Fitness maximization

Jonathan Birch

Adaptationist approaches in evolutionary ecology often take it for granted that natural selection maximizes fitness. Consider, for example, the following quotations from standard textbooks:

> The majority of analyses of life history evolution considered in this book are predicated on two assumptions: (1) natural selection maximizes some measure of fitness, and (2) there exist trade-offs that limit the set of possible [character] combinations. (Roff 1992: 393)

> The second assumption critical to behavioral ecology is that the behavior studied is adaptive, that is, that natural selection maximizes fitness within the constraints that may be acting on the animal. (Dodson et al. 1998: 204)

> Individuals should be designed by natural selection to maximize their fitness. This idea can be used as a basis to formulate optimality models [...]. (Davies et al. 2012: 81)

Yet there is a long history of scepticism about this idea in population genetics. As A. W. F. Edwards puts it:

> [A] naive description of evolution [by natural selection] as a process that tends to increase fitness is misleading in general, and hill-climbing metaphors are too crude to encompass the complexities of Mendelian segregation and other biological phenomena. (Edwards 2007: 353)

Is there any way to reconcile the adaptationist's image of natural selection as an engine of optimality with the more complex image of its dynamics we get from population genetics? This has long been an important strand in the controversy surrounding adaptationism.[1] Yet debate here has been hampered by a tendency to conflate various different ways of thinking about maximization and what it entails. In this article I distinguish, at a deliberately coarse grain of analysis, four varieties of

[1] For excellent introductions to these wider debates, see Lewens (2007, 2009); Godfrey-Smith and Wilkins (2008), and Orzack and Forber (2012).

maximization principle.[2] I then discuss the logical relations between these varieties, arguing that, although they may seem similar at face value, none entails any of the others. I then turn briefly to the status of each variety, arguing that, while each type of maximization principle faces serious problems, the problems are subtly different for each type.

In the last section, I reflect on what is at stake in this debate. Defenders of fitness maximization are often motivated by a desire to defend adaptationist, optimality-based approaches in evolutionary ecology of the sort described in the quotations at the start of this article. I argue, however, that the value of optimality-based approaches as tools for hypothesis generation does not depend on the existence of a universal maximization principle describing the action of natural selection. The need for such a principle only arises for those who hold a more epistemically ambitious view about what these approaches can achieve.

1 Four varieties of maximization

Any maximization principle, to be worthy of the name, must spell out what is meant by a fitness maximum, and must assign a special status to such a point in the dynamics of evolution by natural selection. This, however, leaves many options open regarding the nature of the maximum and its significance in the dynamics. We should not be surprised, then, to find many quite different fitness maximization principles in evolutionary biology.

I suggest that two distinctions lead to a useful taxonomy of such principles. First, we should distinguish between maximization principles that concern *what happens at equilibrium* and those that concern *the direction of change*. Second, we should distinguish between maximization principles that concern the *population mean fitness* and those that concern the *behavioural strategies of individual organisms*.

As a preliminary, I want to introduce Sewall Wright's (1932) adaptive landscape metaphor, to which Edwards alludes in the above quotation. This controversial metaphor looms large in debates about fitness maximization. Wright imagined the mean fitness of a population moving through a multi-dimensional gene frequency

[2] The basic taxonomy here is set out in greater detail in Birch (2016).

space.[3] Flattening this space to three dimensions for ease of visualization, he pictured a landscape characterized by "adaptive peaks" representing mean fitness maxima, and he pictured evolution by natural selection as a "hill-climbing" process that drives a population towards the nearest maximum. In this vision of evolution, natural selection sometimes drives populations to the highest peak (the global maximum) but it may also cause populations to become marooned on local maxima, separated from the global maximum by fitness valleys.

The adaptive landscape metaphor combines two seductive ideas about the dynamics of evolution by natural selection: an idea about equilibrium and an idea about change. First, it pictures the stationary points of evolution by natural selection as points at which mean fitness is maximized, such that any change in the frequency of any allele will decrease mean fitness. Second, it pictures a population out of population-genetic equilibrium as moving reliably upward, in the direction of greater mean fitness.

These two claims are conceptually distinct. To help us keep these ideas separate, let us denote them with the labels "MAX-A" and "MAX-B":

> **MAX-A (*Mean fitness, equilibrium*)**: A population undergoing evolution by natural selection is at a stable population-genetic equilibrium if and only if its mean fitness is maximized, such that any change in allele frequencies will reduce mean fitness.

> **MAX-B (*Mean fitness, change*)**: If a population is not in population-genetic equilibrium, then natural selection will reliably change allele frequencies in a way that leads to greater mean fitness, even if other factors prevent the population from reaching a maximum.

---

[3] Wright originally envisaged "genotypes [...] packed, side by side [...] in such a way that each is surrounded by genotypes that differ by only one gene replacement" (Wright 1988: 116). On such a landscape, populations would be represented by clouds of genotypes. But the version of the metaphor that now features in standard textbooks represents a population as single point moving through a space defined by population gene frequencies (Ridley 2004; Futuyma 2013). See Pigliucci and Kaplan (2006); Kaplan (2008) for discussion of the different versions of the metaphor.

In both MAX-A and MAX-B, the variable that is maximized is the population mean, averaged over genotypes or over individuals, of some fitness measure. In this sense, MAX-A and MAX-B are population-centred: they focus on the properties and dynamics of populations, making no explicit reference to the properties of individuals in those populations. But this is not the only way to think about fitness maximization. Behavioural ecologists commonly start with the assumption that an individual organism will behave as if attempting to maximize its own individual fitness or (in the case of social behaviour) its inclusive fitness. They then ask: which strategy, from the range of feasible options, would it be rational for the organism to adopt, given its apparent goal?

We can say (following Alan Grafen) that behavioural ecologists who think in this way are employing an "individual as maximizing agent" analogy (Grafen 1984, 1999). Agential thinking of this sort is widespread in many areas of evolutionary ecology, including inclusive fitness theory, life history theory and evolutionary game theory (e.g. Maynard Smith 1982; Parker and Smith 1990; Davies et al. 2012). The analogy does not involve any literal attribution of rational agency to non-human organisms. Instead, the thought is that organisms, regardless of their degree of cognitive sophistication, can be modelled *as if* they were rational agents attempting to maximize their individual fitness (or inclusive fitness), because natural selection tends to lead to equilibria at which organisms adopt strategies that maximize their individual fitness (or inclusive fitness) within the set of feasible options. This leads to a third conception of fitness maximization:

> **MAX-C (*Individual fitness, equilibrium)***: A population undergoing evolution by natural selection is at a stable population-genetic equilibrium if and only if all organisms adopt the phenotype that maximizes their individual fitness (or inclusive fitness) within the set of biologically feasible phenotypic options.

This notion of maximization clearly bears some resemblance to MAX-A, in that it posits a close relationship between population-genetic equilibria and fitness maxima, but it differs in that it defines these maxima not in terms of the mean fitness of the population, but rather in terms of optimal strategy choice, within the set of

biologically feasible options, on the part of individual organisms (the reference to "biologically feasible options" makes it clear that we are talking here about optimization *subject to constraints*, not unconstrained maximization). Despite the superficial similarities, this way of thinking about maximization has little to do with Wright's adaptive landscape metaphor. It is much closer to the notion of maximization which appears in economics, in which humans are typically modelled as rational agents maximizing utility subject to constraints.

MAX-C, like MAX-A, is a claim about what happens at equilibrium. However, the equilibrium/change distinction cross-cuts the mean fitness/individual fitness distinction. This leads to our fourth variety, an individual-level analogue of MAX-B concerning the direction of change:

> **MAX-D (*Individual fitness, change*)**: If a population is not in population-genetic equilibrium, then natural selection will reliably drive it in the direction of a point at which all organisms adopt the phenotype that maximizes their individual fitness (or inclusive fitness) within the set of biologically feasible phenotypic options, even if other factors prevent the population from reaching this point.

2 Relations between the varieties

We now have four varieties of fitness maximization on the table (Table 1). I claim that none of them entails any of others. I will defend this claim piecemeal, looking first at the rows in Table 1 and then at the columns. I assume that if there is no entailment along the rows or the columns, then there is no serious prospect of entailment across the diagonals.

	Equilibrium	Change
Mean fitness	**MAX-A**	**MAX-B**
Individual fitness	**MAX-C**	**MAX-D**

Table 1: Four varieties of fitness-maximization

The first non-entailment I want to consider concerns the first row. The key points here can be expressed in terms of the adaptive landscape metaphor. In principle, it might be that adaptive peaks are always stationary points and yet selection might be ineffectual at driving populations up slopes towards them. Conversely, selection might drive populations reliably upward whenever they are out of equilibrium, and yet the population might stably stop at least some of the time at points that are not peaks. Hence MAX-A does not entail MAX-B, nor vice versa.

The broader point here is that claims about what happens at equilibrium do not entail claims about the direction of out-of-equilibrium change, nor vice versa. This carries over to the second row. In principle, it might be that a stable stationary point in the dynamics of evolution by natural selection occurs if and only if all organisms in the population have optimal phenotypes, and yet natural selection is ineffectual at driving populations towards such optima. Conversely, selection might reliably drive populations towards such optima, only to reach a stable stationary point part way there. Hence MAX-C does not entail MAX-D, nor vice versa.

The columns are a little more subtle. MAX-A does not entail MAX-C, nor vice versa, because there can be mean fitness maxima (in allele frequency space) at which suboptimal phenotypes are present in the population. MAX-A says that these points constitute stable population-genetic equilibria, whereas MAX-C says they do not. Consider, for example, the polymorphic equilibrium in the standard model of heterozygote advantage, illustrated by the famous case of sickle-cell anaemia and malarial resistance. In regions with a high incidence of malaria, an allele that causes sickle-cell anaemia in the homozygote (i.e. the genotype with two copies of the allele) is nonetheless present at a low frequency at equilibrium because it causes malarial resistance in the heterozygote (i.e. the genotype with one copy). In the standard model of this situation, the equilibrium is a mean fitness maximum—any change in allele frequencies lowers the mean fitness—but it is not a point at which every organism has an optimal phenotype within the range of feasible options (Hedrick 2011).

This suggests that the relationship between MAX-A and MAX-C, far from being one of logical entailment, is actually one of logical incompatibility: they imply contradictory claims about the status of mean fitness maxima at which suboptimal phenotypes are present. However, MAX-A and MAX-C can be made compatible if interpreted as claims about different evolutionary timescales. MAX-A-type

maximization principles have usually been studied and discussed in the context of models of short-term "microevolution", such as the heterozygote advantage model discussed above. Yet when applying the "individual as maximizing agent" analogy, evolutionary ecologists often have a longer timescale in mind: the timescale of what Hans Metz (2011) and Peter Godfrey-Smith (2012) have called "mesoevolution." The idea here is that we should think of the attainment of phenotypic optimality as occurring over a timescale long enough for populations to escape short-term equilibria, such as the sickle-cell equilibrium, at which suboptimal phenotypes may be present. This move is central to Peter Hammerstein's (1996) "streetcar theory," which I consider below. For now, I simply want to note that MAX-A, read as a claim about the equilibria of short-term microevolution, is logically independent of MAX-C, read as a claim about the "mesoevolutionary" long run.

The broader point here is that there is a logical gap between claims about short-term changes in gene frequency and claims about longer-term phenotypic evolution (cf. Wilkins and Godfrey-Smith 2009). This carries over to the second column. Read as claims about the direction of short-term change, MAX-B and MAX-D seem to disagree about what will happen in cases in which a population stands to increase its mean fitness by reducing the frequency of an optimal phenotype. We see this in the sickle-cell model, in which an initially high frequency of malarial resistance is reduced by selection, owing to the adverse fitness consequences of the same gene in the homozygote. Mean fitness increases, but there is no convergence on universal malarial resistance.

As with MAX-A and MAX-C, however, thinking about timescales can help remove this apparent tension. We can read MAX-D as the claim that *over the long term* the dynamics of a population evolving by natural selection will converge on a point at which the population realizes an optimal phenotypic profile. This claim about long-term convergence is logically independent of MAX-B, read as a claim about the short-term direction of change. It is compatible with selection reliably driving a population in the direction of greater mean fitness in the short term, even if this sometimes means driving it away from phenotypic optimality, provided the population converges on phenotypic optimality in the long run. It is also compatible with the direction of short-term change in mean fitness being highly variable and context-dependent.

3 Status of the varieties: MAX-A and MAX-B

All four varieties of fitness maximization are controversial, but for different reasons. Let us start with MAX-A and MAX-B. While these may look innocuous to biologists trained to think of evolution in terms of adaptive landscapes, they are contentious in population genetics (Ewens 2004; Edwards 2007). MAX-A is challenged by models in which evolution stops at a point that, on any reasonable measure of fitness, is not a mean fitness maximum, even though natural selection is the only evolutionary process at work. Meanwhile, MAX-B is challenged by models in which, on any reasonable measure of fitness, natural selection drives the mean fitness of a population downwards over time.

Models of both sorts have a long history in population genetics. In one-locus models that satisfy various other assumptions (random mating, frequency-independent fitness, selection on viability differences only), the mean fitness does reliably increase and stable equilibria do correspond to mean fitness maxima (Scheuer and Mandel 1959; Mulholland and Smith 1959; Edwards 2000). But relax any of the assumptions of these models and the result is no longer valid. A standard citation in this context is Moran (1964), who constructed a two-locus model in which mean fitness decreases over time, and in which population-genetic equilibrium occurs far from any "adaptive peak." Moran took this result to debunk the very idea of an "adaptive topography." Ewens (1968) and Karlin (1975) reinforced Moran's conclusions with further results along similar lines. The overall message of this work is that both MAX-A and MAX-B are extremely dubious in the multi-locus case (see also Hammerstein 1996; Eshel et al. 1998; Ewens 2004).

Intuitively, the source of the trouble in multi-locus models is that Mendelian segregation, recombination and epistasis complicate the transmission of fitness between parents and offspring. Offspring, while resembling their parents on the whole, inherit a combination of genes that is not a simple replica of either parent. Consequently, a gene that promotes the fitness of a parent can, on finding itself in a new genomic context, detract from the fitness of the offspring by whom it is inherited, with adverse consequences for the population mean fitness. Unfortunately, natural selection only "sees" whether current bearers of an allele are fitter, on average, than non-bearers; it does not "see" what the mean population fitness will be after the vagaries of Mendelian inheritance have taken their course.

In the models referenced above, the fitness of a genotype is assumed to be independent of population gene frequencies. Matters are even worse for mean fitness maximization when we introduce frequency-dependent genotypic fitness. Here, the intuitive problem is that frequency-dependence makes it possible for an allele to be selected even when an increase in its frequency would, via knock-on effects on genotypic fitness values in the next generation, detract from the mean fitness of the population. The moral of over fifty years of work in this area is that, when genotypic fitness depends on gene frequency, the mean fitness does not reliably increase and is rarely maximized at equilibrium. Indeed, in an early treatment of frequency-dependence, Sacks (1967) showed that frequency-dependent selection can lead to a stable equilibrium that is also a fitness minimum. This point has been underlined by recent work in the field of adaptive dynamics, which suggests an important role for fitness-minimization in evolution. The idea is that mean fitness minima act as "evolutionary branching points" at which a population fragments, causing different subpopulations to pursue divergent evolutionary trajectories (Geritz and Metz 1998; Doebeli and Dieckmann 2000; Doebeli 2011).

To be clear, the problem these models pose for MAX-A is not simply that the population stops at a local maximum rather than finding its way to the global maximum. The problem is that the population stops at a point that is *not a maximum at all*, whether local or global. If we insist on employing the "adaptive landscape" metaphor in such cases, we should say that the stopping point lies on a "slope" or in a "valley" rather than on a "peak." Likewise, note that the problem these models pose for MAX-B is not simply that the "uphill push" of natural selection is counteracted by other causes of gene frequency change. The problem is that, even when there is no cause of gene frequency change other than natural selection, the mean fitness still decreases.

4 Fisher's fundamental theorem

From Wright onwards, defenders of MAX-B have often cited R. A. Fisher's fundamental theorem of natural selection (Fisher 1930, 1941) in support of their claims, even though Fisher himself never regarded the theorem as a maximization principle (Edwards 1994). The theorem states that the rate of change in the mean fitness in a population "ascribable to a change in gene frequency" is equal to the

additive genetic variance in fitness. Although there has long been uncertainty over its mathematical validity, later reconstructions show clearly that it is a correct result, given a particular interpretation of what Fisher meant by the rate of change "ascribable to a change in gene frequency" (Price 1972; Ewens 1989; Lessard 1997). Since variance cannot be negative, the theorem seems at first glance to imply that the rate of change in mean fitness cannot be negative either, apparently contradicting the results Moran and others have obtained in specific models.

A lot depends, however, on what is packed into Fisher's rather obscure concept of a rate of change "ascribable to a change in gene frequency." In informal terms, the quantity that Fisher proved can never be negative is a quantity that captures what the total rate of change in mean fitness *would* be, *if* we could hold the average effects of alleles on fitness at their current values as natural selection changes their frequencies. The trouble is that, except in cases of perfectly additive genetics (no dominance, epistasis or linkage), the average effects of alleles depend on genotype frequencies, and therefore on allele frequencies, and therefore on the action of natural selection. So as natural selection changes allele frequencies, it changes the average effects of alleles, creating a gap between the total rate of change in mean fitness and the "partial" rate of change with which Fisher's fundamental theorem is concerned.

There is in fact no theoretical guarantee that the total rate of change in mean fitness will be non-negative. To use a potentially misleading metaphor, the picture we get from the fundamental theorem, when we interpret it correctly, is of natural selection pushing the population "uphill" with one hand while it reshapes the landscape with the other. The total action of natural selection may leave the population higher, lower or at the same level, depending on the details. Of course, as Moran (1964) pointed out, this arguably casts doubt on the utility of the adaptive landscape metaphor.[4]

[4] See Price 1972; Ewens 1989; Frank and Slatkin 1992; Frank 1997; Edwards 1994; Ewens 2004; Plutynski 2006; Okasha 2008; Ewens 2011; Edwards 2014; Grafen 2015; Ewens and Lessard 2015; Birch 2016 for further detail on, and discussion of, these complex issues.

5 Status of the varieties: MAX-C and MAX-D

MAX-C-type maximization principles, which switch the focus from the population mean fitness to individual phenotypes and their fitness consequences, have two main cards up their sleeve to help them deal with the traditional problem cases for MAX-A and MAX-B. First, in cases of strategic interaction, an equilibrium that is not a mean fitness maximum can still be reconciled with MAX-C, as long as it is a Nash equilibrium. For, at a Nash equilibrium, organisms are best-response maximizers: they adopt the phenotype (or a phenotype, in cases of weak Nash equilibrium) that is fitness-optimal conditional on the phenotypes of their social partners. This is true even if the Nash equilibrium is a mean fitness minimum.

Second, polymorphic equilibria in which one of the phenotypes present is clearly suboptimal, such as the sickle-cell equilibrium, can be reconciled with MAX-C provided MAX-C is understood as a claim about the stable equilibria of long-term phenotypic evolution, not the stable equilibria of short-term gene frequency change. The key here is to adopt a particularly demanding conception of stability when defining a stable equilibrium of long-term phenotypic evolution, so that sickle-cell type polymorphic equilibria do not qualify as stable. Crucially, the sickle-cell equilibrium is vulnerable to invasion by a mutant that produces malarial resistance in the heterozygote without producing sickle-cell anaemia in the homozygote. So if we define stability in terms of resistance to invasion in the long run, this equilibrium may not be stable after all.

Peter Hammerstein's (1996) "streetcar theory" has been particularly influential in this context (see also Eshel and Feldman 1984, 2001; Liberman 1988; Hammerstein and Selten 1994; Eshel et al. 1998; Hammerstein 2012). On Hammerstein's picture, "an evolving population resembles a streetcar in the sense that it may reach several temporary stops that depend strongly on genetic detail before it reaches a final stop which has higher stability properties and is mainly determined by selective forces at the phenotypic level" (Hammerstein 1996: 512). The "final stop," he argues, will be a Nash equilibrium. Hence we arrive at a tenable version of MAX-C, provided we interpret "stable" equilibria as only those which correspond to Hammerstein's "final

stops" achieved in the evolutionary long run, as opposed to the "stops along the way" described by standard microevolutionary theory.[5]

Hammerstein's argument, however, does not establish (nor attempt to establish) MAX-D: it characterizes a special sort of long-term stable equilibrium and shows that it corresponds to a fitness maximum in a certain sense, but it does not give us a reason to think that a population evolving by natural selection will reliably converge towards such a point. This is ultimately an empirical matter, because it depends on the rate of mutation and the rate at which the selective environment changes. As Ilan Eshel and Marcus Feldman note, arguments of this general sort predict optimal outcomes in the long run only if "the regime of selection acting on the trait under study remains invariant during the slow process of transitions between genetic [i.e. short-term] equilibria" (Eshel and Feldman 2001: 186). By contrast, "for shorter-lived processes of conflict (e.g. in a newly colonized niche) we expect the population to be close to a short-term stable equilibrium, but not to one that is long-term stable" (Eshel and Feldman 2001: 186).

In light of this, it also seems clear that the streetcar theory, although it does provide support for a specific, long-term version of MAX-C, does not support the idea that natural selection has any tendency to maximize fitness in the absence of other causes of gene frequency change. On the contrary, the argument concedes that natural selection often will not be able to do so unless another cause of gene frequency change, viz. mutation, is powerful enough to circumvent genetic barriers to optimality (cf. Sober 1987). So, to the extent that the streetcar theory supports a version of adaptationism, it is a version that recognizes the importance of both mutation and selection in determining evolutionary outcomes.

6 Formal Darwinism

This is where Alan Grafen's on-going "Formal Darwinism" project enters the scene (Grafen 2002, 2006, 2007, 2014). Grafen aims to show that, even in models in which

[5] This argument does not, however, give us a tenable version of MAX-A, since a Nash equilibrium need not be a mean fitness maximum.

we assume the absence of mutation[6], there are strong formal links between population genetic equilibrium and phenotypic optimality, where the optimal phenotype is defined as that which maximizes inclusive fitness within a set (*X*) of specified alternative options.

The assuming away of mutation in Grafen's models marks one important difference with Hammerstein's project. The other notable difference is that Grafen's formal links concern the direction of short-term change as well as the nature of long-term equilibrium. In broad terms, what Grafen has shown is that, across a wide range of (mutation-free) models, a population is at a point at which there is no "scope for selection" (roughly, no expected change in any gene frequency) and no "potential for positive selection" (roughly, no phenotype in *X* that is selected-for or that would be selected-for if present) if and only if all organisms have the optimal phenotype in *X*. He also proves links (which I will not discuss here) concerning changes in gene frequency in populations in which some or all individuals are suboptimal. Grafen (2014: 166) glosses these results as showing that "there is a very general expectation of something close to fitness maximization, which will convert into fitness maximization unless there are particular kinds of circumstances." [7]

---

[6] The key assumptions of Grafen's framework are that there is "no mutation, no gametic selection, fair meiosis and that all the loci contributing to the p-score have the same mode of inheritance" (Grafen 2002: 82). I previously described the absence of mutation as a "limitation" of the framework (Birch 2016), but I now suspect that this not the right way to think about it. A more charitable reading is that Grafen intentionally assumes away mutation in the hope of proving links between selection and optimality that do not rely on assumptions about mutation, as Hammerstein's (1996) results do.

[7] Others cite Grafen in support of stronger claims. See, for example, West and Burton-Chellew (2013: 1043): "The success of the behavioral ecology approach is built on an extremely solid theoretical grounding (Davies et al. 2012). Darwin (1859) argued that traits that increase fitness will accumulate in populations, leading to organisms that behave as if they are trying to maximize their fitness. Our modern most general genetical interpretation of this is that organisms should behave as if they are trying to maximize their inclusive fitness (Hamilton 1964; Grafen 2006)."

I have criticized the Formal Darwinism project on other occasions, and I cannot do justice to this complex topic here (Birch 2014, 2016). I will, however, explain briefly why I think that some of Grafen's informal glosses, such as that in the above quotation, overstate the implications of his formal results for fitness maximization.

We should first ask: which varieties of maximization are at stake? The maximization in which Grafen is interested is the maximization of individual fitness by individual phenotypes: it does not directly involve population means. In effect, he claims to have shown that versions of MAX-C and MAX-D *would* be true in a world without mutation (and in which various other idealizations he makes in his models, such as the absence of meiotic drive and gametic selection, also obtain). Here I will focus on MAX-C.[8]

A natural reaction to this claim is to ask: how could MAX-C possibly be true in a world without mutation? Assuming away mutation seems to make things worse, not better, for fitness maximization. For in a world without mutation, there is no way to get around the constraints imposed by genetic architecture. A population can get permanently stuck at a sickle-cell type polymorphic equilibrium at which suboptimal phenotypes are present. Yet Grafen proves that all of his formal links between gene frequency change and optimal strategy choice still hold in such a context. This is surprising at face value, and it leaves two possibilities: either these cases are not really incompatible with MAX-C after all, despite the apparent presence of suboptimal phenotypes, or else Grafen's formal links do not really imply a version of MAX-C after all, even though his informal gloss suggests they do.

It takes a bit of untangling to see what is going on here (Grafen 2014; Okasha and Paternotte 2014; Birch 2016). The key is to see that Grafen's links do not explicitly refer to population-genetic equilibrium: instead, they characterize an equilibrium as a point at which there is no "scope for selection" and no "potential for positive selection". It turns out that the sickle-cell equilibrium does not qualify as an equilibrium in this sense, because there is a phenotype—malarial resistance—that is being selected-for. By characterizing evolutionary equilibrium in partly phenotypic terms, Grafen is able to disqualify equilibria in which gene frequencies are stably constant but suboptimal phenotypes are present.

[8] Similar considerations complicate the relationship between Grafen's links and MAX-D, though I will not discuss this issue here.

However, this unorthodox way of thinking about equilibrium has some odd consequences. For example, an initial population composed of 100% heterozygotes, all with the optimal malarial resistance phenotype, qualifies as an equilibrium in the sense that matters for Grafen's links. It qualifies because it has no expected change in gene frequencies in the initial time step and no phenotype that is or would be selected-for, even though selection will inevitably start altering gene frequencies as soon as homozygotes appear (Grafen 2014).

As Grafen himself notes, the way in which the links hold in cases of heterozygote advantage "seem[s] to contain an element of evasion, and call[s] into question the meaning and value of the links themselves" (Grafen 2014: 165). The question is what this means for the relationship between the links and our MAX-C. Here is one way to go: MAX-C is clearly false in sickle-cell type models without mutation, but Grafen's links are true; so, despite appearing at face value to do so, Grafen's links do not imply MAX-C. This is the response I advocated on an earlier occasion (Birch 2016).

However, there is, I think, is another way of reading this: a way more sympathetic to Grafen's aims. This is to say that Grafen, like Hammerstein, has found a way of constructing a non-standard equilibrium concept so that equilibria at which suboptimal phenotypes are present do not qualify as equilibria. The novelty of Grafen's approach is to appeal to phenotypic considerations in constructing the equilibrium concept, where Hammerstein appeals to assumptions about the rate of mutation and the long-run malleability of genetic architectures. If we formulate MAX-C using Grafen's non-standard equilibrium concept, then it comes out true (see "MAX-C**" in Birch 2016). What remains up for debate is whether evolution by natural selection has any reliable tendency to arrive at equilibria, thus construed.

7 Living without maximization

For both Hammerstein and Grafen, the project of pursuing fitness maximization principles, in the face of widespread skepticism from population geneticists, is justified by the need to provide a theoretical foundation for adaptationist, optimality-based approaches in behavioural ecology (and evolutionary ecology more generally). The same need is clearly felt by those behavioural ecologists who have pounced on Grafen's links (too hastily, in my view) as providing "an extremely solid theoretical grounding" for the field (West and Burton-Chellew 2013: 1043).

What drives this need? Why can't behavioural ecologists simply accept the message from population genetics that the dynamics of natural selection are messy and complicated, and revise their models accordingly? The problem with this suggestion is that the vast majority of work in behavioural ecology relies on what Grafen (1984) has termed "the phenotypic gambit": the bet that the evolution of complex phenotypes can be understood in ignorance of the complex genetic architectures that underlie them. Approaches as diverse as inclusive fitness theory, life history theory, multi-level selection theory and evolutionary game theory all have this much in common.

The precise nature of the gambit varies depending on the details of the approach: for example, the inclusive fitness approach aims to understand the evolution of a trait by looking at its fitness effects on an organism and its social partners, the patterns of genetic relatedness between social partners, and the trait's heritability. Maximization-based techniques are often employed, but need not be. In virtually all cases, however, researchers make a fundamental bet that they can explain evolutionary outcomes without detailed knowledge of the genotype-phenotype map. The rationale for this bet is a practical one. We may be living in a "post-genomic" age, but, for the vast majority of traits in the vast majority of species, we still lack the sort of data concerning the genetic architectures underlying complex behaviour that ecologists would need in order to do without the phenotypic gambit. This is what drives the desire to show that the long-term equilibria of the evolutionary process are governed by, in Hammerstein's words, "selective forces at the phenotypic level," which can be understood in the absence of detailed knowledge of genetics.

There is, I think, a real danger that this holy grail of foundational work in behavioural ecology will prove mythical. The early models of Moran and others should already be enough to convince us that there can be no purely theoretical guarantee that evolutionary equilibria will be fitness maxima. This inevitably depends on the ability of mutation to alter genetic arrangements. There may be nothing further to say here except that sometimes this happens and sometimes it doesn't. In some cases, the genetic architecture underlying a trait will preclude its optimization; sometimes it will be favourable. Sometimes an unfavourable architecture will be made more favourable by a change in the genetics; in other cases it may persist longer than the selective environment. It all depends on the details.

I suggest, however, that we can make peace with the phenotypic gambit without having to deny the dependence of real-world evolutionary outcomes on genetic detail. The key is to recognize that, while the gambit really is a gambit—an opening bet—and not a "solid theoretical grounding" for which we have compelling independent evidence, it not always problematic to rest a scientific research programme on a bet of this sort. This depends on the epistemic ambitions of the programme. If optimality modelling aims to yield, by itself, knowledge of the evolutionary processes that have shaped phenotypic traits, then its reliance on a bet is indeed a problem. For this suggests that, even when the hypotheses it generates are true, they are only luckily true (i.e., true because the assumptions of the phenotypic gambit happened to be true in this case), and this undermines the idea that they constitute knowledge.

However, if we see the goal of optimality-based approaches as primarily one of hypothesis generation, the reliance on a bet is unproblematic.[9] After all, it is a bet that has led consistently to the generation of serious and credible evolutionary hypotheses—hypotheses that are plausible given everything we currently know. This is not a trivial achievement. The phenotypic gambit, in all its forms, represents a very well designed heuristic for this purpose. On the one hand, it permits modellers to idealize away potential complications about which they are unavoidably ignorant, while, on the other hand, it demands sensitivity to the knowable empirical facts about fitness effects, population structure, heritabilities, coefficients of relatedness and so on.

The upshot is that whether optimality modellers should be worried about the absence of a theoretical justification for fitness maximization depends on the function they intend their models to serve. The lack of such a justification challenges more epistemically ambitious claims about their function, but it does not undermine their value as sources of credible empirical hypotheses: hypotheses that should not be

[9] Alexandrova (2008) argues that we should understand models in experimental economics as tools for hypothesis generation and constructs a detailed account of how this works. Roughly, the idea is that models provide "open formulae" for causal hypotheses: they generate schemas for causal hypotheses that do not assert anything until we add either a quantifier or a singular instance. I suspect this sort of account would fit many optimality models in evolutionary ecology quite well, but I do not pursue this in detail here.

regarded as knowledge until the underlying genetic architecture of the trait in question—and its compatibility or otherwise with the hypothesis—is known.[10]

I suspect many evolutionary ecologists would want to resist this epistemically modest conception of the function of optimality modelling. But I think we can embrace it while still recognizing the scientific value of this kind of work. Serious and credible evolutionary hypotheses are hard to find, and we should not be dismissive of methodological approaches that have consistently generated them.

Acknowledgements

I thank Anthony Edwards, Warren Ewens, Alan Grafen, Rufus Johnstone, Tim Lewens, Samir Okasha, Steven Orzack, Cedric Paternotte, and John Welch for very helpful discussions and/or email exchanges on these issues. I thank Wiley-Blackwell for permitting the re-use of some material from Birch, J. (2016) "Natural selection and the maximization of fitness," *Biological Reviews*, © 2016 Cambridge Philosophical Society. The present article is intended as a companion piece to this longer article, in which Fisher's fundamental theorem and Formal Darwinism are discussed in greater mathematical detail. This work was supported by a Philip Leverhulme Prize from the Leverhulme Trust.

[10] In a similar vein, Potochnik (2009) distinguishes strong and weak uses of optimality models, where the "strong use" involves the claim that selection was the only important influence on the evolution of the trait, and the "weak use" involves the weaker claim that model accurately represents the role played by selection in the evolution of the trait. However, even Potochnik's "weak use" strikes me as an epistemically ambitious use, since it relies on the idea that optimality models "accurately represent the selection dynamics involved in producing the target evolutionary outcome" (Potochnik 2009: 187). My proposal is more akin to the "even weaker" use that Potochnik attributes to Seger and Stubblefield (1996).

References

Alexandrova, A. (2008). Making models count. *Philosophy of Science*, 75(3):383–404.

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

Birch, J. (2016). Natural selection and the maximization of fitness. *Biological Reviews*. doi: 10.1111/brv.12190

Darwin, C. R. (1859). *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. John Murray, London, 1st edition.

Davies, N. B., Krebs, J. R., and West, S. A. (2012). *An Introduction to Behavioural Ecology*. Wiley-Blackwell, Hoboken, NJ, 4th edition.

Dodson, S. I., Allen, T. F. H., Carpenter, S. R., Ives, A. R., Jeanne, R. L., Kitchell, J. F., and Langston, N. E. (1998). *Ecology*. Oxford University Press, Oxford.

Doebeli, M. (2011). *Adaptive diversification*. Princeton University Press, Princeton, NJ.

Doebeli, M. and Dieckmann, U. (2000). Evolutionary branching and sympatric speciation caused by different types of ecological interactions. *American Naturalist*, 156:S77–S101.

Edwards, A. W. F. (1994). The fundamental theorem of natural selection. *Biological Reviews*, 69:443–474.

Edwards, A. W. F. (2000). *Foundations of Mathematical Genetics*. Cambridge University Press, Cambridge, 2nd edition.

Edwards, A. W. F. (2007). Maximisation principles in evolutionary biology. In Matthen, M. and Stephens, C., editors, *Handbook of the Philosophy of Science: Philosophy of Biology*, pages 335–347. North-Holland, Amsterdam.

Edwards, A. W. F. (2014). R. A. Fisher's gene-centred view of evolution and the fundamental theorem of natural selection. *Biological Reviews*, 81:135–147.

Eshel, I. and 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. and Feldman, M. W. (2001). Optimality and evolutionary stability under short- and long-term selection. In Orzack, S. H. and Sober, E., editors,

Adaptationism and Optimality, pages 161–190. Cambridge University Press, Cambridge.

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

Ewens, W. J. (1968). A genetic model having complex linkage behaviour. *Theoretical and Applied Genetics*, 38:140–143.

Ewens, W. J. (1989). An interpretation and proof of the fundamental theorem of natural selection. *Theoretical population biology*, 36:167–180.

Ewens, W. J. (2004). *Mathematical Population Genetics*. Springer, New York, 2nd edition.

Ewens, W. J. (2011). What is the gene trying to do? *British Journal for the Philosophy of Science*, 62:155–176.

Ewens, W. J. and Lessard, S. (2015). On the interpretation and relevance of the fundamental theorem of natural selection. *Theoretical Population Biology*, 104:59–67.

Fisher, R. A. (1930). *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford, 1st edition.

Fisher, R. A. (1941). Average excess and average effect of a gene substitution. *Annals of Human Genetics*, 11:53–63.

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

Frank, S. A. and Slatkin, M. (1992). Fisher's fundamental theorem of natural selection. *Trends in Ecology and Evolution*, 7:92–95.

Futuyma, D. J. (2013). *Evolution*. Sinauer, Sunderland, MA, 3rd edition.

Geritz, S. A. H., G. M. and Metz, J. A. J. (1998). Evolutionarily singular strategies and the adaptive growth and branching of the evolutionary tree. *Evolutionary Ecology*, 12:35–57.

Godfrey-Smith, P. (2012). Darwinism and cultural change. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 367:2160–2170.

Godfrey-Smith, P. and Wilkins, J. F. (2008). Adaptationism. In Sarkar, S. and Plutynski, A., editors, *A Companion to the Philosophy of Biology*, London, pages 186–201. Blackwell.

Grafen, A. (1984). Natural selection, kin selection and group selection. In Davies, J. R. K. . N. B., editor, *Behavioural Ecology*, pages 62–84. Blackwell, Oxford, 2nd edition.

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*, 238:541–63.

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

Grafen, A. (2014). The formal Darwinism project in outline. *Biology and Philosophy*, 29:155–174.

Grafen, A. (2015). Biological fitness and the fundamental theorem of natural selection. *American Naturalist Nat*, 186(1):1–14.

Hamilton, W. D. (1964). The genetical evolution of social behaviour. *Journal of Theoretical Biology*, 7:1–52.

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

Hammerstein, P. (2012). Towards a darwinian theory of decision making: games and the biological roots of behavior. In Okasha, S. and Binmore, K., editors, *Evolution and Rationality: Decisions, Co-operation and Strategic Behavior*, pages 7–22. Cambridge University Press, Cambridge.

Hammerstein, P. and Selten, R. (1994). Game theory and evolutionary biology. In Aumann, R. J. and Hart, S., editors, *Handbook of Game Theory with Economic Applications*, volume 2, pages 929–993. Elsevier, Amsterdam.

Hedrick, P. W. (2011). *Genetics of Populations*. Jones and Bartlett, Sudbury, MA, 4th edition.

Kaplan, J., editor (2008). The Adaptive Landscape: Metaphors and Models [special issue]. *Biology and Philosophy* 23(5).

Karlin, S. (1975). General two locus selection models: some objectives, rules and interpretations. *Theoretical Population Biology*, 7:364–398.

Lessard, S. (1997). Fisher's fundamental theorem of natural selection revisited. *Theoretical Population Biology*, 52:119–136.

Lewens, T. (2007). Adaptation. In Hull, D. L. and Ruse, M., editors, *The Cambridge Companion to the Philosophy of Biology*, pages 1–21. Cambridge University Press, Cambridge.

Lewens, T. (2009). Seven types of adaptationism. *Biology and Philosophy*, 24:161–182.

Liberman, U. (1988). External stability and ESS: criteria for initial increase of new mutant allele. *Journal of Mathematical Biology*, 26:477–485.

Maynard Smith, J. (1982). *Evolution and the Theory of Games*. Cambridge University Press, Cambridge.

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

Moran, P. A. P. (1964). On the non-existence of adaptive topographies. *Annals of Human Genetics*, 27:383–393.

Mulholland, H. P. and Smith, C. A. B. (1959). An inequality arising in genetical theory. *American Mathematical Monthly*, 66:673–683.

Okasha, S. (2008). Fisher's 'fundamental theorem' of natural selection: a philosophical analysis. *British Journal for the Philosophy of Science*, 59:319–351.

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

Orzack, S. H. and Forber, P. (2012). Adaptationism. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Winter 2012 edition.

Parker, G. A. and Smith, J. M. (1990). Optimality theory in evolutionary biology. *Nature*, 348:27–33.

Pigliucci, M. and Kaplan, J. (2006). *Making Sense of Evolution: The Conceptual Foundations of Evolutionary Biology*. University of Chicago Press, Chicago, IL.

Plutynski, A. (2006). What was Fisher's fundamental theorem of natural selection and what was it for? *Studies in History and Philosophy of Biological and Biomedical Sciences*, 37:59–82.

Potochnik, A. (2009). Optimality modeling in a suboptimal world. *Biology and Philosophy*, 24:183–197.

Price, G. R. (1972). Fisher's fundamental theorem made clear. *Annals of Human Genetics*, 36:129–140.

Ridley, M. (2004). *Evolution*. Wiley-Blackwell, Hoboken, NJ, 3rd edition.

Roff, D. A. (1992). *The Evolution of Life Histories: Theory and Analysis*. Chapman and Hall, New York.

Sacks, J. M. (1967). A stable equilibrium with minimum average fitness. *Genetics*, 56:705–708.

Scheuer, P. A. G. and Mandel, S. P. H. (1959). An inequality in population genetics. *Heredity*, 31:519–524.

Seger, J. and Stubblefield, J. W. (1996). Optimization and adaptation. In Rose, M. and Lauder, G., editors, *Adaptation*, pages 93–123. Academic Press, San Diego.

Sober, E. (1987). What is adaptationism? In Dupré, J., editor, *The Latest on the Best: Essays on Evolution and Optimality on the best: essays on evolution and optimality*, pages 105–118. MIT Press, Cambridge, MA.

West, S. A. and Burton-Chellew, M. N. (2013). Human behavioral ecology. *Behavioral Ecology*, 24:1043–1045.

Wilkins, J. F. and Godfrey-Smith, P. (2009). Adaptationism and the adaptive landscape. *Biology and Philosophy*, 24(2):199–214.

Wright, S. (1932). The roles of mutation, inbreeding, crossbreeding and selection in evolution. In Jones, D. F., editor, *Proceedings of the Sixth International Congress of Genetics*, volume 1, pages 356–366. Brooklyn Botanic Garden, Menasha.

Wright, S. (1988). Surfaces of selective value revisited. *American Naturalist*, 131:115–123.