\Psi}^{(S,I)}\mathbf{S}^{-(S)} & \mathbf{S}^{-(I)} \end{bmatrix} \tag{40}$$

where

$$\mathbf{S}^{(S)} = \mathbf{I}-\mathbf{\Psi}^{(I)}-\mathbf{\Psi}^{(I,S)}(\mathbf{I}-\mathbf{\Psi}^{(S)})^{-1}\mathbf{\Psi}^{(S,I)} \tag{41}$$

is the Schur complement for $\mathbf{\Lambda}^{(S)}$ and

$$\mathbf{S}^{(I)} = \mathbf{I}-\mathbf{\Psi}^{(S)}-\mathbf{\Psi}^{(S,I)}(\mathbf{I}-\mathbf{\Psi}^{(I)})^{-1}\mathbf{\Psi}^{(I,S)} \tag{42}$$

is the Schur complement for $\mathbf{\Lambda}^{(I)}$. The final way is specific to the maternal effect model, since $\mathbf{\Psi}$ (Equation 9) has a 1st order vector autoregressive form (Lütkepohl, 2005) so $\mathbf{\Lambda}$ has inverse

$$\boldsymbol{\Lambda}^{-1} = \begin{bmatrix} \mathbf{I} & \boldsymbol{\Psi}^{(I,S)} & \boldsymbol{\Psi}^{2(I,S)} & \boldsymbol{\Psi}^{3(I,S)} & \dots & \boldsymbol{\Psi}^{n(I,S)} \\ \mathbf{0} & \mathbf{I} & \boldsymbol{\Psi}^{(I,S)} & \boldsymbol{\Psi}^{2(I,S)} & \dots & \boldsymbol{\Psi}^{n(I,S)} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} & \boldsymbol{\Psi}^{(I,S)} & \dots & \boldsymbol{\Psi}^{(n-1)(I,S)} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I} & \dots & \boldsymbol{\Psi}^{(n-2)(I,S)} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{I} \end{bmatrix} \tag{43}$$

We need to show that in the three examples given, the change in trait means given in Equation 15 reduces to that in equation 16 when it is assumed that $\Delta\mathbf{a}^{(I)} = \Delta\mathbf{a}^{(S)}$. Using the inverse in Equation 39, the change in trait means is obtained as:

$$\begin{aligned} \Delta\mathbf{z}^{(I)} = & \left(\boldsymbol{\Lambda}^{-1}\Delta\mathbf{a}\right)^{(I)} \\ & \mathbf{S}^{-(S)}\Delta\mathbf{a}^{(I)} + \mathbf{S}^{-(S)}\boldsymbol{\Psi}^{(I,S)}(\mathbf{I} - \boldsymbol{\Psi}^{(S,S)})^{-1}\Delta\mathbf{a}^{(S)} \end{aligned} \tag{44}$$

so that when $\Delta\mathbf{a}^{(I)} = \Delta\mathbf{a}^{(S)}$:

$$\begin{aligned} \Delta\mathbf{z}^{(I)} = & \mathbf{S}^{-(S)}\Delta\mathbf{a}^{(I)} + \mathbf{S}^{-(S)}\boldsymbol{\Psi}^{(I,S)}(\mathbf{I} - \boldsymbol{\Psi}^{(S,S)})^{-1}\Delta\mathbf{a}^{(I)} \\ = & \mathbf{S}^{-(S)}\left(\mathbf{I} + \boldsymbol{\Psi}^{(I,S)}(\mathbf{I} - \boldsymbol{\Psi}^{(S,S)})^{-1}\right)\Delta\mathbf{a}^{(I)} \end{aligned} \tag{45}$$

In the non-social example, $\boldsymbol{\Psi}^{(I,I)}$ is non zero and there are no social partners, hence

$$\begin{aligned} \Delta\mathbf{z}^{(I)} = & \mathbf{S}^{-(S)}\Delta\mathbf{a}^{(I)} \\ = & (\mathbf{I} - \boldsymbol{\Psi}^{(I)})^{-1}\Delta\mathbf{a}^{(I)} \end{aligned} \tag{46}$$

consistent with Equation 16. In the symmetric 2-player game, $\boldsymbol{\Psi}^{(I,I)} = \boldsymbol{\Psi}^{(S,S)} = \mathbf{0}$ and $\boldsymbol{\Psi}^{(S,I)} = \boldsymbol{\Psi}^{(I,S)}$ and so

$$\begin{aligned} \Delta\mathbf{z}^{(I)} = & \mathbf{S}^{-(S)}(\mathbf{I} + \boldsymbol{\Psi}^{(I,S)})\Delta\mathbf{a}^{(I)} \\ & \left(\mathbf{I} - \boldsymbol{\Psi}^{(I,S)}\boldsymbol{\Psi}^{(S,I)}\right)^{-1}(\mathbf{I} + \boldsymbol{\Psi}^{(I,S)})\Delta\mathbf{a}^{(I)} \\ & (\mathbf{I} - \boldsymbol{\Psi}^{(I,S)})^{-1}\Delta\mathbf{a}^{(I)} \end{aligned} \tag{47}$$

again, consistent with Equation 16. The above holds because,

$$\begin{aligned}
(\mathbf{I}-\boldsymbol{\Psi}^{(I,S)})^{-1} &= \left(\mathbf{I}-\boldsymbol{\Psi}^{(I,S)}\boldsymbol{\Psi}^{(S,I)}\right)^{-1}(\mathbf{I}+\boldsymbol{\Psi}^{(I,S)}) \\
(\mathbf{I}-\boldsymbol{\Psi}^{(I,S)})^{-1}(\mathbf{I}+\boldsymbol{\Psi}^{(I,S)})^{-1} &= \left(\mathbf{I}-\boldsymbol{\Psi}^{(I,S)}\boldsymbol{\Psi}^{(S,I)}\right)^{-1} \\
\left((\mathbf{I}+\boldsymbol{\Psi}^{(I,S)})(\mathbf{I}-\boldsymbol{\Psi}^{(I,S)})\right)^{-1} &= \left(\mathbf{I}-\boldsymbol{\Psi}^{(I,S)}\boldsymbol{\Psi}^{(S,I)}\right)^{-1} \\
\left(\mathbf{I}+\boldsymbol{\Psi}^{(I,S)}-\boldsymbol{\Psi}^{(I,S)}-\boldsymbol{\Psi}^{(I,S)}\boldsymbol{\Psi}^{(I,S)}\right)^{-1} &= \left(\mathbf{I}-\boldsymbol{\Psi}^{(I,S)}\boldsymbol{\Psi}^{(S,I)}\right)^{-1} \\
\left(\mathbf{I}-\boldsymbol{\Psi}^{(I,S)}\boldsymbol{\Psi}^{(I,S)}\right)^{-1} &= \left(\mathbf{I}-\boldsymbol{\Psi}^{(I,S)}\boldsymbol{\Psi}^{(S,I)}\right)^{-1}
\end{aligned} \tag{48}$$

when $\boldsymbol{\Psi}^{(I,S)} = \boldsymbol{\Psi}^{(S,I)}$. In the final, maternal effect case, it is easier to derive Equation 16 using the inverse form in Equation 43. Assuming that evolutionary change in all generations has been equal to $\Delta\mathbf{a}^{(I)}$ then:

$$\begin{aligned}
\Delta\mathbf{z}^{(I)} &= \left(\boldsymbol{\Lambda}^{-1}\Delta\mathbf{a}\right)^{(I)} \\
&= \textstyle\sum_{m=0}^{n=\infty}\boldsymbol{\Psi}^{m(I,S)}\Delta\mathbf{a}^{(I)} \\
&= (\mathbf{I}-\boldsymbol{\Psi}^{(I,S)})^{-1}\Delta\mathbf{a}^{(I)}
\end{aligned} \tag{49}$$

consistent with Equation 16. The final line is obtained since we are taking the infinite sum of a geometric series.

The derivation of cost and benefit vectors in Equation 21 can most easily be obtained using the inverse of $\boldsymbol{\Lambda}$ in the form presented in Equation 40:

$$\begin{aligned}
\boldsymbol{\beta}_{\widetilde{\mathbf{z}}} &= \boldsymbol{\Lambda}^{-\top}\boldsymbol{\beta}_z \\
&= \begin{bmatrix} \mathbf{S}^{-(S)} & (\mathbf{I}-\boldsymbol{\Psi}^{(I,I)})^{-1}\boldsymbol{\Psi}^{(I,S)}\mathbf{S}^{-(I)} \\ (\mathbf{I}-\boldsymbol{\Psi}^{(S,S)})^{-1}\boldsymbol{\Psi}^{(S,I)}\mathbf{S}^{-(S)} & \mathbf{S}^{-(I)} \end{bmatrix}^{\top}\begin{bmatrix} \boldsymbol{\beta}^{(I)} \\ \boldsymbol{\beta}^{(S)} \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{S}^{-\top(S)} & \mathbf{S}^{-\top(S)}\boldsymbol{\Psi}^{\top(S,I)}(\mathbf{I}-\boldsymbol{\Psi}^{\top(S,S)})^{-1} \\ \mathbf{S}^{-\top(I)}\boldsymbol{\Psi}^{\top(I,S)}(\mathbf{I}-\boldsymbol{\Psi}^{\top(I,I)})^{-1} & \mathbf{S}^{-\top(I)} \end{bmatrix}\begin{bmatrix} \boldsymbol{\beta}^{(I)} \\ \boldsymbol{\beta}^{(S)} \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{S}^{-\top(S)}\left(\boldsymbol{\beta}^{(I)}+\boldsymbol{\Psi}^{\top(S,I)}(\mathbf{I}-\boldsymbol{\Psi}^{\top(S,S)})^{-1}\boldsymbol{\beta}^{(S)}\right) \\ \mathbf{S}^{-\top(I)}\left(\boldsymbol{\beta}^{(S)}+\boldsymbol{\Psi}^{\top(I,S)}(\mathbf{I}-\boldsymbol{\Psi}^{\top(I,I)})^{-1}\boldsymbol{\beta}^{(I)}\right) \end{bmatrix}
\end{aligned} \tag{50}$$

where expansion of the Schur complements gives Equation 21. In the symmetric two-player game this simplifies to Equations 51 and 52:

$$\boldsymbol{\beta}_C = -\left(\mathbf{I} - \boldsymbol{\Psi}^{\top(I,S)}\boldsymbol{\Psi}^{\top(I,S)}\right)^{-1}(\boldsymbol{\beta}^{(I)} + \boldsymbol{\Psi}^{\top(I,S)}\boldsymbol{\beta}^{(S)}) \tag{51}$$

$$\boldsymbol{\beta}_B = \left(\mathbf{I} - \boldsymbol{\Psi}^{\top(I,S)}\boldsymbol{\Psi}^{\top(I,S)}\right)^{-1}(\boldsymbol{\beta}^{(S)} + \boldsymbol{\Psi}^{\top(I,S)}\boldsymbol{\beta}^{(I)}) \tag{52}$$

McGlothlin's (2014) selection gradients are only given in univariate form without derivation, but we take the multivariate form to be:

$$\begin{aligned} \boldsymbol{\beta}_{C_M} &= (\mathbf{I} + \boldsymbol{\Psi}^{\top(I,S)})^{-1}(\boldsymbol{\beta}^{(I)} + \boldsymbol{\Psi}^{\top(I,S)}\boldsymbol{\beta}^{(S)}) \\ &= -(\mathbf{I} + \boldsymbol{\Psi}^{\top(I,S)})^{-1}\left(\mathbf{I} - \boldsymbol{\Psi}^{\top(I,S)}\boldsymbol{\Psi}^{\top(I,S)}\right)\boldsymbol{\beta}_C \\ &= -(\mathbf{I} - \boldsymbol{\Psi}^{\top(I,S)})\boldsymbol{\beta}_C \end{aligned} \tag{53}$$

$$\begin{aligned} \boldsymbol{\beta}_{B_M} &= (\mathbf{I} + \boldsymbol{\Psi}^{\top(I,S)})^{-1}(\boldsymbol{\beta}^{(S)} + \boldsymbol{\Psi}^{\top(I,S)}\boldsymbol{\beta}^{(I)}) \\ &= (\mathbf{I} + \boldsymbol{\Psi}^{\top(I,S)})^{-1}\left(\mathbf{I} - \boldsymbol{\Psi}^{\top(I,S)}\boldsymbol{\Psi}^{\top(I,S)}\right)\boldsymbol{\beta}_B \\ &= (\mathbf{I} - \boldsymbol{\Psi}^{\top(I,S)})\boldsymbol{\beta}_B \end{aligned} \tag{54}$$

where in each case the final line can be obtained by taking the inverse of both sides of Equation 48 to show:

$$(\mathbf{I} - \boldsymbol{\Psi}^{(I,S)}) = (\mathbf{I} + \boldsymbol{\Psi}^{(I,S)})^{-1}\left(\mathbf{I} - \boldsymbol{\Psi}^{(I,S)}\boldsymbol{\Psi}^{(S,I)}\right) \tag{55}$$

In the maternal effect model the inverse of $\boldsymbol{\Lambda}$ in the form presented in Equation 43 allows a simpler derivation:

$$
\begin{aligned}
\boldsymbol{\beta}_{\tilde{\mathbf{z}}} = \quad & \boldsymbol{\Lambda}^{-\top}\boldsymbol{\beta}_z \\
= \quad & \begin{bmatrix}
\mathbf{I} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\
\boldsymbol{\Psi}^{\top(I,S)} & \mathbf{I} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\
\boldsymbol{\Psi}^{2\top(I,S)} & \boldsymbol{\Psi}^{\top(I,S)} & \mathbf{I} & \mathbf{0} & \dots & \mathbf{0} \\
\boldsymbol{\Psi}^{3\top(I,S)} & \boldsymbol{\Psi}^{2\top(I,S)} & \boldsymbol{\Psi}^{\top(I,S)} & \mathbf{I} & \dots & \mathbf{0} \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
\boldsymbol{\Psi}^{n\top(I,S)} & \boldsymbol{\Psi}^{(n-1)\top(I,S)} & \boldsymbol{\Psi}^{(n-2)\top(I,S)} & \boldsymbol{\Psi}^{(n-3)\top(I,S)} & \dots & \mathbf{I}
\end{bmatrix}
\begin{bmatrix}
\boldsymbol{\beta}^{(I)} \\ \boldsymbol{\beta}^{(S)} \\ \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{0}
\end{bmatrix} \\
= \quad & \begin{bmatrix}
\boldsymbol{\beta}^{(I)} \\
\boldsymbol{\Psi}^{\top(I,S)}\boldsymbol{\beta}^{(I)} + \boldsymbol{\beta}^{(S)} \\
\boldsymbol{\Psi}^{2\top(I,S)}\boldsymbol{\beta}^{(I)} + \boldsymbol{\Psi}^{\top(I,S)}\boldsymbol{\beta}^{(S)} \\
\boldsymbol{\Psi}^{3\top(I,S)}\boldsymbol{\beta}^{(I)} + \boldsymbol{\Psi}^{2\top(I,S)}\boldsymbol{\beta}^{(S)} \\
\vdots \\
\boldsymbol{\Psi}^{n\top(I,S)}\boldsymbol{\beta}^{(I)} + \boldsymbol{\Psi}^{(n-1)\top(I,S)}\boldsymbol{\beta}^{(S)}
\end{bmatrix}
\end{aligned}
\tag{56}
$$

which gives Equations 24 and 25.

Although not discussed in the main manuscript, here we consider an alternative way to partition the causal graph in maternal effect models where only downstream paths from the mother are considered as having a causal effect on offspring fitness. To achieve this we use the transform:

$$
\boldsymbol{\Lambda} = \begin{bmatrix} \mathbf{I} & -\boldsymbol{\Psi}^{(I,S)} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \qquad \boldsymbol{\Lambda}^{-1} = \begin{bmatrix} \mathbf{I} & \boldsymbol{\Psi}^{(I,S)} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \tag{57}
$$

However, it is important to realise that because the transform does not capture the complete causal model defined by Equation 10 then Equation 14 does not hold. However, Kirkpatrick & Lande (1989) derived $\mathrm{COV}(\mathbf{a}^{(I)}, \mathbf{z}^{\top})$:

$$\mathrm{COV}(\mathbf{a}^{(I)}, \mathbf{z}^{\top}) = \left[\mathbf{G} \left(\mathbf{I} - \frac{1}{2} \boldsymbol{\Psi}^{\top(I,S)} \right)^{-1} \quad \frac{1}{2} \mathbf{G} \left(\mathbf{I} - \frac{1}{2} \boldsymbol{\Psi}^{\top(I,S)} \right)^{-1} \right] \tag{58}$$

which gives

$$\begin{aligned} \mathrm{COV}(\mathbf{a}, \mathbf{z}^{\top}) \boldsymbol{\Lambda}^{\top} = & \left[\mathbf{G} \left(\mathbf{I} - \tfrac{1}{2} \boldsymbol{\Psi}^{\top(I,S)} \right)^{-1} \quad \tfrac{1}{2} \mathbf{G} \left(\mathbf{I} - \tfrac{1}{2} \boldsymbol{\Psi}^{\top(I,S)} \right)^{-1} \right] \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\boldsymbol{\Psi}^{\top(I,S)} & \mathbf{I} \end{bmatrix} \\ = & \left[\mathbf{G} \left(\mathbf{I} - \tfrac{1}{2} \boldsymbol{\Psi}^{\top(I,S)} \right)^{-1} - \tfrac{1}{2} \mathbf{G} \left(\mathbf{I} - \tfrac{1}{2} \boldsymbol{\Psi}^{\top(I,S)} \right)^{-1} \boldsymbol{\Psi}^{\top(I,S)} \quad \tfrac{1}{2} \mathbf{G} \left(\mathbf{I} - \tfrac{1}{2} \boldsymbol{\Psi}^{\top(I,S)} \right)^{-1} \right] \\ = & \left[\mathbf{G} \left(\mathbf{I} - \tfrac{1}{2} \boldsymbol{\Psi}^{\top(I,S)} \right)^{-1} \left(\mathbf{I} - \tfrac{1}{2} \boldsymbol{\Psi}^{\top(I,S)} \right) \quad \tfrac{1}{2} \mathbf{G} \left(\mathbf{I} - \tfrac{1}{2} \boldsymbol{\Psi}^{\top(I,S)} \right)^{-1} \right] \\ = & \left[\mathbf{G} \quad \tfrac{1}{2} \mathbf{G} \left(\mathbf{I} - \tfrac{1}{2} \boldsymbol{\Psi}^{\top(I,S)} \right)^{-1} \right] \end{aligned} \tag{59}$$

such that the phenotypic effects of more distant maternal ancestors are considered as responsible for building up a (non-standard) covariance between the breeding values of the focal individual and the maternal phenotypes (the right hand partition of the above matrix). Under this scenario,

$$\begin{aligned} \boldsymbol{\beta}_{\tilde{\mathbf{z}}} = & \boldsymbol{\Lambda}^{-\top} \boldsymbol{\beta}_{\mathbf{z}} \\ = & \begin{bmatrix} \boldsymbol{\beta}^{(I)} \\ \boldsymbol{\beta}^{(S)} + \boldsymbol{\Psi}^{\top(I,S)} \boldsymbol{\beta}^{(I)} \end{bmatrix} \end{aligned} \tag{60}$$

As an independent check,

$$\begin{aligned} \boldsymbol{\Delta}\mathbf{a}^{(I)} = & \mathbf{G}\boldsymbol{\beta}_C + \tfrac{1}{2} \mathbf{G} \left(\mathbf{I} - \tfrac{1}{2} \boldsymbol{\Psi}^{\top(I,S)} \right)^{-1} \boldsymbol{\beta}_B \\ = & \mathbf{G}\boldsymbol{\beta}_I + \tfrac{1}{2} \mathbf{G} \left(\mathbf{I} - \tfrac{1}{2} \boldsymbol{\Psi}^{\top(I,S)} \right)^{-1} \boldsymbol{\Psi}^{\top(I,S)} \boldsymbol{\beta}^{(I)} + \tfrac{1}{2} \mathbf{G} \left(\mathbf{I} - \tfrac{1}{2} \boldsymbol{\Psi}^{\top(I,S)} \right)^{-1} \boldsymbol{\beta}^{(S)} \\ = & \tfrac{1}{2} \mathbf{G} \left(\mathbf{I} - \tfrac{1}{2} \boldsymbol{\Psi}^{\top(I,S)} \right)^{-1} \left(2 \left(\mathbf{I} - \tfrac{1}{2} \boldsymbol{\Psi}^{\top(I,S)} \right) + \boldsymbol{\Psi}^{\top(I,S)} \right) \boldsymbol{\beta}^{(I)} + \tfrac{1}{2} \mathbf{G} \left(\mathbf{I} - \tfrac{1}{2} \boldsymbol{\Psi}^{\top(I,S)} \right)^{-1} \boldsymbol{\beta}^{(S)} \\ = & \mathbf{G} \left(\mathbf{I} - \tfrac{1}{2} \boldsymbol{\Psi}^{\top(I,S)} \right)^{-1} \boldsymbol{\beta}^{(I)} + \tfrac{1}{2} \mathbf{G} \left(\mathbf{I} - \tfrac{1}{2} \boldsymbol{\Psi}^{\top(I,S)} \right)^{-1} \boldsymbol{\beta}^{(S)} \end{aligned} \tag{61}$$

as given in Kirkpatrick & Lande (1989).

In Equation 33 we derive the conditions under which a system of traits will evolve so that they benefit recipients. The derivation makes use of the property

$\mathbf{a}^\top \mathbf{b} = \cos(\theta) \cdot ||\mathbf{a}|| \cdot |\mathbf{b}||$ where $||\mathbf{a}||$ is the length of $\mathbf{a}$ and θ is the angle between $\mathbf{a}$ and $\mathbf{b}$:

$$\begin{aligned}
0 < & \quad \boldsymbol{\beta}_B^\top \Delta \mathbf{a} \\
0 < & \quad \boldsymbol{\beta}_B^\top \mathbf{G} \left[r\boldsymbol{\beta}_B - \boldsymbol{\beta}_C \right] \\
0 < & \quad \boldsymbol{\beta}_B^\top \mathbf{G} r\boldsymbol{\beta}_B - \boldsymbol{\beta}_B^\top \mathbf{G} \boldsymbol{\beta}_C \\
0 < & \quad \boldsymbol{\beta}_B^\top \mathbf{G}^{1/2} \mathbf{G}^{1/2} r\boldsymbol{\beta}_B - \boldsymbol{\beta}_B^\top \mathbf{G}^{1/2} \mathbf{G}^{1/2} \boldsymbol{\beta}_C \\
0 < & \quad r \cdot ||\boldsymbol{\beta}_B^*|| \cdot ||\boldsymbol{\beta}_B^*|| - ||\boldsymbol{\beta}_B^*|| \cdot ||\boldsymbol{\beta}_C^*|| \cdot \cos(\theta) \\
||\boldsymbol{\beta}_C^*|| \cdot \cos(\theta) < & \quad r \cdot ||\boldsymbol{\beta}_B^*||
\end{aligned} \tag{62}$$

In Equation 34 we derive the conditions under which a system of traits will evolve so that they are costly to actors:

$$\begin{aligned}
0 < & \quad \boldsymbol{\beta}_C^\top \Delta \mathbf{a} \\
0 < & \quad \boldsymbol{\beta}_C^\top \mathbf{G} \left[r\boldsymbol{\beta}_B - \boldsymbol{\beta}_C \right] \\
0 < & \quad \boldsymbol{\beta}_C^\top \mathbf{G} r\boldsymbol{\beta}_B - \boldsymbol{\beta}_C^\top \mathbf{G} \boldsymbol{\beta}_C \\
0 < & \quad \boldsymbol{\beta}_C^\top \mathbf{G}^{1/2} \mathbf{G}^{1/2} r\boldsymbol{\beta}_B - \boldsymbol{\beta}_C^\top \mathbf{G}^{1/2} \mathbf{G}^{1/2} \boldsymbol{\beta}_C \\
0 < & \quad r \cdot ||\boldsymbol{\beta}_C^*|| \cdot ||\boldsymbol{\beta}_B^*|| \cdot \cos(\theta) - ||\boldsymbol{\beta}_C^*|| \cdot ||\boldsymbol{\beta}_C^*|| \\
0 < & \quad r \cdot ||\boldsymbol{\beta}_B^*|| \cdot \cos(\theta) - ||\boldsymbol{\beta}_C^*|| \\
||\boldsymbol{\beta}_C^*|| < & \quad r \cdot ||\boldsymbol{\beta}_B^*|| \cdot \cos(\theta)
\end{aligned} \tag{63}$$