Hamilton's Two Conceptions of Social Fitness*

Jonathan Birch†

Abstract

Hamilton introduced two conceptions of social fitness, which he called neighbour-modulated fitness and inclusive fitness. Although he regarded them as formally equivalent, a re-analysis of his own argument for their equivalence brings out two important assumptions on which it rests: weak additivity and actor's control. When weak additivity breaks down, neither fitness concept is appropriate in its original form. When actor's control breaks down, neighbour-modulated fitness may be appropriate, but inclusive fitness is not. Yet I argue that, despite its more limited domain of application, inclusive fitness provides a distinctively valuable perspective on social evolution.

1 Introduction

W. D. Hamilton is rightly celebrated as the progenitor of modern social evolution theory. This symposium marks the 50th anniversary of the publication of his seminal article, "The Genetical Evolution of Social Behaviour" (1964). It is a paper bursting with ideas, many of which were hugely innovative at the time. Perhaps the best known is the principle now called Hamilton's rule, which states that a social behaviour will be favoured by natural selection if and only if $rb > c$, where c is the fitness cost to the organism that performs the behaviour, b is the fitness benefit the trait confers on another organism, and r is the coefficient of relatedness.

I have discussed this idea elsewhere (Birch 2014a; Birch and Okasha 2015), but here I want to focus on two other major innovations. For in the same paper, Hamilton introduced two alternative ways of thinking about fitness in the context of social evolution. He called them *inclusive fitness* and *neighbour-modulated fitness* (Hamilton 1964, 5-6), and they continue to be the most commonly used fitness concepts in social evolution research.

Hamilton chose to focus on developing the inclusive fitness approach, and this continues to be the better known of the two. By the mid-1990s, however, the neighbour-modulated fitness approach had inconspicuously grown into a full-blown rival framework (Taylor and Frank 1996; Frank 1998), and in recent years it has become preferred methodology of many social evolution theorists (Taylor et al. 2007; Wenseleers et al. 2010; Frank 2013).

*This paper is based on my contribution to the symposium "50 Years of Inclusive Fitness", held at the 2014 PSA meeting in Chicago, IL, November 6-9 2014. I am very grateful to my fellow contributors (Ullica Segerstrale, Patrick Forber, Rory Smead, Dave Queller and Samir Okasha) and also to Andrew Buskell, Ellen Clarke, James Marshall and a reading group at ANU. This work was supported by a Philip Leverhulme Prize from the Leverhulme Trust.

†Department of Philosophy, Logic and Scientific Method, London School of Economics and Political Science, Houghton Street, London, WC2A 2AE, UK.

This has led to discussion of the relationship between the two fitness concepts. The key questions are: When are they equivalent for the purpose of calculating gene frequency change? When do they come apart? And when the choice between them is not forced by considerations of accuracy, which fitness concept is preferable? Here I revisit some of Hamilton's early papers in order to bring his own work to bear on these questions. In short, I will argue that although the neighbour-modulated fitness concept has a wider domain of application than the inclusive fitness concept, the latter remains distinctively valuable in those cases to which it does apply.

2 The Conceptual Contrast

To understand the difference between the two fitness concepts, consider two perspectives on what happens when altruism evolves due to genetic relatedness between organisms. One is to view relatedness as a source of *correlated interaction*: when r is high, bearers of the genes for altruism are more likely to interact with organisms who express that same gene, and hence are more likely to receive the benefits of altruism. Thus bearers of the genes for altruism may have greater reproductive success, on average, than non-bearers. The other is to view relatedness as a source of *indirect reproduction*: when r is high, recipients provide actors with an indirect means of securing genetic representation in future generations. Thus the genes for altruism may spread if the representation an altruist secures through helping its relatives exceeds the representation it loses by sacrificing a fraction of its own reproduction success.

The first perspective is captured in Hamilton's neighbour-modulated fitness approach, which analyses the correlations between an individual's genotype and its social neighbourhood in order to predict when these correlations will give bearers of the genes for altruism greater reproductive output, on average, than non-bearers (Hamilton 1964; Taylor and Frank 1996; Frank 1998). The second perspective is captured in Hamilton's inclusive fitness approach, which adds up all the fitness effects causally attributable to a social actor, weighting each component by a coefficient of relatedness, in order to calculate the net effect of a social behaviour on the actor's genetic representation in the next generation (Hamilton 1964; Frank 1998; Grafen 2006)

To pre-empt some misunderstandings, I should explain what these fitness concepts are *not*. First, neighbour-modulated fitness is not simply a new name for classical individual fitness. Neighbour-modulated fitness assumes that an individual's reproductive success can be decomposed into a sum of components, each attributable to a particular neighbourhood phenotype, plus a "baseline" component that is independent of what these neighbours do. Since the classical Darwinian fitness concept does not make any such assumption about the causal structure of fitness, it would be incorrect to simply equate the two (cf. Marshall 2015, 57-58).

Second, inclusive fitness is not simply an organism's classical individual fitness plus the classical individual fitness of its relatives, with the latter weighted by relatedness. This was never Hamilton's conception, and he avoided it with good reason. As Grafen (1982, 1984) emphasizes, any adequate fitness concept must be such that if bearers of one allele are, on average, fitter than bearers of another allele, then the former should be selected. A simple weighted sum violates this constraint, essentially because it allows the same offspring to be counted multiple times, once in computing the fitness of its parents, and again (and again, and again...) in computing the fitness of any collateral relatives.

This multiple-counting means that organisms with "bushier" family trees can be much "fitter" than organisms with more sparse family trees, even though the bushiness of one's family tree makes no difference in itself to the future representation of one's genes in a population.

What Hamilton (1964) saw from the outset is that inclusive fitness must be defined in a way that avoids multiple-counting. His solution was to assume that every organism's reproductive output can be written as a sum of components, each causally attributable to the behaviour of a specific actor. Given this assumption, we can make sure that each component is counted once and only once, by insisting that each component counts only towards the fitness of the actor who was causally responsible for it. As Hamilton (1964) himself put it:

> Inclusive fitness may be imagined as the personal fitness which an individual actually expresses in its production of adult offspring as it becomes after it has been stripped and augmented in a certain way. It is stripped of all components which can be considered as due to the individuals social environment, leaving the fitness he would express if not exposed to any of the harms or benefits of that environment. This quantity is then augmented by certain fractions of the quantities of harm and benefit which the individual himself causes to the fitness of his neighbours. The fractions in question are simply the coefficients of relationship. (Hamilton 1964, 8)

Inclusive fitness is thus an inherently causal notion: a weighted sum of the fitness components for which a given social actor is causally responsible.

Correlated interaction and indirect reproduction sound like very different processes, and neighbour-modulated and inclusive fitness sound like very different concepts. Despite this, the two are often considered formally equivalent, in the sense that they reliably yield identical predictions regarding the direction of gene frequency change (Wenseleers et al. 2010; Gardner et al. 2011; Queller 2011; Marshall 2015). Yet there have always been dissenters from the consensus. For example, Maynard Smith (1983) contrasted "the exact 'neighbour-modulated fitness' approach" with "the more intuitive 'inclusive fitness' method" (1983, 315). He advocated inclusive fitness on the grounds that he considered it easier to apply, but he thought it less accurate. More recently, Frank has advocated the neighbour-modulated approach, claiming that "inclusive fitness is more limited and more likely to cause confusion" (Frank 2013, 1172).

Hamilton (1964) claimed that the direction of selection can be calculated using either fitness concept, but he provided no formal argument for their equivalence. He did, however, include such an argument in his (1970) paper on selfishness and spite. The (1970) paper is quite brilliant: drawing on the work of Price (1970), Hamilton articulates clearly and concisely the basic insights he had presented in a rather dense way in earlier work. But perhaps the argument is a little *too* concise. As I will show, it leaves important assumptions unarticulated.

3 Hamilton's (1970) Argument Reconsidered

Here I reconstruct Hamilton's (1970) argument for the formal equivalence of his two fitness concepts.[1] In doing so, I want to draw attention to two assumptions Hamilton

[1]Frank (1998) also reconstructs and discusses Hamilton's model, though without drawing attention to its assumptions.

leaves implicit, since they point to important limitations of this equivalence.[2]

Consider a finite population of N numbered individuals. Let W_i^{tot} represent the total reproductive success of the i^{th}individual ("the recipient"), and let s_{ij} represent the additive effect of the social behaviour of the j^{th}individual ("the actor") on the reproductive success of the i^{th}individual.

We should pause here to consider the meaning of the s_{ij}. Hamilton (1970) simply glosses the s_{ij} as "additive effects". However, I suggest that, to do justice to Hamilton's explicitly causal conception of social fitness, we should interpret s_{ij} in explicitly *causal* terms, as the *causal* effect of the j^{th}individual on the reproductive success of the i^{th}individual. Roughly, it is the amount by which i^{th}individual's reproductive success would have differed had it never interacted with the j^{th}individual.

We can now return to the model. Let r_{ij} represent the coefficient of relatedness. More precisely, let $r_{ij} = \text{Cov}(q_i, q_j)/\text{Var}(q_j)$, where q_i is the recipient's individual gene frequency for a particular allele (i.e. its total number of copies of the allele divided by its ploidy) and q_j is the actor's individual gene frequency. This is equivalent to the simple regression (across all interacting pairs) of q_i on q_j

We can express W_i^{tot} as a sum of additive causal effects attributable to distinct social actors, plus a term representing its baseline non-social fitness (α_i), plus a residual component (ϵ_{W_i}) that represents deviations from fitness additivity:

$$W_i^{\text{tot}} = \alpha_i + \sum_j s_{ij} + \epsilon_{W_i}. \tag{1}$$

Let us define the neighbour-modulated fitness of the i^{th}individual as its total reproductive success *as predicted by this additive causal model*, neglecting the residual ϵ_{W_i}:

$$W_i^{\text{NMF}} = \alpha_i + \sum_j s_{ij} \tag{NMF}$$

If the residuals are all zero (i.e. $\epsilon_{W_i} = 0$ for all i), then an individual's neighbour-modulated fitness can be equated with its reproductive success. If $\epsilon_{W_i} \neq 0$, then neighbour-modulated fitness can still be used to calculate gene frequency change accurately provided the residuals do not co-vary with any genes. However, if these residuals do co-vary with genes, neighbour-modulated fitness may mislead. At this point, a substantive assumption of "additivity" is required. We must assume that the causal structure of social interaction is such that ϵ_{W_i} is either zero or, if non-zero, irrelevant to the direction and magnitude of gene frequency change.

Now let us define the inclusive fitness of the j^{th}individual as its baseline fitness (α_j) plus the sum of all the additive causal effects (s_{ij}) for which it is responsible, weighted in each case by the coefficient of relatedness r_{ij}:

$$W_j^{\text{IF}} = \alpha_j + \sum_i s_{ij} r_{ij} \tag{IF}$$

We can now ask: under what conditions does $\text{Cov}(W_i^{\text{NMF}}, q_i)$ equal $\text{Cov}(W_j^{\text{IF}}, q_j)$? This is the crucial question as regards the "formal equivalence" of the two fitness concepts. It turns out that, given one further important assumption, these quantities are equal.

[2]Readers wishing to avoid mathematical details may skip this section—but please note the two key assumptions stated verbally at the end of it!

First, we use our formal definitions of NMF and IF to split each covariance into a non-social and social component:

$$\mathrm{Cov}(W_i^{\mathrm{NMF}}, q_i) = \overbrace{\mathrm{Cov}(\alpha_i, q_i)}^{\text{non-social}} + \overbrace{\mathrm{Cov}\left(\sum_j s_{ij}, q_i\right)}^{\text{social}} \tag{2}$$

$$\mathrm{Cov}(W_j^{\mathrm{IF}}, q_j) = \overbrace{\mathrm{Cov}(\alpha_j, q_j)}^{\text{non-social}} + \overbrace{\mathrm{Cov}\left(\sum_i s_{ij} r_{ij}, q_j\right)}^{\text{social}}. \tag{3}$$

The non-social component is the same in both cases (since i and j are being used to label elements of the same set, the difference in indices is merely notational). Hence only the equivalence of the two social components needs to be established. Following Hamilton, let us call these components (respectively) the "neighbour-modulated fitness effect" and the "inclusive fitness effect". Hamilton further simplifies matters by assuming α to be a constant (of value 1), so that the non-social component is zero in both cases, but this assumption is dispensable to the argument.

The neighbour-modulated fitness effect can be rewritten as

$$\mathrm{Cov}\left(\sum_j s_{ij}, q_i\right) = \sum_j \left\{ \frac{1}{N} \sum_i (q_i - \overline{q}) s_{ij}) \right\}. \tag{4}$$

Now note that, from the definition of relatedness as the simple regression of q_i on q_j, it follows that

$$q_i - \overline{q} = r_{ij}(q_j - \overline{q}) + \epsilon_{q_i}, \tag{5}$$

where ϵ_q denotes the extent to which the recipient's actual genotype deviates from the regression prediction. Assume now that

$$\mathrm{Cov}\left(\sum_j s_{ij}, \epsilon_{q_i}\right) = 0. \tag{6}$$

This key assumption, which Hamilton (1970) makes implicitly, amounts to assuming that the recipient's individual gene frequency predicts its social fitness only via correlations with actors, and not via any other pathway (e.g. via conferring an ability on the recipient to make better use of the help of others). This entitles us to substitute $r_{ij}(q_j - \overline{q})$ for $(q_i - \overline{q})$ in equation 4, yielding

$$\mathrm{Cov}\left(\sum_j s_{ij}, q_i\right) = \sum_i \left\{ \frac{1}{N} \sum_j (q_j - \overline{q}) r_{ij} s_{ij}) \right\}. \tag{7}$$

The right-hand side of (7) can now be rewritten once again as a covariance:

$$\mathrm{Cov}\left(\sum_j s_{ij}, q_i\right) = \mathrm{Cov}\left(\sum_i s_{ij} r_{ij}, q_j\right). \tag{8}$$

Comparing this result to (2) and (3), we see that

$$\mathrm{Cov}(W_i^{\mathrm{NMF}}, q_i) = \mathrm{Cov}(W_j^{\mathrm{IF}}, q_j), \tag{9}$$

as we hoped to prove.

We can now see that Hamilton's (1970) argument for the formal equivalence of neighbour-modulated and inclusive fitness relies on two implicit assumptions, both of which amount to assumptions of uncorrelated residuals:

- *Additivity*: Deviations from fitness additivity (ϵ_{W_i}) are either zero or, if non-zero, irrelevant to the direction and magnitude of gene frequency change.
- *Actor's control*: The recipient's genotype predicts its social fitness only via its correlation with actor genotypes, and not via any other pathway (e.g. by enabling it to make better use of the help received).

Hamilton is not alone in relying on these assumptions. More recently, Grafen (2006, 543-549) has provided an argument for formal equivalence that improves on Hamilton's: in particular, it accommodates uncertainty, and it accommodates the various social "roles" an actor can occupy. Nevertheless, Grafen's argument still relies on the assumptions of additivity and actor's control. The only difference is that, while Hamilton left these assumptions implicit, Grafen makes them explicit.[3]

4 Actor's Control and Additivity

Actor's control points to one important qualification of Hamilton's equivalence result. If the recipient's genotype predicts the fitness effects it receives in ways that are *not* fully explained by correlations with actor genotypes, the result will be a situation in which $\mathrm{Cov}(\sum_j s_{ij}, \epsilon_q) \neq 0$. This is a situation in which neighbour-modulated fitness remains valid, but in which inclusive fitness could lead to errors.

Such a situation may seem hard to visualize. But all it needs is for there to be some genotype that, in addition to disposing an organism to express a social behaviour, also affects its ability to receive the benefit of that behaviour when expressed in others. Consider, for example, a genotype that disposes its bearer to produce an alarm call. In so doing, it reveals the organism's location to nearby predators, adversely affecting its ability to benefit from the alarm calls of others. In this scenario, the benefit of receiving an alarm call for a recipient does not just depend on properties of the actor. It also depends on whether or not the recipient has itself produced an alarm call.

Let us turn now to additivity. We should distinguish here between strong and weak varieties. If the deviation from additivity ϵ_{W_i} is zero for all individuals, then we can say that the structure of social interaction is *strongly additive*. If ϵ_{W_i} is non-zero for at least some individuals but makes no difference to changes in gene frequency, we can say that the structure of social interaction is *weakly additive*. Either way, we are talking about a property of the causal structure of social interaction in a population, not a property of any particular organism or gene.

It is clear that real social interactions frequently violate strong additivity. For recall what it requires: that an individual's social fitness can be written, without remainder,

[3]Frank (1998) derives a qualified equivalence result that holds under similar assumptions in a framework that incorporates class structure (discussed in Birch 2013).

as a sum of components, each reflecting the causal influence of a particular social actor. This is unlikely to be the case when cooperation takes the form of collaborative tasks involving multiple actors, because task-structure tends to creates situations in which the total payoff cannot be expressed as a sum of components, each corresponding to the difference made by a single actor's contribution (Birch 2012).

Weak additivity, however, is compatible with substantial deviations from the additive causal model. Its tenability in any particular case depends on whether these deviations co-vary with genes. This makes the empirical status of weak additivity difficult to assess. We are often in a position to know empirically that a social interaction violates strong additivity, since this depends only on the causal structure of the interaction, but we are less often in a position to know whether or not the deviations from strong additivity co-vary with any genes. I will not try to settle this empirical question here. Note, however, that failures of weak additivity are clearly possible in principle whenever there are deviations from strong additivity. We should therefore be cautious about assuming weak additivity when strong additivity is violated.

These considerations broadly support the view that, as Grafen puts it, "the assumption of additivity ... is not in general a realistic assumption. In many applications, non-additivity is an important part of the problem" (Grafen 2006, 543; see also Queller 1985, 2011; Marshall 2015). What does this mean for our two conceptions of social fitness, and for the relationship between them?

The immediate challenge is not to the formal equivalence of the two fitness concepts, but rather to the generality of both. Hamilton formulated both fitness concepts in terms of an additive causal model. If the model is inappropriate in some biological scenario, then both fitness concepts as Hamilton originally conceived them are inappropriate in that scenario.

Yet it would be wrong to conclude that the two fitness concepts are on a par when it comes to accommodating deviations from additivity. A key difference is that neighbour-modulated fitness, because it does not assume actor's control, has more leeway for accommodating effects that depend on the behaviour of multiple actors. Neighbour-modulated fitness requires that an individual's fitness can be expressed as a sum of effects attributable to properties of its social neighbourhood, but it does not require that each property is controlled by a single actor. This means that, as Queller (1985, 2011) has shown, we can augment the basic additive causal model with "synergistic effects" that depend in complex ways on the phenotypes of multiple actors (see also Marshall 2015, 66-67). This extended version of the neighbour-modulated fitness approach can handle cases in which deviations from the additive causal model arise from synergistic interactions.

By contrast, the inclusive fitness concept relies fundamentally on the assumption that each fitness effect can be attributed to a single controlling actor, whose inclusive fitness it counts towards. Since synergistic effects are not controlled by any single actor, there is no principled answer to the question of whose inclusive fitness they promote. In these contexts, inclusive fitness, as Hamilton conceived it, is no longer a well-defined property of an individual organism.

Thus, the two fitness concepts are threatened in different ways by failures of weak additivity. Put simply: for neighbour-modulated fitness, the problem is a technical one that can usually be surmounted by expanding the causal model of fitness. For inclusive fitness, however, the problem runs deeper, for it is a problem of a conceptual nature. The reassignment of fitness components to controlling actors that Hamilton envisaged is no longer possible when the additive causal model fails.

5 Inclusive Fitness, Adaptation and Selection-*for*

The bottom line is that, although neighbour-modulated and inclusive fitness are equivalent in their predictions when weak additivity and actor's control obtain, the inclusive fitness concept has a more restricted domain of application.

One might take this as an argument for abandoning the inclusive fitness concept. Yet I maintain that, for all its disadvantages, the inclusive fitness concept provides a distinctively valuable perspective on social evolution. This is because it provides a perspective from which we can make sense of altruistic (or indeed spiteful) behaviours as *adaptations* that have been selected-*for*, rather than as traits that were selected merely because they correlated with favourable social neighbourhoods (cf. Gardner 2009; West and Gardner 2013).

To make sense of this idea, let us briefly revisit Sober's (1984) selection-for/selection-of distinction:

> To say that there is selection for a given property means that having that property *causes* success in survival and reproduction. But to say that a given object was selected [or "selected-of"] is merely to say that the result of the selection process was to increase the representation of that kind of object. (Sober 1984, 100, his italics)

As Sober (1984, 197) observes, it seems intuitively important to some trait's being an adaptation that it has been selected-for, and not merely selected-of.

Within this framework, we can see the dilemma that altruistic (or spiteful) traits presented to biologists prior to Hamilton. On the one hand, these traits apparently could not be adaptations, because they make no causal contribution to their bearer's success in survival and reproduction. On the other hand, if they were not adaptations, their existence seemed to defy explanation.

Hamilton's two conceptions of social fitness resolve this dilemma in subtly different ways. Neighbour-modulated fitness offers one resolution: the property of being an altruist is selected not because it causally promotes its bearers' fitness but because, in populations with the right kind of structure, it systematically correlates with receiving the benefits of altruism in others. On this picture, it is the extrinsic property of having an advantageous social neighbourhood that is selected-for, whereas having the trait oneself is selected only because it correlates with this extrinsic property. If an individual could suppress its own altruism without altering its social environment, it would increase its fitness by doing so. From a neighbour-modulated fitness perspective, then, we can see why the property of being an altruist is sometimes selected, but it is only ever selected-*of*.

By contrast, the inclusive fitness approach offers a resolution that puts selection-*for* back at the centre of the picture. From an inclusive fitness perspective, a social trait that detracts from the actor's viability or fecundity may still contribute causally to its fitness, and evolve for that reason, if the benefits of expressing the trait fall systematically on genetic relatives. For this reason, I suggest, the inclusive fitness concept provides the more satisfying resolution to the dilemma. It shows how being an altruist can be an adaptation, not just a correlate of having a favourable social environment.

6 Inclusive Fitness Maximization

A different way of arguing for the distinctive value of inclusive fitness is to argue that organisms *maximize* this quantity, in a certain sense of the term (Grafen 2006). Hamilton (1964, 1) himself made such a claim, writing that that populations satisfying the assumptions of his model "should tend to evolve behaviour such that each organism appears to be attempting to maximize its inclusive fitness".

Pinning down the sense of "maximization" at stake here is crucial in order to evaluate Hamilton's claim, since there are senses of the term on which inclusive fitness is clearly not maximized (Birch 2016). In economics, it is common to model humans as "maximizing agents" who make strategic choices that, within a set of feasible options, maximize a quantity known as utility. Along similar lines, behavioural ecologists often assume that organisms behave in ways that maximize, within a set of feasible options, their inclusive fitness. Following Grafen (1984, 1999), we can call this an "individual as maximizing agent" analogy.

What enables inclusive fitness to play this role as the putative "maximand" of animal behaviour is its focus on which actors control which phenotypes. Because an actor's inclusive fitness is a weighted sum of the fitness effects for which it is causally responsible, we can put ourselves in the position of the actor and ask: "How should I behave, in order to maximize my expected inclusive fitness?" Since this quantity is under the actor's control, this can serve as an informal route to predictions of how we should expect an organism to behave. By contrast, we cannot usefully ask the same question with regard to neighbour-modulated fitness, because an individual's neighbour-modulated fitness contains components over which it may have no control. All we can do is put ourselves in the position of a recipient and ask: "What genotypes are correlated with good outcomes, as far as my neighbour-modulated fitness is concerned?" But this heuristic is less intuitive, because considerations of causation and control are replaced by considerations of statistical auspiciousness (cf. Gardner 2009; Marshall 2015).

Hamilton's claim, as quoted above, appears to invoke an "individual as maximizing agent" analogy. However, Hamilton provided no formal argument for its validity. What he actually showed was that, within his one-locus model, the mean inclusive fitness of the population increases until a local maximum is reached. Population geneticists have constructed numerous counterexamples to this sort of mean fitness maximization (reviewed in Birch 2016), so this feature of Hamilton's model cannot be considered a general truth about the operation of natural selection.

Grafen's (2006; 2014) "Formal Darwinism project" can be regarded as a sophisticated attempt to vindicate the "individual as maximizing agent" analogy that Hamilton verbally gestured towards. It aims to do this by forging links between formal representations of gene frequency change and optimal strategy choice. In a nutshell, Grafen aims to prove that "natural selection always changes gene frequencies in the direction of increasing inclusive fitness; and that a population genetic equilibrium in which no feasible mutations can spread implies that the individuals in the population are each acting so as to maximize their inclusive fitness." (Grafen 2006, 543)

I have criticized the Formal Darwinism project elsewhere, and I cannot do justice to this complex topic here (Birch 2014b, 2016). In short, Grafen does prove what he aimed to prove, given a very specific and unorthodox understanding of the concept of "equilibrium". But the "equilibrium" concept that features in his links, defined in terms of "scope" and "potential" for selection, is neither necessary nor sufficient for a population-

genetic equilibrium in the usual sense. The true relationship between population-genetic equilibria and inclusive fitness maxima is much more complicated than Grafen's links initially suggest.

Despite my doubts about Formal Darwinism, I remain convinced that the inclusive fitness concept remains valuable. One reason is that a heuristic need not be completely or even mostly reliable in order to warrant its continued use. If the "individual as maximizing agent" analogy sometimes generates fruitful hypotheses, as it surely has done, then this provides a pragmatic justification for its use as a method of hypothesis generation. We do not need to put empirical projects on hold while we wait for a theoretical argument to reassure us that the hypotheses thus generated will be correct.

More fundamentally, however, we can reject the idea that organisms in any sense maximize their inclusive fitness and yet retain the idea that social traits are selected because they causally contribute to this quantity. This milder claim is already enough to make inclusive fitness valuable. For it allows us to see how social traits that detract from their bearers' classical fitness can be selected-for, not just selected-of, and it thus allows us to see how such traits can be adaptations, regardless of whether or not they maximize inclusive fitness.

References

Birch, Jonathan. 2012. "Collective Action in the Fraternal Transitions." *Biology and Philosophy* 27:363–80.

— — —. 2013. "Kin Selection: A Philosophical Analysis." PhD diss., University of Cambridge.

— — —. 2014a. "Hamilton's Rule and Its Discontents." *British Journal for the Philosophy of Science* 65:381–411.

— — —. 2014b. "Has Grafen Formalized Darwin?" *Biology and Philosophy* 29:175–80.

— — —. 2016. "Natural Selection and the Maximization of Fitness." *Biological Reviews*. doi: 10.1111/brv.12190.

Birch, Jonathan and Samir Okasha. 2015. "Kin Selection and Its Critics." *BioScience* 65:22–32.

Frank, Steven A. 1998. *Foundations of Social Evolution*. Princeton, NJ: Princeton University Press.

— — —. 2013. "Natural Selection. VII. History and Interpretation of Kin Selection Theory." *Journal of Evolutionary Biology* 26:1151–84.

Gardner, Andy. 2009. "Adaptation as Organism Design." *Biology Letters* 5:861–64.

Gardner, Andy, Stuart A. West, and Geoff Wild. 2011. "The Genetical Theory of Kin Selection." *Journal of Evolutionary Biology* 24:1020–43.

Grafen, Alan. 1982. "How Not to Measure Inclusive Fitness." *Nature* 298:425–26.

— — —. 1984. "Natural Selection, Kin Selection and Group Selection." In *Behavioural Ecology* (2nd ed.), ed. John R. Krebs and Nicholas B. Davies, 62–84. Oxford: Blackwell.

— — —. 1999. "Formal Darwinism, the Individual-as-Maximising-Agent Analogy, and Bet-Hedging." *Proceedings of the Royal Society B: Biological Sciences* 266:799–803.

— — —. 2006. "Optimization of Inclusive Fitness." *Journal of Theoretical Biology* 238:541–63.

— — —. 2014. "The Formal Darwinism Project in Outline." *Biology and Philosophy* 29:155–74.

Hamilton, William D. 1964. "The Genetical Evolution of Social Behaviour." *Journal of Theoretical Biology* 7:1–52.

— — —. 1970. "Selfish and Spiteful Behaviour in an Evolutionary Model." *Nature* 228:1218–20.

Marshall, James A. R. 2015. *Social Evolution and Inclusive Fitness Theory: An Introduction*. Princeton, NJ: Princeton University Press.

Maynard Smith, John. 1983. "Models of Evolution." *Proceedings of the Royal Society B: Biological Sciences* 219:315–325.

Price, George R. 1970. "Selection and Covariance." *Nature* 227:520–21.

Queller, David C. 1985. "Kinship, Reciprocity, and Synergism in the Evolution of Social Behaviour." *Nature* 318:366–67.

— — —. 2011. "Expanded Social Fitness and Hamilton's Rule for Kin, Kith and Kind." *Proceedings of the National Academy of Sciences USA* 108:10792–99.

Sober, Elliott. 1984. *The Nature of Selection: Evolutionary Theory in Philosophical Focus*. Chicago, IL: University of Chicago Press.

Taylor, Peter D. and Steven A. Frank. 1996. "How to Make a Kin Selection Model." *Journal of Theoretical Biology* 180:27–37.

Taylor, Peter D., Geoff Wild and Andy Gardner. 2007. Direct Fitness or Inclusive Fitness: How Shall We Model Kin Selection? *Journal of Evolutionary Biology* 20:301–9.

Wenseleers, Tom, Andy Gardner and Kevin R. Foster. 2010. Social Evolution Theory: A Review of Methods and Approaches. In *Social Behaviour: Genes, Ecology and Evolution*, ed. Tamás Székely, Allen J. Moore, and Jan Komdeur, 132–58. Cambridge: Cambridge University Press.

West, Stuart A. and Andy Gardner. 2013. Inclusive Fitness and Adaptation. *Current Biology* 23:R577–84.