Inclusive Fitness as a Criterion for Improvement

Jonathan Birch

Department of Philosophy, Logic and Scientific Method,
London School of Economics and Political Science,
London, WC2A 2AE, United Kingdom.

Email: j.birch2@lse.ac.uk
Webpage: http://personal.lse.ac.uk/birchj1

This article appears in a special issue of *Studies in History & Philosophy of Biological & Biomedical Sciences* on Optimality and Adaptation in Evolutionary Biology, edited by Nicola Bertoldi.

16 July 2019

Abstract: I distinguish two roles for a fitness concept in the context of explaining cumulative adaptive evolution: fitness as a predictor of gene frequency change, and fitness as a criterion for phenotypic improvement. Critics of inclusive fitness argue, correctly, that it is not an ideal fitness concept for the purpose of predicting gene-frequency change, since it relies on assumptions about the causal structure of social interaction that are unlikely to be exactly true in real populations, and that hold as approximations only given a specific type of weak selection. However, Hamilton took this type of weak selection, on independent grounds, to be responsible for cumulative assembly of complex adaptations. In this special context, I argue that inclusive fitness is distinctively valuable as a criterion for improvement and a standard for optimality. Yet to call inclusive fitness a criterion for improvement and a standard for optimality is not to make any claim about the frequency with which inclusive fitness optimization actually occurs in nature. This is an empirical question that cannot be settled by theory alone. I close with some reflections on the place of inclusive fitness in the long running clash between 'causalist' and 'statisticalist' conceptions of fitness.

Key words: inclusive fitness, adaptation, natural selection, evolution, causality

1. What do we want from a fitness concept?

For neo-Darwinians, the evolution of complex adaptation is a gradual process. Complex traits do not pop into existence fully formed; they are assembled via the accumulation of small improvements. Darwin (1859), in a famous passage on the origin of the eye, expresses this idea powerfully and beautifully:

> If we must compare the eye to an optical instrument, we ought in imagination to take a thick layer of transparent tissue, with a nerve sensitive to light beneath, and then suppose every part of this layer to be continually changing slowly in density, so as to separate into layers of different densities and thicknesses, placed at different distances from each other, and with the surfaces of each layer slowly changing in form. Further we must suppose that there is a power always intently watching each slight accidental alteration in the transparent layers; and carefully selecting each alteration which, under varied circumstances, may in any way, or in any degree, tend to produce a distincter image. We must suppose each new state of the instrument to be multiplied by the million; and each to be preserved till a better be produced, and then the old ones to be destroyed. In living bodies, variation will cause the slight alterations, generation will multiply them almost infinitely, and natural selection will pick out with unerring skill each improvement. Let this process go on for millions on millions of years; and during each year on millions of individuals of many kinds; and may we not believe that a living optical instrument might thus be formed as superior to one of glass, as the works of the Creator are to those of man? (Darwin, 1859, pp. 188–189)

Most models in population genetics do not attempt to model this process. They focus instead on changes in allele frequency over what is, comparatively speaking, a shorter timescale: the timescale of so-called *microevolution*. A number of authors have noted the disconnect between the microevolutionary timescale of most population genetics models and the longer timescale over which complex adaptations are cumulatively assembled (Eshel and Feldman, 1984, 2001; Neander, 1995; Hammerstein, 1996, Eshel et al., 1998; Godfrey-Smith and Wilkins, 2009; Metz, 2011). Over the short term, evolution by natural selection leads to equilibria that are stable given the alleles currently competing in the population, but that may eventually be destabilized by the appearance of novel mutants. Cumulative adaptive evolution occurs over many successive episodes of short-term allele frequency change, as one novel mutant responsible for a small improvement arises by chance and spreads through the population, eventually followed by another, and another, and so on, in a process memorably described by Dawkins (1996) as ‘climbing Mount Improbable’.

The timescale of cumulative adaptive evolution is clearly a longer timescale than that of microevolution to a short-term stable equilibrium, but nor are we talking here about the timescale of *macroevolution* (Gould, 2002; Sterelny, 2007), since these changes are typically still occurring within a single species. Metz (2011) and Godfrey-Smith (2012) have suggested the term *mesoevolution* for this intermediate timescale. We might also call it, less technically, the *evolutionary medium term*.

In the context of explaining the cumulative, step-by-step assembly of complex adaptation over the medium term, what explanatory roles do we want a fitness concept to play? I suggest there are two main roles. First, we want a fitness concept to help us predict, within a single microevolutionary step, the trajectory of the population through gene-frequency space. If the bearers of a mutant allele *B* are currently fitter, on average, than the bearers of a wild type allele *A*, it should be the case that *B* will (probably, allowing for drift) increase in frequency relative to *A*. If all alleles present in the population are currently such that their bearers are, on average, equal in fitness, it should be the case that, for the moment at least, natural selection does not alter allele frequencies.

Second, and perhaps less obviously, we want a fitness concept to provide a stable criterion, throughout the whole cumulative process, for what constitutes an improvement to an organism's whole phenotype. In other words, a fitness concept should tell us, whenever a novel mutant appears, whether or not that mutant has improved the organism's phenotype relative to the wild type. To play this role, fitness must be a property, *X*, of an individual organism such that new mutant phenotypes are systematically selected-for (in the sense of Sober, 1984) over the wild type if and only if they make a positive causal contribution, in contrast with the wild type, to *X*. A fitness concept that provides a criterion for improvement also provides a standard for optimality: roughly speaking, we can say that a phenotype is at least locally optimal when there is no more room for improvement.

The relationship between these roles—predictor of gene-frequency change and criterion for phenotypic improvement—is a subtle one.[1,2] A good predictor of gene-frequency change is not necessarily a good criterion for improvement, because a good criterion for improvement has to be a property of an individual organism, not of an allele or a genotype. Moreover, a good criterion for improvement has to apply at every stage in a multi-step process of cumulative adaptation, and therefore has to deliver reliably correct verdicts as to what selection will favour despite changes over time in the environment, the composition of the population, and the particular phenotypic differences that selection is targeting at any given moment. A good predictor of gene-frequency change will not necessarily be well suited to this role.

[1] The distinction maps imperfectly on to Sober's (2001) well-known distinction between the 'two faces of fitness'. The 'predictor' role is reminiscent of Sober's 'mathematical predictor', while the 'criterion for improvement' role is perhaps, at first sight, reminiscent of Sober's 'ecological descriptor'. However, the distinction I have in mind derives from consideration of different timescales in a process of cumulative adaptation, not mathematics vs. ecology. The notion of a criterion for improvement can be given a mathematical representation; indeed, as I explain in Sections 4 and 5, I see this as an attractive way of understanding the concept of inclusive fitness.

[2] The distinction may also, at first sight, call to mind Mills and Beatty's (1979, p. 275) distinction between 'short-term' and 'long-term' fitness, where the former notion refers to an organism's expected number of *immediate offspring* and the latter refers to the number of descendants it leaves after *some specified number of subsequent generations* (e.g. its number of grandoffspring, or great-grandoffspring). Despite the superficially similar terminology, Mills and Beatty's distinction cross-cuts mine. A good predictor of gene-frequency change may sometimes have to take into account numbers of grandoffspring, etc. (as in, for example, sex ratio models), whereas a good criterion for phenotypic improvement may sometimes still be 'short-term' in the sense of neglecting grandoffspring, etc., in cases in which effects on later generations are not relevant to the direction of selection-for.

For a simple example, consider the exponential growth rate of a genotype, or 'Malthusian parameter' (Fisher, 1930; Charlesworth, 1970; Crow & Kimura, 1970). This is a commonly seen fitness measure in population genetics, and a useful one for predicting and explaining changes in gene frequency. Yet it makes no sense to say, of a series of small changes to a complex phenotype, that they qualify as improvements by virtue of increasing their bearer's exponential growth rate. A growth rate in this sense, being a property of a genotype or allele rather than an individual organism, is not an intelligible criterion for whole-phenotype improvement.

Conversely, a good criterion for phenotypic improvement need not be an especially good predictor of gene-frequency change. As noted above, it must be such that, in a process of cumulative adaptation, mutants are systematically selected-for over the wild type if and only if, relative to the wild type, they make a positive causal contribution to it. It will therefore not be *irrelevant* to the prediction of gene-frequency change. But it need not deliver exact, accurate, quantitative predictions. It is enough that it delivers reliably correct qualitative verdicts about the direction of selection-for. It may be able to do this while making various approximations that limit its predictive power. Moreover, it need only apply in the special context of a process of cumulative adaptation. It may therefore make assumptions which are valid in that context, but which limit its usefulness outside that context.

These considerations bear directly on the status of one of the most contested fitness concepts in evolutionary theory: Hamilton's concept of *inclusive fitness* (Hamilton, 1964, 1970). It is hard to explain to non-specialists the strength of feeling this concept evokes. To its critics, such as Nowak, Wilson and colleagues, it is an arcane, obsolete construct, based on lazy approximations and implausible assumptions, that holds back progress in the field (Nowak et al., 2010; Wilson, 2012; Allen et al., 2013). To its defenders, such as Gardner and West, it is the insight of a genius, and an essential concept for understanding social adaptation (Gardner, 2009, 2017; Gardner et al., 2011; West and Gardner, 2013).

My argument in this paper is that we can better understand the nature of this disagreement, and the way forward, by distinguishing the two roles for a fitness concept outlined above. As a predictor of gene-frequency change, inclusive fitness as Hamilton conceived it is valid only in special cases, and typically only as a first-order approximation. Therefore, if one starts with the assumption that a fitness concept should be an exact, quantitative predictor of gene-frequency change in as many cases as possible, it becomes hard to see the value of inclusive fitness. The concept comes into its own, however, as a criterion for phenotypic improvement in the context of the cumulative assembly of social adaptations. For this reason, it would be a great loss to social evolution research if the inclusive fitness concept were declared obsolete.

2. Inclusive fitness as a predictor of gene-frequency change

Inclusive fitness is, at first sight, a strange quantity. Here is how Hamilton (1964) defined 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 individual's 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, p. 8)

For Hamilton, inclusive fitness was, first and foremost, a property of an individual organism. We can, of course, talk of the mean inclusive fitness of a population, or of the bearers of a particular trait. But these are derivative notions: the fundamental notion is a property of an individual (cf. Pence and Ramsey, 2015). It also an inherently *causal* notion: a weighted sum of the effects on reproductive success for which the focal organism is *causally responsible* (Figure 1).

[FIGURE 1]

Inclusive fitness is committed, at a conceptual level, to the validity of an additive causal model of fitness (Birch, 2016a). The procedure Hamilton describes in the above paragraph involves crediting components of reproductive output to the actors whose social behaviour was causally responsible for them, rather than crediting them to the organisms that actually produced the offspring. For example, the larvae produced by a queen in a social insect colony should be credited not to the queen but to the workers who rear them. If this procedure is to avoid problems of double-counting, it must be that the reproductive success of an organism can be written as a *sum* of components, each of which is attributable to the social behaviour of a single social actor. Let us call this assumption *fitness additivity*. Moreover, it must be that the value of each component depends only on the genotype of the actor, and not on the genotype of the recipient, an assumption known as *actor's control* (Grafen, 2006).

The reliance of inclusive fitness, as Hamilton defined it, on fitness additivity and actor's control is something even its most committed defenders should acknowledge, and at least some of them do. Grafen (2006, p. 544), for example, writes that 'the question of how to define inclusive fitness in the absence of additivity has not been settled, and so fundamental theory on the non-additive case can hardly yet begin'.

Critics of inclusive fitness argue, plausibly, that these assumptions are unlikely to be exactly true in real evolving populations (Nowak et al., 2010; Allen et al., 2013; Allen and Nowak, 2016). Consider, for example, a genotype that disposes its bearer to produce an alarm call (Queller, 1985). In so doing, it reveals the organism's location to nearby predators, reducing its ability to benefit from the alarm calls of others. In this scenario, the benefit of receiving an alarm call does not just depend on the behaviour (and genotype) of the actor. It also depends on a fact about the recipient's behaviour (i.e. whether or not the recipient has also produced an alarm call) that is dependent on its genotype. Actor's control is violated.

A more formal way to put the problem is to say the alarm call has *second-order* effects on reproductive output. Suppose the wild type strategy is to produce an alarm call with probability zero on seeing a predator (i.e. never), and that this is competing against a mutant strategy that produces an alarm call with probability q. Possessing the mutant strategy rather than the wild type imposes a viability cost on the actor proportional to q, and confers a viability benefit proportional to q on each recipient. These are its *first-order* effects. But if a recipient itself possesses the mutant strategy, there will also be a *second-order* effect, proportional to q^2, that at least partially cancels out the benefit to that recipient, because an organism which produces an alarm call is less able to reap the benefit of an alarm call produced by others. It is this second-order effect—an effect that depends on the genotype of both actor and recipient—that makes trouble for the actor's control assumption.

Defenders of inclusive fitness should accept this too. Fitness additivity and actor's control are strong assumptions; they are unlikely to be exactly true in real populations. Again, Grafen (2006, p. 543) acknowledges this, writing, for example, that 'the assumption of additivity is made throughout this paper, but is not in general a realistic assumption. In many applications, non-additivity is an important part of the problem.' Critics of inclusive fitness may reply, with some justification, that this point is absent from some of the more forthright defences of inclusive fitness, such as Abbot et al. (2011).[3]

This, however, is not the end of the story. Defenders of inclusive fitness have often noted that its assumptions can be justified as *first-order approximations* if we assume a specific form of weak selection, usually known as δ-weak selection (Grafen, 1985, 2006; Wild and Traulsen, 2007). To assume δ-weak selection is to assume that the character of interest is a quantitative character, and that the alternatives competing in the population are a wild type and a mutant that differs only very slightly from the wild type. For example, a δ-weak selection model of an alarm call scenario might pit a wild type strategy in which an organism makes an alarm call with probability q against a mutant strategy in which an organism makes an alarm call with probability $q + \delta$, where δ is a very small increment such that $\delta^2 \approx 0$ (Wild and Traulsen, 2007).

With this assumption in place, we can reinterpret inclusive fitness in terms of *marginal* or *differential* causal effects rather than *total* causal effects. That is, instead of defining an

[3] There are certainly ways to formulate *Hamilton's rule* so that it works without assuming actor's control; see the discussions of 'HRG' in Birch (2014, 2017a, b) and Birch and Okasha (2015). But Hamilton's rule is a population-level statistical result, and it should be distinguished from *inclusive fitness*, which Hamilton clearly conceptualized as a property of an individual organism. A criterion for improvement must be a property of an individual organism, and inclusive fitness *qua* property of an individual organism indispensably relies on an additive causal model of reproductive success (Birch, 2016a).

actor's inclusive fitness as a baseline non-social component plus a weighted sum of the *total* effects of its social behaviour on reproductive success, we instead define an actor's inclusive fitness as a baseline component that includes all the effects of the wild type, to which we then add a weighted sum of the *differential* effects of its actual behavioural phenotype on reproductive success, *relative* to a default scenario in which the actor expresses the wild type behaviour. On this marginal interpretation, fitness additivity and actor's control can be reinterpreted as assumptions about marginal effects: what is assumed is that the marginal effects of expressing the mutant phenotype *rather than the wild type* are additive and actor controlled.

The result is that the assumptions that initially seemed too strong become reasonable as approximations. To illustrate these points let us return to our alarm call example. The original problem was that making an alarm call reduces the benefit an organism receives from an alarm call expressed in others, leading to a violation of actor's control. But now consider the marginal effect of making an alarm call with probability $q + \delta$ rather than probability q, for some very small increment δ. This will have a first-order marginal effect (proportional to δ) on one's own reproductive success and a first-order marginal effect on the reproductive success of nearby recipients. It will also have a second-order effect (proportional to δ^2) on the benefit one receives from a marginal increase in the probability with which another nearby individual makes an alarm call. However, this second-order effect, which is the source of the trouble for the actor's control assumption, is precisely the kind of effect that the assumption of δ-weak selection entitles us to regard as approximately zero, since it relies on the product of two tiny phenotypic differences.

To put the point more formally, let B denote the first-order benefit conferred by an alarm call on a recipient, let $-C$ denote the first-order cost, let $-D$ denote the reduction in the benefit that accrues to the recipient when it has itself produced an alarm call, and let r denote the coefficient of relatedness between actors and recipients. The marginal effect of the mutant phenotype on the actor is $-\ \delta C$. The marginal effect of the mutant phenotype on a recipient who possesses the wild type strategy is δB, and the marginal effect on a recipient who also possesses the mutant strategy is $\delta B - \delta^2 D$. However, on the assumption that $\delta^2 \approx 0$, the marginal effect is approximately δB regardless of the genotype of the recipient. The second-order effect disappears, actor's control is restored, and we can write the overall marginal effect of the strategy on the actor's inclusive fitness as $r\delta B - \delta C$.

3. Why weak selection?

To critics of inclusive fitness, this appeal to δ-weak selection seems *ad hoc*: to justify two questionable assumptions, we invoke another assumption that appears no less questionable. Why think that selection is usually δ-weak? Why think it is ever δ-weak? It may even seem rather absurd: why go to such trouble to show that inclusive fitness delivers *approximately* accurate calculations of gene-frequency change under a narrowly circumscribed range of conditions, when other, simpler fitness measures, such as an organism's personal reproductive success, can deliver *exact* calculations under a wider range of conditions (Nowak et al., 2010; Allen et al., 2013)? However, I maintain that this move is neither *ad hoc* nor absurd. It is fairer, I think, to see it as an assumption grounded in some important background commitments of inclusive fitness theory—commitments that can be traced back to Hamilton, but which critics of inclusive fitness do not necessarily share.

At the heart of the Hamiltonian tradition is a version of adaptationism that takes complex adaptation, or 'organism design', to be the explanatory target of social evolution research (Gardner, 2009, 2017; Grafen, 2014). This is combined with an empirical commitment to a gradualist picture of how complex adaptation arises. Fisher (1930), a major influence on Hamilton, took complex adaptation to result from the gradual accumulation of mutations with tiny phenotypic effects.[4] Fisher posited small-effect mutations on the grounds that large-effect mutations are much less likely to cause adaptive improvements. In support of this, he offered two iconic arguments: one involving an informal analogy with a microscope, the other involving a more formal geometric model (Fisher, 1930, pp. 38-41).

To paraphrase (and simplify) the informal argument, suppose you are attempting to focus a microscope by turning an adjustment knob. Knowing nothing of microscopes, you have no idea which way to turn the knob, so you turn it in a random direction. If the adjustment is very small, there is a 50% chance it will improve the focus, because any very small adjustment in the right direction will help. But the larger the adjustment gets, the lower the probability it will be an improvement, because it becomes ever more likely that an adjustment, even if it happens to be in the right direction, will overshoot the target.

Using a geometric model in which a population is displaced from the optimum in phenotypic space and must find its way back to the optimum through random gene substitutions, Fisher showed that the probability of an improvement, which falls off with the size of the adjustment even in the one-dimensional case, falls off more rapidly in the case of an adjustment in two dimensions, and falls off very rapidly indeed when we are adjusting at random in many dimensions, as in the case of a mutation that affects many aspects of the phenotype. The chance of improvement is greatest, at 50%, for a mutation that affects the phenotype by an infinitesimal amount.

Fisher's argument has not been without its critics. Kimura (1983) argued that, in finite populations, mutants with larger effects on the phenotype have a greater chance of going to fixation, because mutants with small effects are prone to drifting out of existence. Orr (1998) showed that both Fisher and Kimura could be partially vindicated in relation to different

[4] Darwin was also a gradualist about complex adaptation, as shown clearly by the passage quoted at the start of this article. However, it would be tendentious to project on to Darwin the whole package of Fisher's views about the typical strength of selection and the typical effect size of adaptive substitutions. Darwin, knowing little about the mechanisms of inheritance, at times appears to invoke very strong selection as a way of preserving adaptive phenotypes in the face of erosion by blending inheritance (Lewens, 2010a).

stages of the process of cumulative adaptation: the typical effect size of a mutation fixed at an early stage in the process, when the phenotype is far from the optimum, is much larger than Fisher thought; but, as the phenotype gets closer to optimality, Fisher's concern about 'overshooting' becomes increasingly salient and the typical effect size of a fixed mutation becomes progressively smaller.

But although Fisher's argument remains a source of debate (for more recent contributions, see Waxman and Welch, 2005; Martin and Lenormand, 2006), what matters for our purposes is that a commitment to Fisherian gradualism is at the heart of Hamilton's theory of social evolution. Consider, for example, the following *credo* from Hamilton's collected papers:

> I was and still am a Darwinian gradualist for most of the issues of evolutionary change. Most change comes, I believe, through selected alleles that make small modifications to existing structure and behaviour. If one could understand just this case in social situations, who cared much what might happen in the rare cases where the gene changes were great and happened not to be disastrous? Whether under social or classical selection, defeat and disappearance would, as always, be the usual outcome of genes that cause large changes. I think that a lot of the objection to so-called 'reductionism' and 'bean-bag reasoning' directed at Neodarwinist theory comes from people who, whether through inscrutable private agendas or ignorance, are not gradualists, being instead inhabitants of some imagined world of super-fast progress. Big changes, strong interlocus interactions, hopeful monsters, mutations so abundant and so hopeful that several may be under selection at one time—these have to be the stuff of their dreams if their criticisms are to make sense. (Hamilton, 1996, pp. 27-28).

Thus a focus on δ-weak selection is grounded in the core commitments of Hamilton's program. The subset of selection processes for which inclusive fitness is a valid fitness concept is the same subset Hamilton and his successors take, on independent grounds, to be responsible for the cumulative assembly of complex adaptations.

4. Inclusive fitness as a criterion for improvement

For all this, critics may still be bemused: the fact that inclusive fitness delivers at least approximately accurate calculations of gene frequency change under δ-weak selection does not give it any advantage over other predictively accurate fitness concepts, even granting the point that δ-weak selection is type of selection with special evolutionary significance. The question remains: why not use a simpler fitness concept, such as personal reproductive success, that does not require us to assume weak selection or neglect second-order effects? Bemusement of this sort is palpable in the writings of Nowak and colleagues (Nowak et al., 2010; Allen et al., 2013).

This is where the second role for a fitness concept becomes important. In the context of explaining cumulative adaptation, we want a fitness concept that can serve as a *criterion for phenotypic improvement*. As explained in Section 1, a criterion for improvement is a property, X, of an individual organism such that new mutants are systematically selected-for over the wild type if and only if they make a positive causal contribution (in contrast with the wild type) to *X*. I contend that the distinctive advantage of inclusive fitness over other fitness concepts is that it is a particularly good candidate for property *X*.

Consider, for example, a process of social evolution in which natural selection gradually shapes various different aspects of a complex social strategy involving the conditional expression of different actions in different contexts. In one context, C_1, the strategy produces actions that benefit the actor; in another context, C_2, the strategy produces actions that confer benefits on genetically related recipients. Mutants periodically arise (one at a time) that alter some aspect of the strategy very slightly, implying δ-weak selection.

Suppose natural selection targets different aspects of the phenotype at different times: the strategy is initially shaped by selection for enhanced benefits for the actor in C_1, then goes through a stage in which it is shaped by selection for greater benefits conferred on genetically related recipients in C_2, and then finally goes through a streamlining stage in which the cost to the actor of conferring benefits on relatives in C_2 is gradually reduced. A more realistic scenario would involve the shaping of all these aspects of the phenotype as and when relevant mutants arise, but for the purpose of fixing ideas it helps to think to selection targeting different aspects in discrete stages.

In such a process, the actor's personal reproductive success is not a suitable criterion for improvement, because mutants that causally detract from this quantity are selected-for during the second stage. The personal reproductive success of the recipients is not a suitable criterion either, because mutants that may be neutral or deleterious with respect to this quantity are selected-for in the initial and final stages. An appropriate criterion for improvement must be more 'inclusive' than the personal reproductive success of any single organism. It must include the social effects of an organism's behaviour on the reproductive output of other organisms, so that it counts mutants in the first and third stages as improvements by virtue of their actor-directed effects, and counts mutants in the second stage as improvements by virtue of their recipient-directed effects. It must therefore be a quantity that is appropriately sensitive to the reproductive success of both actor and recipient, weighting these quantities in such a way as to deliver accurate qualitative verdicts as to whether or not the mutant will be selected-for over the wild type.

Inclusive fitness is that quantity. Provided selection is δ-weak, it provides the correct weighting of actor-directed and recipient-directed effects. All and only those mutants that causally promote (in contrast to the wild type) the inclusive fitness of the actor will be selected-for. Because inclusive fitness relies on assumptions that are only justified as approximations, it may not be the best fitness concept for calculating the precise magnitude of gene frequency change at any particular moment in time. But that is not the job we are asking it to perform. Its job is to produce correct verdicts regarding the direction of selection-for: to be such that, throughout the entire medium-term process of cumulative social adaptation, all and only those mutants that causally promote it are selected-for.

The cumulative assembly of social adaptations by δ-weak selection thus constitutes a special context in which inclusive fitness is both valid and valuable. It is valid because its assumptions are reasonable as first-order approximations when selection is δ-weak. It is valuable because, unlike other fitness concepts, it provides a stable criterion, throughout the whole process, for what constitutes an improvement to the phenotype.

5. Should we expect inclusive fitness to be optimized?

This distinctive role for inclusive fitness as a criterion for improvement points to a close relationship between inclusive fitness and the concept of optimality, because a criterion for improvement implies a standard with respect to which the optimality (or suboptimality) of a phenotype can be judged. For social evolution researchers thinking about cumulative adaptation over mesoevolutionary timescales, it makes sense to conceptualize a locally optimal trait, within a specified set of alternatives, as one that leaves no room for further cumulative improvement under δ-weak selection, and it makes sense to conceptualize a 'suboptimal' trait as one that does leave room for such improvement. Since, as I have argued, inclusive fitness provides the appropriate criterion for what constitutes an improvement, it also provides the appropriate standard for optimality. An optimal trait, within a set of alternatives, is one that at least locally maximizes inclusive fitness; a suboptimal trait is one that does not.

On this point, my view to be well aligned with those of Grafen (2014), Gardner (2009, 2017), and West and Gardner (2013), all of whom emphasize the special role of inclusive fitness in explaining 'organism design'. However, in calling inclusive fitness the *criterion* for improvement and the *standard* for optimality, I am not making any empirical claim that we should expect cumulative improvement to occur in any particular case, nor am I suggesting that we should expect optimality, approximate optimality, or anything remotely close to optimality, to be reliably achieved in natural populations. After all, there are numerous well-known obstacles to cumulative improvement: if there is too much dominance or epistasis, if there is too much drift, if the environment changes too rapidly, or if the mutation rate is too high or too low, cumulative adaptation can stall or fail to get off the ground at all.

Many of these obstacles are discussed by Godfrey-Smith (2009), who characterizes a 'paradigm' Darwinian population as one in which cumulative adaptation is apt to occur, and a 'marginal' Darwinian population as one that departs from a paradigm Darwinian population in one or more ways. The conditions for cumulative adaptation highlighted by Godfrey-Smith include the fidelity of transmission, the smoothness of the fitness landscape, and the strength of selection in relation to drift. It is worth noting here that selection can be strong in relation to drift even if selection is δ-weak, although the prospect of advantageous mutations drifting to extinction is particularly concerning under δ-weak selection, as noted by Kimura (1983) and discussed briefly above.

There can be no theoretical guarantee that these obstacles will ever be overcome in nature. When we find a phenotype that is approximately locally optimal among a range of variants, this provides some inductive evidence that the conditions under which the trait evolved were favourable to cumulative optimization (although tests of this sort should be performed and interpreted carefully; see Orzack and Sober 1994; Orzack and Sober 2001). But there can be no mathematical proof that the optimization of inclusive fitness is widespread, likely, or indeed ever instantiated in the natural world.

Here I suspect I part ways with Grafen, who writes, for example, that the theoretical results of his Formal Darwinism project support 'a very general expectation of something close to [inclusive] fitness maximization, which will convert into [inclusive] fitness maximization unless there are particular kinds of circumstances' (Grafen 2014, p. 166). It is, admittedly, hard to discern the empirical commitments of such a hedged claim, and Grafen should not necessarily be read as asserting that the 'particular kinds of circumstances' to which he alludes are rare. However, one gets a general sense that Grafen, like many inclusive fitness theorists, regards cumulative optimization as a process that, on theoretical grounds, we should expect to occur frequently in natural populations, leading in many cases to at least approximately optimal outcomes.

To explain why I am sceptical of this claim, a comment on Grafen's Formal Darwinism project is needed. The aim of the project is to derive mathematical links between a formal representation of natural selection (a version of the Price equation), and a formal representation of phenotypic optimality that uses the apparatus (borrowed from economics) of optimization programs. I find the project admirable and useful but do not think it has proved—or could prove—the above claim, or any empirical claim about the regularity with which we should expect the cumulative optimization of phenotypes to occur in nature.

I have discussed the Formal Darwinism project on several occasions and I will not repeat these discussions here, although two concerns do merit repeating (Birch, 2014b, 2016b, 2018). First, Grafen's framework involves an assumption of 'perfect transmission, that is, no mutation, no gametic selection, fair meiosis, and that all contributing loci have the same mode of inheritance' (Grafen, 2002, p. 77). The assuming away of mutation implies that his model (or 'meta-model') is one in which no cumulative adaptation can occur, so his formal links between gene frequency change and phenotypic optimality are results that are proven to hold only in a world in which no cumulative adaptation is possible (Birch, 2014b, 2016b). Second, Grafen's framework avoids making any assumptions about the genetic architecture underlying phenotypic traits, such as the absence of dominance and epistasis. It follows that his formal results would hold even in a world in which dominance and epistasis were ubiquitous, imposing severe genetic constraints on optimization (Okasha and Paternotte, 2014; Birch 2014b). These two observations should lead us to doubt whether these formal results can be said to support a 'general expectation of something close to inclusive fitness maximization', even in a qualified sense. In fact, they do not imply that inclusive fitness maximization will ever occur.

There is a broader point to be made here: we should not overstate the ability of purely theoretical arguments to support empirical generalizations, no matter how hedged, about natural populations (Orzack, 2014). We can say that if a social adaptation has originated via the gradual accumulation of small-effect mutations under δ-weak selection, then inclusive fitness provides the appropriate criterion for improvement and the appropriate standard for optimality with respect to that adaptation. This points to an important and distinctive role for the concept of inclusive fitness. But we should add the qualification that, in such a case, the distance from optimality of the end product will depend on many different variables, including how readily small mutations arise, how reliably they are inherited, how constant they are in their average effects on the phenotype, how effective selection is at retaining them in the face of drift and other evolutionary processes, and how long the whole process of cumulative improvement has been able to operate in a sufficiently stable environment. These are empirical matters that no amount of theoretical work can settle.

6. Fitness and causality

In closing, I want to connect the foregoing discussion to a family of long-running debates in the philosophy of biology about the relationship between statistics and causality in evolutionary theory—debates in which the rival camps have come to be known as 'statisticalists' and 'causalists'. These debates were initiated in the early 2000s by Matthen and Ariew (2002), and Walsh et al. (2002), who argued that evolutionary theory should be interpreted as a statistical theory rather than a dynamical theory. I say 'debates' because this literature has tended to run together a variety of issues that, while not unrelated, can be usefully distinguished (Otsuka, 2016). These include the relationship between selection and drift; the question of whether natural selection should be regarded as a force, a cause, or merely as a statistical trend; the question of whether evolutionary models provide causal or non-causal explanations; and—most relevantly to the arguments of this paper—the question of whether fitness is a causal property of organisms or merely a statistical predictor of change.[5]

On this last question, the arguments of this paper might be seen as providing some support to the causalist view. Inclusive fitness is, after all, an inherently causal property: a weighted sum of the *effects* a focal individual has on the reproduction of itself and others. It is a property of an individual organism, but not an intrinsic property: it is a property constituted in part by causal relations between the focal organism and its genetic relatives. Brandon and Ramsey (2007) have already noted that social evolution theory is a rich source of problem cases for

[5] Some notable contributions to these debates include Rosenberg and Bouchard (2005); Millstein (2006); Shapiro and Sober (2007); Walsh (2007, 2010); Matthen and Ariew (2009); Lewens (2010b); Otsuka et al. (2011), and Ariew et al. (2015).

those who deny a link between fitness and causality, since many models of social evolution rely on explicitly causal decompositions of fitness. My arguments here reinforce their point, and provide more insight into why causal notions of fitness are so deep-rooted in this area.

Moreover, we can see that Hamilton's decision to formulate inclusive fitness in explicitly causal terms was not an arbitrary or dispensable choice, but rather a way of enabling the concept to play the explanatory role of a criterion for improvement. Recall that a criterion for improvement is charged with reliably telling us the direction of selection-for during successive episodes of δ-weak selection. It should be a property of an individual organism such that all and only those mutant phenotypes that causally promote it, relative to the wild type, are selected-for over the wild type. Since the explanatory role for a criterion for improvement is specified in causal terms ('causally promotes', 'selected-for'), it should not come as a surprise to find that the role is best fulfilled by an explicitly causal fitness concept.

However, I would not interpret this as an unqualified victory for the causalist, because my arguments also undercut the idea of their being any unique fitness concept well suited to all explanatory roles. I have suggested that inclusive fitness earns its keep in social evolution theory as a criterion for phenotypic improvement, despite its limitations as a predictor of gene-frequency change. This is an argument for pluralism about fitness concepts, not an argument for the superiority of inclusive fitness in all contexts. It is compatible with the most versatile and accurate predictors of gene-frequency change being statistical, non-causal properties of genotypes (recall here the example, from Section 1, of the Malthusian parameter)—and these properties have just as strong a claim to the name 'fitness' as inclusive fitness.

Acknowledgements

This is a contribution to a special issue on "Adaptation and Optimality". I thank Nicola Bertoldi for putting the issue together, and I thank Philippe Huneman, Charles Pence, an anonymous referee and audiences at the London School of Economics, the University of Groningen and the University of Pittsburgh for their helpful comments and suggestions.

Funding

This work was supported by a Philip Leverhulme Prize from the Leverhulme Trust.

References

Abbot, P., & 156 others. (2011). Inclusive fitness and eusociality. *Nature*, *471*, E1–E2.

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

Allen B., & Nowak, M. A. (2016). There is no inclusive fitness at the level of the individual. *Current Opinion in Behavioral Sciences*, *12*, 122-128.

Ariew, A., Rice, C., & Rohwer, Y. (2015). Autonomous-statistical explanations and natural selection. *British Journal for the Philosophy of Science*, *66*, 635-658.

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

Birch, J. (2014b). Has Grafen formalized Darwin? *Biology and Philosophy*, *29*, 175-180.

Birch, J. (2016a). Hamilton's two conceptions of social fitness. *Philosophy of Science, 83*, 848-860.

Birch, J. (2016b). Natural selection and the maximization of fitness. *Biological Reviews*, *91*, 712-727.

Birch, J. (2017a). *The philosophy of social evolution*. Oxford: Oxford University Press.

Birch, J. (2017b). The inclusive fitness controversy: finding a way forward. *Royal Society Open Science*, *4*, 170355. http://rsos.royalsocietypublishing.org/content/4/7/170335 (Accessed 17 January 2018).

Birch, J. (2018). Fitness maximization. In R. Joyce, (Ed.), *The Routledge Handbook of Evolution and Philosophy*. London: Routledge, pp. 49-63.

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

Brandon, R. N. & Ramsey, G. (2007). What's wrong with the emergentist statistical interpretation of natural selection and random drift? In D. L. Hull & M. Ruse (Eds), *The Cambridge companion to the philosophy of biology*. Cambridge: Cambridge University Press, pp. 66-84.

Charlesworth, B. (1970). Selection in populations with overlapping generations. I. The use of Malthusian parameters in population genetics. *Theoretical Population Genetics*, *1*, 352-370.

Crow, J. F., & Kimura, M. (1970). *An introduction to population genetics theory*. London: Harper & Row.

Darwin, C. (1859). *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. 1st edition. London: John Murray.

Dawkins, R. (1996). *Climbing Mount Improbable*. New York: W. W. Norton & Company.

Eshel, I., & Feldman, M. W. (1984). Initial increase of new mutants and some continuity properties of ESS in two locus systems. *American Naturalist*, *124*, 631–640.

Eshel, I., & Feldman, M. W. (2001). Optimality and evolutionary stability under short- and long-term selection. In S. H. Orzack & E. Sober (Eds) *Adaptationism and optimality*. Cambridge: Cambridge University Press, pp. 161-190.

Eshel, I., Feldman, M. W., & Bergman, A. (1998). Long-term evolution, short-term evolution and population genetic theory. *Journal of Theoretical Biology*, *191*, 391–396.

Fisher, R. A. (1930). *The genetical theory of natural selection*. Oxford: Clarendon Press.

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

Gardner, A. (2017). The purpose of adaptation. *Interface Focus*, *7*, 20170005.

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

Grafen, A. (1985). A geometrical view of relatedness. *Oxford Surveys in Evolutionary Biology*, *2*, 28-89.

Grafen, A. (2002). A first formal link between the Price equation and an optimization program. *Journal of Theoretical Biology*, *217*, 75-91.

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

Grafen, A. (2014). The Formal Darwinism Project in outline. *Biology & Philosophy*, *29*, 155-174.

Godfrey-Smith, P. (2009). *Darwinian populations and natural selection*. Oxford: Oxford University Press.

Godfrey-Smith, P., & Wilkins, J. F. (2009). Adaptationism and the adaptive landscape. *Biology & Philosophy*, *24*, 199–214.

Hammerstein, P. (1996). Darwinian adaptation, population genetics and the streetcar theory of evolution. *Journal of Mathematical Biology*, *34*, 511–532.

Hamilton, W. D. (1964). The genetical evolution of social behaviour I and II. *Journal of Theoretical Biology*, *7*, 1-52.

Hamilton, W. D. (1970). Selfish and spiteful behaviour in an evolutionary model. *Nature* , *228*, 1218-1220.

Hamilton, W. D. (1996). *Narrow Roads of Gene Land: The Collected Papers of W. D. Hamilton. Vol. 1: Evolution of Social Behaviour*. New York: W. H. Freeman and Company.

Kimura, M. (1983). *The neutral theory of molecular evolution*. Cambridge: Cambridge University Press.

Lewens, T. (2010a). Natural selection then and now. *Biological Reviews*, 85, 829-835.

Lewens, T. (2010b). The natures of selection. *British Journal for the Philosophy of Science*, *61*, 313-333.

Martin, G., & Lenormand, T. (2006) A general multivariate extension of Fisher's geometrical model and the distribution of mutation fitness effects across species. *Evolution*, *60*, 893-907.

Matthen, M., & Ariew, A. (2002). Two ways of thinking about fitness and natural selection. *Journal of Philosophy*, *99*, 55-83.

Matthen, M., & Ariew, A. (2009). Selection and causation. *Philosophy of Science*, *76*, 201-204.

Metz, J. A. J. (2011). Thoughts on the geometry of meso-evolution: collecting mathematical elements for a postmodern synthesis. In F. A. C. C. Chalub and J. F. Rodrigues (Eds), *The Mathematics of Darwin's Legacy*. Basel: Birkhäuser, pp. 193–232.

Mills, S. K., & Beatty, J. H. (1979). The propensity interpretation of fitness. *Philosophy of Science*, *46*, 263-286.

Millstein, R. L. (2006). Natural selection as a population-level causal process. *British Journal for the Philosophy of Science*, 57, 627-653.

Neander, K. (1995). Explaining complex adaptations: a reply to Sober's reply to Neander. *British Journal for the Philosophy of Science*, *46*, 583–587.

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

Okasha, S., Paternotte, C. (2014). Adaptation, fitness and the selection-optimality links. *Biology and Philosophy*, *29*, 225–232.

Orr, H. A. (1998). The population genetics of adaptation: The distribution of factors fixed during adaptive evolution. *Evolution*, *52*, 935-949.

Orzack, S. H. (2014). A commentary on "the Formal Darwinism project": There is no grandeur in this life view of life. *Biology and Philosophy*, *29*, 259-270.

Orzack, S. H., & Sober, E. (1994). Optimality models and the test of adaptationism. *American Naturalist*, *143*, 361–380.

Orzack, S. H., & Sober, E. (Eds) (2001) *Adaptationism and Optimality*. Cambridge: Cambridge University Press.

Otsuka, J. (2016). A critical review of the statisticalist debate. *Biology & Philosophy*, *31*, 459-482.

Otsuka, J., Turner, T., Allen, C., & Lloyd, E. (2011). Why the causal view of fitness survives. *Philosophy of Science*, *78*, 209-224.

Pence, C., & Ramsey, G. (2015). Is organismic fitness at the basis of evolutionary theory? *Philosophy of Science, 82*, 1081-1091.

Queller, D. C. (1985). Kinship, reciprocity, and synergism in the evolution of social behaviour. *Nature*, *318*, 366-367.

Rosenberg, A., & Bouchard, F. (2005). Matthen and Ariew's obituary for fitness: Reports of its death have been greatly exaggerated. *Biology & Philosophy*, *20*, 343-353.

Shapiro, L., & Sober, E. (2007). Epiphenomenalism: The do's and the don'ts. In G. Wolters and P. K. Machamer (Eds), *Thinking about causes: From Greek philosophy to modern physics*. Pittsburgh, PA: University of Pittsburgh Press, pp. 235-264.

Sober, E. (1984). *The nature of selection: Evolutionary theory in philosophical focus*. Chicago, IL: University of Chicago Press.

Sober, E. (2001). The two faces of fitness. In R. S. Singh, C. B. Krimbas, D. B. Paul, J. Beatty (Eds) *Thinking about evolution: Historical, philosophical and political perspectives*. Cambridge: Cambridge University Press, pp. 309-321.

Walsh, D. M. (2007). The pomp of superfluous causes: The interpretation of evolutionary theory. *Philosophy of Science*, *74*, 281-303.

Walsh, D. M. (2010). Not a sure thing: Fitness, probability, and causation. *Philosophy of Science*, *77*, 147-171.

Walsh, D. M., Lewens, T., & Ariew, A. (2002). The trials of life: Natural selection and random drift. *Philosophy of Science*, 69, 452-473.

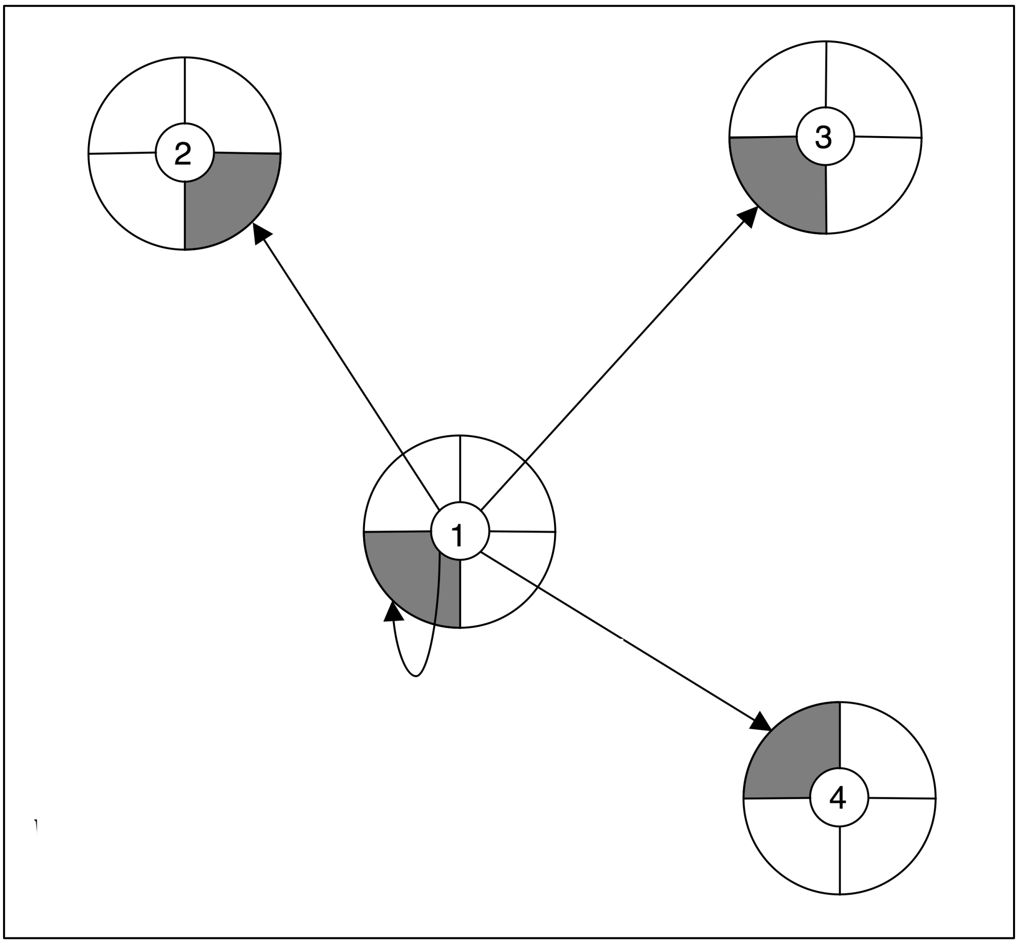
Waxman, D., & Welch, J. J. (2005). Fisher's microscope and Haldan's ellipse. *American Naturalist*, *166*, 447-457.

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

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

Wilson, E. O. (2012). *The social conquest of Earth*. New York: Liveright.

FIGURE 1:



CAPTION: *Inclusive fitness*. An individual organism's inclusive fitness is a weighted sum of the effects of its behaviour on reproductive success. In this illustration, organism 1's behaviour affects the reproductive success of itself and of organisms 2, 3, and 4 (as shown by the arrows; the shaded regions represent components of reproductive success caused by the behaviour of organism 1). Organism 1's inclusive fitness consists of a baseline non-social component, plus the effect on its own reproductive success caused by its own behaviour, plus its effects on organisms 2, 3, and 4, weighted in each case by a coefficient of relatedness (Figure © the author, reprinted from Birch, 2017a).