Natural selection and the maximization of fitness

Jonathan Birch[*,†]

Christ's College, University of Cambridge, St Andrew's Street, Cambridge, CB2 3BU, UK

[*]E-mail:j.birch2@lse.ac.uk; Tel.: +44 (0)20 7107 7334.

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

ABSTRACT

The notion that natural selection is a process of fitness maximization gets a bad press in population genetics, yet in other areas of biology the view that organisms behave as if attempting to maximize their fitness remains widespread. Here I critically appraise the prospects for reconciliation. I first distinguish four varieties of fitness maximization. I then examine two recent developments that may appear to vindicate at least one of these varieties. The first is the 'new' interpretation of Fisher's fundamental theorem of natural selection, on which the theorem is exactly true for any evolving population that satisfies some minimal assumptions. The second is the Formal Darwinism project, which forges links between gene frequency change and optimal strategy choice. In both cases, I argue that the results fail to establish a biologically significant maximization principle. I conclude that it may be a mistake to look for universal maximization principles justified by theory alone. A more promising approach may be to find maximization principles that apply conditionally and to show that the conditions were satisfied in the evolution of particular traits.

Key words: fitness maximization, natural selection, population genetics, evolutionary ecology, Fisher's fundamental theorem, Formal Darwinism.

CONTENTS

I. INTRODUCTION: CONFLICTING PARADIGMS

Evolutionary ecologists often take it for granted that natural selection can be regarded as a process of fitness maximization. Consider, for example, the following quotations from 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, p. 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, p. 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, Krebs & West, 2012, p. 81)

Given the pivotal role such assumptions are playing, one might be forgiven for assuming that they could be straightforwardly justified by population genetics. This, however, is far from the case. Most population geneticists doubt whether any justification of these assumptions is possible. This sceptical consensus is aptly summarized by A. W. F. Edwards (2007):

> [A] naive description of evolution 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, p. 353)

A theoretical rift thus separates the two fields. Evolutionary ecologists should be worried: on the face of it, if the population geneticists are correct, then evolutionary ecologists have built entire research programmes on foundational assumptions that are at best unjustified and at worst provably false. But population geneticists too should be uneasy. For once all talk of hill-climbing and fitness maximization is jettisoned, it becomes less clear why the dynamics of gene frequency change lead so regularly to the brilliantly well-adapted organisms we find in the natural world.

The prospects for reconciliation between these two contrasting approaches to understanding evolution have long been a source of debate. Scepticism about fitness maximization runs deep in population genetics, thanks to numerous results that seem to undermine the suggestion that natural selection maximizes fitness. However, two recent developments motivate an updated assessment. The first is the 'new' interpretation of Fisher's fundamental theorem of natural selection, on which the theorem is exactly true for any evolving population that satisfies some minimal assumptions (Price, 1972; Ewens, 1989, 2004, 2011; Edwards, 1994, 2014; Lessard, 1997; Plutynski, 2006; Okasha, 2008; Bijma, 2010). The second development is the Formal Darwinism project of Alan Grafen and colleagues, which forges links between formal representations of gene frequency change and optimal strategy choice (Grafen, 1999, 2000, 2002, 2003, 2006*a*, *b*, 2007, 2014; Gardner & Grafen, 2009; Gardner & Welch, 2011; Batty *et al*., 2014).

My aim herein is to examine the consequences of these developments for the notion of fitness maximization. My focus will be on theoretical and conceptual issues. In concentrating

on theory and concepts, I do not mean to imply that these are the only relevant considerations in understanding the relationship between population genetics and evolutionary ecology. To understand fully the reasons for the divergence of these two paradigms, we should also take account of empirical disagreements, as well as historical and sociological factors. Nevertheless, some of the reasons are theoretical and conceptual in nature, and, as Grafen and others have emphasized, a resolution of these issues—if possible—would be an important step towards bridging the divide between the two approaches.

Ultimately, I argue that neither Fisher's fundamental theorem nor Formal Darwinism establishes a maximization principle with biological meaning, and I conclude that it may be a mistake to look for universal maximization principles justified by theory alone. In the final section I suggest that a more promising approach may be to find maximization principles that apply conditionally, and to show that the conditions were satisfied in the evolutionary history of particular traits.

II. FOUR VARIETIES OF FITNESS MAXIMIZATION

Before we can assess the prospects for fitness maximization, we need a firmer grip on what it means. While the meaning of 'fitness' is open to debate, I intend to focus here on the concept of 'maximization' (for a review of recent work on the nature of fitness, see Rosenberg & Bouchard, 2010). For even if we could agree about the appropriate measure of fitness in any given context, there would still be room for disagreement regarding what it is for that measure to be 'maximized'. In fact, there are four broad varieties of 'maximization' that we must distinguish.

(1) Equilibrium *versus* change

Fitness maximization is often conceptualized using Sewall Wright's (1932) 'adaptive landscape' metaphor. Wright envisioned 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 seductive vision, 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 involves two senses of 'maximization' that it is helpful to distinguish. First, it involves a claim about equilibria: the stationary points of evolution by natural selection are 'adaptive peaks' at which mean fitness is maximized. Second, it involves a claim about change: a population out of population-genetic equilibrium moves reliably upward, in the direction of greater mean fitness.

These two claims are logically distinct. In principle, it might be that adaptive peaks are always stationary points and yet selection might be ineffectual at driving populations towards them. Conversely, selection might drive populations reliably upward, and yet the population might regularly stop at points that are not peaks. 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 is at a stable population-genetic equilibrium if and only if its mean fitness is maximized.
>
> **MAX-B (*Mean fitness, change*)**: if a population is not in population-genetic equilibrium, then natural selection will reliably drive it in the direction of greater mean fitness, even if other factors prevent the population from reaching a maximum.

Note that, in the eyes of some theorists, MAX-B does not amount to a 'maximization principle' at all, since it does not imply that fitness is maximized at equilibrium (*cf*. Ewens, 2014). MAX-B ascribes a directional 'hill-climbing' bias to the process of natural selection, but it does so without asserting that any particular outcome will result from this bias. Since my aim in this section is to disambiguate various senses of 'maximization' without commenting on the appropriateness of the usage, I will continue to refer to MAX-B as a 'maximization principle'. Nevertheless, it is important to note the contrast with MAX-A, which does assert that fitness is maximized at equilibrium.

(2) Challenges to MAX-A and MAX-B

While MAX-A and MAX-B 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. 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 & Mandel, 1959; Mulholland & 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, Feldman & Bergman, 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 any individual is assumed to be independent of population gene frequencies. Generally speaking, matters are even worse for mean fitness maximization when we introduce frequency-dependent 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 the fitness of other genotypes, detract from the mean fitness of the population. The moral of over 50 years of work in this area is that, when fitness depends on gene frequency, the mean fitness does not reliably increase and is rarely maximized at equilibrium (there are conditions under which mean fitness is maximized, but these conditions are restrictive; see Sigmund, 1987; Asmussen, Cartwright and Spencer 2004). 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 fitness minima act as 'evolutionary branching points' at which a population fragments, causing different subpopulations to pursue divergent evolutionary trajectories (Geritz, Mesze & Metz, 1998; Doebeli & Dieckmann, 2000; Doebeli, 2011).

Note that 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 they pose for MAX-B is not simply that the 'uphill push' of natural selection is counteracted by other evolutionary processes (drift, mutation, etc.). The problem is that even when there is no cause of gene frequency change other than natural selection, the mean fitness still decreases.

If there any way to salvage MAX-A or MAX-B? There are various moves we might make here, though none is uncontroversial. For example, we might try to defend a version of MAX-A by arguing that, although mean fitness maxima are not the only stable stationary points, they possess special stability properties that other stationary points do not possess. The tenability of this claim depends on the nature of these 'special stability properties'.

Two standard stability concepts are Lyapunov stability and asymptotic stability. Roughly speaking, an equilibrium is 'Lyapunov stable' if a population that starts close to the equilibrium stays close indefinitely; an equilibrium is 'asymptotically stable' if, in addition to being Lyapunov stable, it is such that the population will converge to it from any starting point (for global asymptotic stability) or from any local starting point (for local asymptotic stability). These stability concepts, however, provide little comfort for the defender of MAX-A. For once we move beyond the special case of frequency-independent selection at a single locus, equilibria which are not fitness maxima may also have these stability properties.

‘Evolutionary branching points’ are an example: they are asymptotically stable despite being fitness minima.

A different way of conceptualizing stability concentrates on the ability of an equilibrium to resist invasion by feasible mutants (*cf*. Maynard Smith & Price, 1973). Evolutionary branching points are not stable in this sense; they are in fact vulnerable to invasion by any nearby mutant (Doebeli, 2011, p. 20). Equilibria maintained by genetic constraints may also be unstable in this sense in the long run, since they may be vulnerable to invasion by mutants which circumvent the constraints. Consequently, we might be tempted to conjecture that, in any scenario in which fitness is not maximized at equilibrium, this equilibrium will be vulnerable to invasion in the long run. This line of thought is central to Hammerstein’s (1996) ‘streetcar theory’ of evolution, which we revisit below.

But what of MAX-B? Since this concerns the short-term dynamics of natural selection—i.e. the ‘direction’ in which selection ‘pushes’—considerations regarding the long-term malleability of genetic architecture do nothing to support it. However, there is one recent theoretical development that may appear to provide support for MAX-B. This is the ‘new’ interpretation of R. A. Fisher’s ‘fundamental theorem of natural selection’, on which the theorem (thought for many years to be, at best, approximately true as a claim about the action of natural selection) is exactly true, but concerns a ‘partial change’ in mean fitness rather than the total change. A tempting thought is that the theorem thus reveals an underlying tendency on the part of natural selection to increase mean fitness, a tendency often masked by other ‘partial changes’. I examine and ultimately criticize this thought in Section III.

(3) Individuals as maximizing agents

In both MAX-A and MAX-B, the variable that is maximized is the population mean of some appropriate fitness measure. Hence the processes of maximization described by MAX-A and

MAX-B are processes that (if they happen at all) happen to a population rather than to any individual organism.

MAX-A and MAX-B capture the senses of 'maximization' that are most commonly at stake in population genetics. But they are not the only senses of 'maximization' that matter in evolutionary biology. Ecologists commonly start with the assumption that an individual organism will act as if attempting to maximize its own individual fitness (or 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 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 behavioural ecology, including evolutionary game theory and optimality modelling (e.g. Maynard Smith, 1982; Parker & Maynard Smith, 1990; Davies *et al*., 2012). It also surfaces in informal arguments that appeal to the 'inclusive fitness interests' of an organism.

To invoke such an analogy is not to presuppose rational agency on the part of the organisms in question. 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 fitness, because natural selection tends to lead to equilibria at which organisms adopt strategies that at least approximately maximize fitness (or inclusive fitness) within the set of feasible options (Grafen, 1984, 1999).

This leads to a third sense of fitness-maximization, which we can characterize as follows:

> **MAX-C (*Individual fitness, equilibrium*)**: a population is at a stable population-genetic equilibrium if and only if all organisms adopt the phenotype that maximizes their individual fitness within the set of feasible options.

This notion of maximization 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 feasible options) on the part of individual organisms.

As with MAX-A and MAX-B, this leaves open the question of how the key notions should be formalized. Note, however, that the formal apparatus needed to capture the notion of an organism 'adopting a phenotype that maximizes its individual fitness within the set of feasible options' will be different from that needed to capture the notion of a population moving towards the peaks on an adaptive landscape. The notion of 'maximization' at work in MAX-C has little to do with the adaptive landscape metaphor. It is much closer to the notion of 'maximization' which appears in economics, in which human agents are typically modelled as rational utility maximizers. Given this, we should arguably look to economics for an appropriate set of mathematical tools with which to formalize MAX-C (see Section IV; see also Grafen, 1999; Okasha, Weymark & Bossert, 2014).

For defenders of fitness maximization, the main attraction of MAX-C over MAX-A is that it makes room for cases of strategic interaction in which a population evolves to a Nash equilibrium. In such cases, every agent plays the 'best response' to its opponent's strategy; it maximizes its pay-off conditional on what its opponent does. At a Nash equilibrium, mean fitness is rarely maximized; indeed, sometimes it is minimized (Doebeli & Hauert, 2005). Yet there is still a sense in which each organism at a Nash equilibrium is adopting the fitness-maximizing phenotype within the range of feasible options, conditional on what its opponents are doing. An organism may therefore exhibit a meaningful form of fitness maximization at equilibrium (i.e. 'conditional maximization' or 'best-response

maximization') even though the population as a whole remains far from any mean fitness maximum.

(4) MAX-C and the 'streetcar theory'

Although MAX-C has important advantages over MAX-A, one might still fear that it is vulnerable to simple counterexamples. Consider, for instance, a case of heterozygote advantage, such as the famous case of sickle-cell anaemia and malarial resistance (see Hedrick, 2011, or any other population genetics textbook). At the observed polymorphic equilibrium, it is clearly not the case that all organisms have a fitness-maximizing phenotype. How can we reconcile such cases with MAX-C?

One approach is to adopt a particularly demanding conception of stability, so that equilibria at which suboptimal phenotypes are present do not qualify as stable. Hammerstein (1996), as noted above, makes a move along these lines (see also Eshel & Feldman, 1984, 2001; Liberman, 1988; Hammerstein & Selten, 1994; Marrow, Johnstone and Hurst 1996; 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, p. 512). The 'final stop', he argues, will be a Nash equilibrium. Hence we arrive at 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.

Hammerstein's argument leaves open the question of whether genetic barriers to optimality will actually be circumvented in nature. This is ultimately an empirical matter, because it depends on the rate of mutation and the rate at which the selective environment

changes. As Eshel and Feldman (2001, p.186) note, their results 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 equilibria". 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".

The upshot is that 'streetcar theory' and related approaches do not provide unequivocal support for MAX-C: they support it only in conjunction with the empirical assumption that the trait under study evolved in the same environment for long enough and that its genetic architecture was malleable enough for barriers to optimality to be overcome. This may be the most we can expect from theory alone. After all, theory alone cannot establish the long-run invariance of selective environments or the long-run malleability of genetic arrangements.

Before jumping to this conclusion, however, we must consider one important recent development. This is the Formal Darwinism project of Grafen and colleagues, which forges links between formal representations of gene frequency change and optimal strategy choice, in the hope of vindicating something close to MAX-C (Grafen, 2007, 2014). I discuss this project in detail in Section IV.

One final variant of fitness maximization deserves a place in our classification. Note that MAX-C, like MAX-A, is a claim about what happens at equilibrium. In principle, however, the equilibrium/change distinction cross-cuts the mean fitness/individual fitness distinction. The result is that we can formulate 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

within the set of feasible options, even if other factors prevent the population from reaching this point.

As we will see, MAX-D provides an alternative way of conceptualizing the goal of Formal Darwinism. On the face of it, however, it too struggles to accommodate sickle-cell-type cases in which natural selection fails to drive a population in the direction of fitness-maximizing phenotypes, and may even drive it further away.

(5) Summary and preview

We have now distinguished four varieties of fitness maximization (summarized in Table 1), and we have seen why all four are controversial. It is now time to dive into the details. In the foregoing discussion, we noted two recent developments that may appear to vindicate a maximization principle. The first is the 'new' interpretation of Fisher's fundamental theorem of natural selection, which concerns the mean population fitness, and which has sometimes been taken to establish MAX-B. The second is Grafen's Formal Darwinism project, which focusses on individual fitness-maximizing behaviour, and which may be interpreted as aiming to establish MAX-C or MAX-D. In both cases, I argue that the relevant theoretical results, interesting as they may be from a mathematical point of view, do not establish a biologically significant maximization principle.

III. THE STATUS OF FISHER'S 'FUNDAMENTAL THEOREM OF NATURAL SELECTION'

(1) Old and new interpretations

Fisher's fundamental theorem of natural selection (FTNS) states (in Fisher's own terms) that 'the rate of increase in [mean] fitness of any organism [i.e. population] at any time is equal to its [additive] genetic variance in fitness at that time' (Fisher, 1930, p. 35). Because variance

cannot be negative, the theorem appears to imply that the mean fitness of a population evolving by natural selection cannot decrease over time. It therefore appears to support MAX-B.

The status of the fundamental theorem, and its relationship to MAX-B, depends a great deal on how these words and their accompanying mathematics are interpreted. Early commentators, notably including Sewall Wright, took Fisher's result to support a picture of evolution in which the action of natural selection is described by a potential, driving populations to adaptive peaks by the steepest route (Edwards, 1994). Fisher, for his part, vigorously opposed this reading of the theorem (Fisher, 1941; Edwards, 1994).

Wright's interpretation was understandable given Fisher's obscure original presentation, but it was problematic. Firstly, even if the theorem were true as a claim about the total change in mean fitness, it would not imply that mean fitness is maximized at equilibrium, let alone that the action of selection is described by a potential. For the claim that mean fitness never decreases over time (even if true) does not imply that it will arrive at a maximum rather than stabilizing at some suboptimal stationary point. Secondly, if interpreted as a claim about the total change in mean fitness, the theorem is subject to counterexamples in which the mean fitness of a population decreases (Section II; see also Ewens, 2004). Fisher was not infallible, of course, but he was surely well aware of the problems for naive fitness maximization, having himself constructed an example in which the mean fitness of an evolving population decreases (Fisher, 1941; Edwards, 1994). It may seem tempting, given this, to infer that Fisher only ever intended the FTNS to apply approximately or in special cases; but this fits uneasily with his claim to have derived a biological analogue of the second law of thermodynamics (Fisher, 1930, p. 36).

Price (1972) suggested a novel interpretation of the fundamental theorem that reconciles Fisher's apparently conflicting statements. On this 'new' interpretation, introduced to

mainstream population genetics by Ewens (1989), the theorem is concerned not with the overall change in mean fitness but only with the partial change in mean fitness attributable to natural selection. It should thus be read as stating (in Price's terms) that:

> In any species at any time, the rate of change of fitness *ascribable to natural selection* is equal to the [additive] genetic variance in fitness at that time. (Price, 1972, p. 132; my italics, his square brackets)

If this statement were true, MAX-B would stand vindicated. The difficulty is that, as Price himself saw and Ewens (1989, 2011) has emphasized, it is problematic to interpret the 'partial change' described by the FTNS as the 'partial change ascribable to natural selection'. Hence, although the FTNS is a true statement about a 'partial change', this 'partial change' does not carry the biological meaning that Fisher hoped to attach to it. To see why, let us turn briefly to the formal details of the FTNS.

(2) A brief explanation of the FTNS

The aim of this subsection is to provide a concise explanation, rather than a formal derivation, of Fisher's fundamental theorem. Fisher's FTNS was a continuous-time, many-allele, many-locus result, but for ease of exposition I explain the theorem in the context of a discrete-time, two-allele, one-locus model. Even in this simple case, some technicality is unavoidable, but I have kept this to a minimum. For further mathematical detail, including derivations of the theorem in continuous time for many alleles and many loci, see Ewens (1989, 2004, 2011).

The key to understanding the FTNS is to understand Fisher's notion of an 'average effect'. Informally, we can talk of the average effect of an allele (say, A_1) as the average change in fitness we would observe if we were to take an individual at random and increase its dosage of A_1 by one copy. More formally, we can follow Fisher (1930, 1941) in defining

the average effect of an allele by means of a least-squares minimization procedure. Let W_{ij} represent the fitness (and P_{ij} the frequency) of the genotype A_iA_j in the initial generation (in our simple two-allele case, *i* and *j* can only take the values 1 or 2). We can write W_{ij} as a sum of the average effects (labelled β_i and β_j, respectively) of the alleles in A_iA_j, plus a residual component ε:

$$W_{ij} = \beta_i + \beta_j + \varepsilon \ . \qquad (1)$$

Following Fisher, we say that (by definition) the average effects of the alleles A_1 and A_2 (i.e. β_1 and β_2) are constants which minimize the sum of squares of residuals (across all genotypes) in the above regression model. That is, they are those constants such that the following quantity is minimized:

$$\sum_i \sum_j P_{ij}(W_{ij} - \beta_i - \beta_j)^2 \ . \qquad (2)$$

If there are more than two alleles in competition, the basic definition of the 'average effect' in terms of least-squares minimization is unchanged, but the computation of these coefficients becomes more demanding as the number of alleles increases.

Since our least-squares procedure for evaluating β_1 and β_2 guarantees that $\bar{\varepsilon} = 0$, we can write the mean fitness in the population, $\bar{W}$, as a function of genotype frequencies and average effects only (i.e. averaging makes the 'ε' disappear):

$$\bar{W} = \sum_i \sum_j P_{ij}(\beta_i + \beta_j) \ . \qquad (3)$$

Equation (1) simplifies to the following, where P_i is the frequency of the *i*th allele:

$$\bar{W} = 2\sum_i P_i\beta_i \ . \qquad (4)$$

Now consider the second generation. Let P'_i represent the new frequency of the the *i*th allele, let W'_{ij} represent the new fitness of A_iA_j, and let β'_i and β'_j represent the new average effects. The line of reasoning that led to equation (4) also tells us that the mean fitness in the second generation can be expressed as follows:

$$\bar{W}' = 2\sum_i P'_i\beta'_i \ . \qquad (5)$$

Hence the total change in mean fitness between the two generations can be expressed as follows:

$$\Delta\bar{W} = 2\sum_i (P'_i\beta'_i - P_i\beta_i) \ . \qquad (6)$$

Having derived this expression, we now need to put it to one side. For this total change is not the change with which Fisher's fundamental theorem is concerned. Instead, the theorem concerns only a partial change in mean fitness.

To understand this notion of 'partial change', suppose that the average effects (i.e. the β_i) are unchanged between the two generations. Given this supposition, $P'_i\beta'_i - P_i\beta_i$ would simplify to $(P'_i - P_i)\beta_i$. In reality, however, average effects often do change over time; so a 'fundamental theorem' that required that they do not change would have limited relevance to real evolving populations. But here Fisher, at least on Price's interpretation, makes a subtle move. The key is to note that, even if the actual average effects of alleles do change, we can still stipulatively define a quantity—call it $\Delta_F\bar{W}$ —that captures what the change in mean fitness would have been on the supposition that average effects had retained their initial values:

$$\Delta_F\bar{W} = 2\sum_i (P'_i - P_i)\beta_i \qquad (7)$$

Fisher called this quantity (in a continuous-time model) the 'genic rate of increase' in mean fitness, reflecting his (contentious) view that it captured the component of the total change that is ascribable to changes in gene frequency rather than to environmental change. We can, more neutrally, call it the *Fisherian partial change* in mean fitness. This is the quantity that the FTNS concerns. Note that, if the average effects of alleles do change between generations, equation (7) is still correct and the Fisherian partial change is still well defined; it is just that, in such cases, it cannot be equated with the total change in mean fitness.

We are now in a position to state the FTNS itself (or at least its analogue in the discrete-time case, since Fisher's theorem was a continuous-time result). The theorem states that the Fisherian partial change in mean fitness ($\Delta_F \bar{W}$) is equal to the additive genetic variance in fitness in the initial generation (σ_A^2) divided by the mean fitness in the initial generation ($\bar{W}$). That is:

$$\Delta_F \bar{W} = \frac{\sigma_A^2}{\bar{W}}. \qquad (8)$$

In the one-locus case, the theorem is not difficult to prove, given the definition of additive genetic variance and its close affinity (*via* the method of least-squares regression) with the notion of an average effect. I refer the reader to Ewens (1989, 2004, 2011) for details of the derivation and its extension to the many-allele, many-locus case.

(3) Does the FTNS vindicate MAX-B?

Stated carefully, the FTNS is mathematically sound. But questions remain as to its biological significance. The switch from the total change in mean fitness to the Fisherian partial change is necessary if the theorem is to hold when the average effects of alleles change over time. A consequence of this move, however, is that there is no serious prospect of the FTNS vindicating MAX-A, since the latter concerns the overall stationary points of the evolutionary

process, and these stationary points depend on the total dynamics, not just on a partial change. But could the FTNS still vindicate MAX-B?

This turns on whether or not the Fisherian partial change ($\Delta_F \bar{W}$) sustains a biological interpretation as the component of the overall change that is causally attributable to natural selection. If it does, then its sign will tell us something about the direction in which natural selection is 'pushing' or 'driving' the population. But does it? There is disagreement on this point in the theoretical literature. Frank (1997, 1998) and Grafen (2002, 2003) appear to regard the identification of $\Delta_F \bar{W}$ with the partial change attributable to natural selection as unproblematic. By contrast, Ewens (1989, 2004, 2011) explicitly rejects this identification. Price (1972) initially subscribes to it, but later in the same paper expresses reservations.

I side with the sceptics here. The identification of the Fisherian partial change with the change causally attributable to natural selection rests on the assumption that the part of the total change which it neglects (i.e. the part which depends on changes in the average effects of alleles) is not affected by natural selection. In many cases, however, the average effects of alleles are functions of gene frequency. As a consequence, there are many cases in which natural selection, by virtue of altering gene frequencies, also alters the average effects of alleles, influencing the course of evolution in a way that is not captured in the Fisherian partial change. In short, the FTNS fails to establish MAX-B because the Fisherian partial change (which is always positive) is not equivalent to the partial change causally attributable to natural selection (which is not).

Are there any ways to rescue the biological significance of the FTNS in light of this problem? One possibility is suggested by Okasha (2008) who, drawing inspiration from Price (1972), suggests that, although the Fisherian partial change cannot simply be identified wth the change due to natural selection, it may be identified with the change due to natural selection acting in a constant environment. The thought is that, by introducing the

qualification ‘acting in a constant environment’, we implicitly exclude any effect natural selection may have on the average effects of alleles, because selection only alters the average effects of alleles *via* its effects on the environment.

However, the notion of ‘environment’ must be understood very broadly for this move to succeed. In some cases, the average effects of alleles change without any accompanying change in the ecological environment, or even any change in the fitness of genotypes. Sometimes, the average effects of alleles change simply because genes interact non-additively (i.e. with dominance or epistasis), so that their average effects on fitness depend on the relative frequencies of the various genotypic contexts in which they might find themselves. For instance, in a case of heterozygote advantage, the average effects of the relevant alleles depend on the frequency of heterozygotes. For the Price/Okasha response to work, we must regard any change in genotype frequencies, if it leads to a change in the frequency of epistatic or dominance effects, as a form of environmental change.

This is a contentious move. As Price himself pointed out, it is far from intuitive to regard changes in genotype frequencies as changes in the environment. Admittedly, as Okasha points out in response, it seems more intuitive if we adopt (as Fisher did) a ‘gene’s-eye view’ of evolution (see also Edwards 2014). From the perspective of single allele, the rest of the organism’s genome, including the other allele at the same locus, can be viewed as part of the ‘environment’ with which it interacts. With this in mind, it makes more sense to conceptualize a change in the frequency of dominance or epistatic effects in a population as a change in the ‘genic environment’ experienced by the average of copy of an allele, and this appears to have been the notion of ‘environmental change’ with which Fisher was working (Price 1972; Sterelny and Kitcher 1988; Grafen 2003; Okasha 2008; Edwards 2014).

Yet even if we accept this broad conception of environmental change, so that any change in the average effects of alleles implies a change in the environment, we do not thereby arrive

at a biologically significant maximization principle. If we identify the Fisherian partial change with the change attributable to natural selection acting in a constant genic environment, then we may interpret the FTNS as supporting the following modified version of MAX-B, which we can call MAX-B*:

> **MAX-B***: If a population is not in population-genetic equilibrium, then natural selection *acting in a constant genic environment* will reliably drive the population in the direction of greater mean fitness, even if other factors prevent the population from reaching a maximum.

The trouble is that, in cases in which dominance, epistasis, or frequency-dependent fitness make a difference to the average effects of alleles, MAX-B* is a vacuous claim. Natural selection cannot 'act in a constant genic environment' in these circumstances, because in acting it alters genotype frequencies, and in altering genotype frequencies it alters the genic environment. MAX-B* is therefore nothing but a restatement of the claim that, when gene interactions and frequency-dependence are absent, MAX-B holds. The proviso 'acting in a constant genic environment' repackages what we already knew about MAX-B's limitations in a way that makes these limitations appear less restrictive, but this is not a new maximization principle of any biological significance.

None of this implies that the FTNS is not a valuable mathematical theorem. My claim is that it does not yield a biologically significant maximization principle, but this is compatible with it providing important mathematical insights in other ways. For example, Ewens (2011) suggests that the Fisherian partial change can be interpreted as the component of the total change that is mathematically independent of the mating scheme. This is because average effects depend on how genes are allocated to genotypes, which in turn depends on the mating scheme; so we need to know the mating scheme in order to calculate the total change in mean fitness. By contrast, nothing in $\Delta_F \overline{W}$ depends on the mating scheme, so in principle we can

calculate $\Delta_F \bar{W}$ in ignorance of who mates with whom. On Ewens's reading, the partition Fisher achieves by separating out $\Delta_F \bar{W}$ from the rest of the change in mean fitness is a partition between what we can and cannot know in ignorance of the mating scheme: $\Delta_F \bar{W}$ captures the component of the change we can know, the rest we cannot. It does not reflect a causal partition between the 'change due to natural selection' and the change due to other processes. On this reading, the FTNS does embody a genuine insight. For our purposes, however, what matters is that, whatever other interest it may have, it does not vindicate MAX-B.

IV. GRAFEN'S 'FORMAL DARWINISM' PROJECT

I turn now to Formal Darwinism. This section provides a brief introduction to Grafen's project. The following section subjects it to critical scrutiny. I should note that, although I will discuss Grafen's results and proofs (mostly) verbally, Grafen himself presents them within a complex formalism. This formality is understandable: the point of the Formal Darwinism project is to vindicate the 'individual as maximizing agent' analogy without falling back (as previous authors have done) on verbal arguments. The downside to this rigour is that it renders the arguments inaccessible to the uninitiated. A (relatively) informal précis will do for our purposes.

(1) Ingredients

(*a*) *The Price equation*

The goal of Formal Darwinism is to forge links between formal representations of gene frequency change and optimal strategy choice. The 'population genetics' half is provided by a version of the Price equation (Price, 1970) formulated in terms of '*p*-scores' (Grafen, 1985):

$$\Delta\bar{p} = \frac{1}{\bar{w}}\left[\mathrm{Cov}\left(w_i, p_i\right) + \mathrm{E}\left(w_i \Delta p_i\right)\right] , \qquad (9)$$

where w_i denotes the individual fitness of the *i*th individual, p_i denotes its *p*-score (explained below), Δp_i denotes the change in *p*-score between the *i*th individual and its offspring, $\bar{p}$ denotes the mean *p*-score in the population, and $\bar{w}$ denotes the mean individual fitness. $\Delta\bar{p}$ is the change in the average *p*-score between generations, which is expressed as the sum of $\mathrm{Cov}(w_i, p_i)$, the covariance in fitness between individual fitness and *p*-score in the initial population, and $\mathrm{E}(w_i \Delta p_i)$, the average across individuals in the initial population of the quantity $w_i \Delta p_i$.

Here fitness is conceptualized as a property of an individual organism, *viz*. its actual number of offspring. This marks a difference from most population genetics models, in which fitness is a parameter (written as *W* in the preceding section) that attaches to a genotype rather than an individual. I note in passing that Grafen employs a more complicated formulation of the Price equation which, unlike Price's original formulation, incorporates stochasticity (Grafen, 2000). There is room for debate about the best way to accommodate stochasticity within Price's framework; see Rice (2008) for an alternative approach. I neglect this complication here for ease of exposition.

A *p*-score, as Grafen defines it, can be any quantity 'that an offspring inherits by averaging together the gametic contributions of its parents' (Grafen, 1985, p. 33). One example of a *p*-score is an individual's individual gene frequency with respect to a particular allele, i.e. its total number of copies of the allele in question, divided by its ploidy. Any weighted sum of individual gene frequencies will also constitute a *p*-score. Importantly, there is no formal requirement that a *p*-score has a biological interpretation as a weighted sum of individual gene frequencies—all that matters formally is that we can average the parental

values to get the offspring value. Hence we could, in principle, define a *p*-score simply by assigning numbers to individuals arbitrarily and equating an individual's *p*-score with its arbitrary number, subject to the sole constraint that the number assigned to each individual must equal the average of the numbers assigned to its parents. Grafen explicitly allows for such "hypothetical *p*-scores, in which we assign a number to each individual, without asking whether there is an actual set of allelic weights that does produce that set of numbers" (Grafen, 2006*b*, p. 553).

Grafen assumes unbiased transmission of *p*-scores from parents to offspring, which amounts to an assumption of "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, p. 82). This entitles him to work with the following simplified version of the Price equation:

$$\Delta\overline{p} = \frac{1}{\overline{w}}\left[\mathrm{Cov}\left(w_i, p_i\right)\right] \qquad (10)$$

(*b*) *Optimization programmes*

As far as Formal Darwinism is concerned, population genetics is only half the story. In order to forge formal links between concepts of natural selection and concepts of optimization, we also need to formalize the latter. To this end, Grafen employs the formal apparatus of optimization programmes, originally developed by economists to model rational choice (Mas-Colell, Whinston & Green, 1995).

The two central concepts that define an optimization programme are the strategy set and the maximand. The strategy set, *X*, is a set each member of which corresponds to a possible phenotype. If we are interested in behaviour, these phenotypes will be behavioural strategies, but the phenotypes can also be non-behavioural. In principle, the set of possible phenotypes can be as inclusive as we like. Grafen's formalism is intentionally permissive in this respect, because his aim is to prove links that hold regardless of the choice of *X*. In practice, the most

relevant set will usually consist of the set of phenotypes actually present in the population plus the set of biologically feasible alternatives.

The maximand, m, is formally defined as a function that maps X onto the set of real numbers. The only constraints on the nature of the maximand are that (*i*) we can assign a real-numbered value for that property to every individual in the population and (*ii*) an individual's value functionally depends on its phenotype. Many properties could satisfy these criteria, and we could define an optimization programme for any of them. Ideally, however, we want to find a maximand that is (*a*) biologically significant, and (*b*) actually solved (or at least approximately solved) by the behaviour of real organisms at population-genetic equilibrium. This constrains the choice of maximand a great deal.

In the abstract, an optimization programme takes the following form:

$$x \max\left[m(x) \right],\ x \in X \ . \qquad (11)$$

A solution to the optimization programme, x^*, is a phenotype that satisfies the two constraints embodied in equation (11): it is a member of X and, among the members of X, it maximizes (i.e. maps on to the highest value of) m. Grafen (2002, 2006*b*) adds several layers of complexity to his optimization programmes that we do not need to consider here.

(*c*) *'Scope for selection' and 'potential for positive selection'*

The Price equation captures the basic idea of evolution by natural selection, while the notion of an optimization programme captures what it takes for an organism to behave 'as if attempting to maximize' some quantity. Grafen's strategy is to prove links between claims about natural selection (formulated in terms of the Price equation) and claims about fitness-maximizing behaviour (formulated in terms of an optimization programme). The links themselves are formulated in terms of two crucial bridging notions: scope for selection, and potential for positive selection in relation to X.

Scope for selection can be informally defined (within a discrete-generations model) as follows:

> **Scope for selection**: there is a possible *p*-score such that the (expected) change in the population mean for that *p*-score between the current generation and the next generation is non-zero.

The word 'possible' is significant here. Recall that the only formal requirement on a *p*-score is that an offspring's value is an average of its parents' gametic contributions. There is no formal requirement that a *p*-score can be interpreted in terms of actual alleles in the population: *p*-scores may be merely 'hypothetical' (Grafen, 2006*b*, p. 553). These hypothetical *p*-scores still count for the purposes of determining whether there is scope for selection in Grafen's sense. The implication is that there can be scope for selection in a population even if no change in allele frequencies takes place. Indeed, if a population contains fitness differences then there is guaranteed to be some possible *p*-score that co-varies with fitness, irrespective of whether or not fitness co-varies with any actual alleles (Okasha & Paternotte, 2012, 2014).

The second bridging notion is that of 'potential for positive selection in relation to *X*'. This notion explicitly blends the language of population genetics ('positive selection') with the language of optimization programmes ('in relation to *X*'). Informally, it is defined as follows:

> **Potential for positive selection in relation to *X***: there is a phenotype in *X* that would have been favoured by selection if it had been present in the population.

As we noted above, the set of possible phenotypes, *X*, can be as inclusive as we like: there are no 'biological feasibility' constraints built into the formalism. However, a sensible choice of *X* will limit membership to the actual phenotypes in the population of interest plus the set of

biologically feasible alternatives, and I will tacitly assume this choice of X in the discussion below.

(2) Links

With the necessary formalism in place, it is relatively straightforward to prove links (conditional on a few important assumptions, discussed below) between the Price equation and an optimization programme in which the maximand (m) is the expected relative fitness of the focal individual (i.e. the expected value of its individual fitness as a fraction of the population mean). Grafen (2002) proves four such links:

> **LINK 1**: if each individual acts optimally in relation to X, then there is no scope for selection and no potential for positive selection in relation to X.
>
> **LINK 2**: if each individual acts sub-optimally in relation to X, but equally so, then there is no scope for selection but there is potential for positive selection in relation to X.
>
> **LINK 3**: if individuals vary in the value of the maximand they attain, then there is scope for selection, and the change in every gene frequency and in the additive genetic value of every character equals its covariance across individuals with the value of the maximand.
>
> **LINK 4**: if there is no scope for selection and no potential for positive selection in relation to X, then each individual in the population acts optimally in relation to X.

Links 1, 2 and 3 jointly imply Link 4, so the inclusion of Link 4 in the list is not compulsory. But it is worth including, because it helps bring out the implications of the links.

A number of assumptions are required if the links are to hold. We have already mentioned that, in formulating the 'population genetics' side of the formalism, Grafen assumes "no mutation, no gametic selection, fair meiosis and that all the loci contributing to the p-score

have the same mode of inheritance". The assumption of 'no mutation' stands out as somewhat concerning. For as Hammerstein (1996) emphasizes, the long-run outcomes of evolution depend a great deal on the extent to which mutation is able to modify the genetic architecture of phenotypic traits. Since these long-run outcomes are a central concern of Formal Darwinism, it must be regarded as a limitation of Grafen's framework that it currently assumes the absence of mutation.

A further important assumption is that all individuals face the same optimization programme: that is, the strategy set X and the mapping function are the same for every organism, so that if we were to swap the phenotypes of any pair of organisms we would also swap their attained values of the maximand. Grafen introduces this assumption under the name 'pairwise exchangeability'. If the maximand in question is individual fitness, it amounts to the assumption that the expected fitness of every organism depends on its phenotype in the same way; this in turn requires that there are no systematic fitness-affecting differences in the local environments organisms occupy. In essence, we are assuming that "all individuals face the same environmental challenges, and so are having to solve the same problems" (Grafen, 2002, p. 79).

Grafen's (2002) arguments also assume the absence of frequency-dependent selection and of social interactions. The rationale is broadly the same in both cases: if individual fitness is to be a suitable maximand, an organism's fitness must depend on its own phenotype but not on the phenotype of other organisms. Frequency dependence and social interaction both introduce ways in which an organism's fitness in a given environment can depend on the phenotypes of others, and Grafen thus assumes their absence. Of course, there can be no doubt that both phenomena are often present in real populations, so the necessity of assuming their absence severely limits the scope of Grafen's (2002) arguments.

In more recent work, Grafen and colleagues have sought to relax these assumptions. Grafen (2006*b*) proves that the four links do obtain in cases of social interaction, but only when an organism's inclusive fitness (Hamilton, 1964), rather than its individual reproductive output, is taken as the maximand. In the special case in which social interactions are absent, an individual's inclusive fitness is equal to its direct reproductive output, so the non-social version of Grafen's links are recovered as a special case of the social version. This suggests that, if any quantity can serve as a universal maximand for animal behaviour, it is inclusive fitness rather than individual fitness (a point stressed by West & Gardner, 2013).

However, Grafen (2006*b*) does not relax the assumption of frequency independence, and this remains largely unfinished business. Gardner & Welch (2011), in a paper primarily aimed at extending the 'individual as maximizing agent' analogy to the level of individual genes (an issue I do not discuss here), suggest one way in which the assumption of frequency independence can be relaxed. Their suggestion, in informal terms, is that we regard the actual phenotypic composition of a population as part of the environment that all agents share. We can thereby recover a function-like mapping from phenotypic states to attained maximand values in a given environment, because trait frequencies are now considered to be part of this environment (note the parallel here with Fisher's inclusive conception of environmental change, discussed in Section III).

It may appear to be a drawback of this approach that any change to the phenotypic composition of the population will potentially change the mapping: a strategy that was optimal, relative to the population's initial phenotypic composition, can cease to be optimal a generation later. But this may simply be an unavoidable feature of frequency-dependent selection. When fitness is frequency dependent, talk of the optimality (or otherwise) of a strategy only makes sense relative to a specification of what the other individuals in the population are doing (*cf*. Parker & Maynard Smith, 1990).

V. WHAT DO GRAFEN'S LINKS ACTUALLY SHOW?

(1) Which variety of maximization is at stake?

I now turn to the significance of Grafen's results for the idea of fitness maximization. First of all, let us consider how they relate to our four varieties. In Grafen's view, the links jointly provide "a secure logical foundation for the commonplace biological principle that natural selection leads organisms to act as if maximizing their 'fitness'" (Grafen, 2002, p. 75). Remarks such as this make it clear that Grafen's target is an 'individual as maximizing agent' analogy, not a population-level maximization thesis concerning mean fitness. The target can therefore be located in the second row in Table 1.

But which column? Do the links concern what happens at equilibrium or what happens out of it? The answer is that they concern both. As Grafen (2006*b*, p. 543) puts it:

> A broad interpretation [of the links] is that natural selection always changes gene frequencies in the direction of increasing inclusive fitness; and that a population genetic equilibrium in which no feasible mutations can spread implies that the individuals in the population are each acting so as to maximize their inclusive fitness.

The first claim concerns out-of-equilibrium change, while the second concerns what happens at equilibrium. As Grafen has noted elsewhere, it is significant that "the links between the optimization program and the population genetics model do not hold only at solutions to the optimization [programme]. There are also out-of-equilibrium links" (Grafen, 2002, p. 89). This concern with out-of-equilibrium situations—with "selection in progress" (Grafen, 2007, p. 1253)—differentiates Formal Darwinism from Hammerstein's 'streetcar theory' and related work, which solely concerns the stable stopping points of the process. Nevertheless, a concern with what happens at equilibrium clearly remains central. In particular, Links 1 and

4 are concerned only with what happens when scope for selection and potential for positive selection are absent. Only Links 2 and 3 relate to 'selection in progress'.

On the face of it, then, Grafen's links may be relevant to both MAX-C and MAX-D. I will therefore consider how the links relate to both. In saying this, I am not attributing to Grafen the view that the links do establish either principle. Grafen himself is cautious about attaching any 'official' interpretation to the links, and (to his credit) he is upfront about the fact that the 'broad interpretation' given above needs qualifying in various respects. That said, Grafen does not hide his sympathies for fitness maximization, and in a recent overview article glosses the links 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" (Grafen, 2014, p. 166). In my view, this overstates the implications of the links. I contend that, if 'fitness maximization' here refers to our MAX-C or MAX-D, then the links do not in fact support such an expectation.

(2) Do the links vindicate MAX-C?

Do Grafen's links succeed in vindicating MAX-C? First, note that since the links rely on idealizing assumptions—including the absence of mutation, gametic selection and meiotic drive—they could not directly support MAX-C as a claim about evolution in the real world. The most they could show is that in an hypothetical world without mutation, gametic selection and meiotic drive, MAX-C would be true (A. Grafen, personal communication). Call this reformulated principle MAX-C*:

> **MAX-C***: assuming the absence of mutation, gametic selection and meiotic drive, a population is at a stable population-genetic equilibrium if and only if all organisms adopt the phenotype that maximizes fitness within the range of feasible options.

Since its assumptions are usually false, MAX-C* connects less directly to real-world evolutionary processes than MAX-C. Nevertheless, it would (if true) still represent an interesting and in some ways surprising result. After all, there is reason to think that mutation makes it easier for evolution to arrive at optimal phenotypes, since it provides a way for evolution to overcome genetic barriers to optimality (*cf*. Hammerstein's 'streetcar theory'). Intuitively, we would expect these genetic barriers to be insurmountable in a world without mutation. So, if Grafen's links do support MAX-C*, they support a principle that, taken at face value, looks harder to defend than our original MAX-C.

But do Grafen's links actually establish MAX-C*? Here I am sceptical. My scepticism derives from the fact that the links do not explicitly state conditions for a stable population-genetic equilibrium. Instead, they use the notions of 'no scope for selection' and 'no potential for positive selection' as proxies. Grafen describes these concepts as "equilibrium concepts" (Grafen, 2006*b*, p. 553) and suggests that they jointly imply "a kind of population genetic equilibrium" (Grafen, 2006*b*, p. 552), but I find this interpretation questionable. As I see it, the absence of scope for selection and potential for positive selection is neither necessary nor sufficient for a population to be at a stable population-genetic equilibrium.

To see why the absence of scope for selection and potential for positive selection are unnecessary for a stable population-genetic equilibrium, consider again the polymorphic equilibrium of the standard sickle-cell anaemia model. Here we have a stable population-genetic equilibrium, yet we have both scope for selection and potential for positive selection (a point noted by Grafen, 2002, 2007, 2014, and Okasha & Paternotte, 2014). We have scope for selection because there are differences in individual fitness among members of the population. Hence although there may be no actual *p*-score that correlates with fitness, there is a possible *p*-score that does. Intuitively, this possible *p*-score corresponds to a gene that encodes malarial resistance in the heterozygote without encoding sickle-cell anaemia in the

homozygote, or one which simply encodes malarial resistance in all its bearers. We have potential for positive selection because there is a phenotype in *X* that co-varies positively with fitness (namely malarial resistance), and positive covariance between a (potential or actual) phenotype in *X* and fitness is all that is required for potential for positive selection (Grafen, 2002, p. 87).

It might be objected here that although the absence of scope and potential for selection are unnecessary for an equilibrium that is stable in the short term, they are necessary if we want an equilibrium that is stable in the long run. The thought (echoing Hammerstein) would be that an equilibrium at which there is still scope and potential for selection (such as the polymorphic sickle-cell equilibrium) will be vulnerable in the long run to invasion by mutants that modify the genetic architecture underlying the phenotype of interest. But while this idea is central to Hammerstein's alternative approach to these issues, it is not a promising move in the present context. This is because, as we noted above, Grafen's formal framework assumes the absence of mutation. As a consequence, it also assumes the absence of mutants that can circumvent genetic constraints. In this mutation-free setting, there is no reason to think that a polymorphic equilibrium maintained by dominance (or epistasis) should be any less stable than a monomorphic equilibrium at which all phenotypes are optimal. In a world without mutation, the 'streetcar' move gains no traction.

To see why the absence of scope for selection and potential for positive selection are insufficient for a stable population-genetic equilibrium, consider the following hypothetical example, based on cases discussed by Grafen (2002, 2014): a population contains three genotypes, *AA*, *BB* and *AB*. These genotypes produce phenotypes P_1, P_2 and P_3, respectively. The optimal phenotype is P_3, which is produced by the *AB* heterozygote only. Initially, the population is 100% *AB*, and every individual has the optimal phenotype. In the next generation, assuming random mating and fair meiosis, we expect to find 25% *AA* (P_1),

25% *BB* (P_2), and 50% *AB* (P_3), and it will no longer be the case that every individual has the optimal phenotype.

At the initial 'all heterozygotes' point, the population is highly unstable, and yet there is no scope for selection and no potential for positive selection in relation to *X*, where *X* is the set (P_1, P_2, P_3). There is no scope for selection because there are no fitness differences in the initial generation, and so no possible *p*-score co-varies with fitness in that generation. There is no potential for positive selection because no phenotype co-varies positively with fitness (there is zero variance in both), and there is no possible phenotype in *X* that would co-vary positively with fitness if introduced (because all phenotypes are already optimal). Yet the population is not at a stable equilibrium, because gene frequencies are expected to change as soon as homozygotes arise. As Grafen (2014, p. 158) points out, examples such as this show vividly that "even once a population consists of all optimal phenotypes [...] it can, for reasons that are entirely clear if the genetic system is known, move far away from that state".

To be clear, Grafen cannot be accused of trying to sweep these awkward cases under the rug. On the contrary, he has drawn attention to them on several occasions (Grafen, 2002, 2007, 2014) and admits in recent work that they may "call into question the meaning and value of the links" (Grafen, 2014, p. 164). The question is whether it is possible to maintain, in light of these cases, that the links still support a form of fitness maximization. There can be no doubt that cases like these complicate the relationship between Formal Darwinism and fitness maximization, and that they do so in a way that undermines the suggestion that Grafen's links support MAX-C*. The problem is not that cases of heterozygote advantage are counterexamples to Grafen's formal links (they are not!), but that they expose a logical gap between a stable population-genetic equilibrium and a point in the dynamics at which there is neither scope nor potential for selection. Since the latter is neither necessary nor sufficient for the former, there is no valid inference from Grafen's four links to MAX-C*.

If Grafen's links do not imply MAX-C or MAX-C*, what do they imply? I suggest the closest proposition to MAX-C that the links do support is the following:

> **MAX-C****: assuming the absence of mutation, gametic selection and meiotic drive, a population is at a point at which there is neither scope nor potential for selection if and only if all organisms adopt the fitness-optimal phenotype within the range of feasible options.

MAX-C** may be the closest claim to MAX-C that it is possible to derive from formal arguments alone, without invoking more substantial assumptions (e.g. assumptions about the long-run malleability of genetic systems). However, one might question whether MAX-C** is really enough, if the aim is to derive a maximization principle capable of vindicating the assumptions quoted at the start of this review. For once we recognize the gap between a stable population-genetic equilibrium and a point at which there is neither scope nor potential for selection, the biological significance of MAX-C** becomes less clear, as does its relevance to evolutionary ecology.

In effect, Grafen's links show that there is a way of characterizing a special point in the dynamics such that, in a world without mutation, gametic selection and meiotic drive, a population's being at that point entails (and is entailed by) its being phenotypically optimal. But it is important to see that this special point in the dynamics will not always correspond to a stable population-genetic equilibrium in the usual sense of the term. What the evolutionary ecologist would like, ideally, is a robust link between phenotypic optimality and population-genetic equilibrium. Since MAX-C** does not provide such a link, its biological significance remains open to question.

(3) Do they vindicate MAX-D?

The message of the preceding section is that Grafen's links do not establish MAX-C. They do support a substantially modified version of that principle, MAX-C**, but the biological significance of this principle is questionable. Let us now turn to MAX-D, the parallel claim about change. Do the links vindicate the idea that natural selection reliably drives populations in the direction of a point at which all agents are behaving optimally? Or that, in a world without mutation, gametic selection or meiotic drive, it would do so?

In this respect, Links 2 and 3 are pivotal, for these are the links that explicitly concern out-of-equilibrium situations. Link 2 tells us that if all organisms are behaving suboptimally, but equally so, there is no scope for selection (no possible *p*-score co-varies with fitness) but there is potential for positive selection: there is a feasible phenotype that, if present, would be favoured. Link 3 tells us that if individuals vary in the value of the maximand (i.e. fitness or inclusive fitness) they attain, then the change in every gene frequency equals its covariance across individuals with the value of the maximand. Do these links imply that natural selection will drive a population towards a point at which all agents are behaving optimally, or at least close to optimally? Not necessarily.

The difficulty is that the links are neutral as to whether or not the dynamics of evolution display convergence towards any point at all, optimal or otherwise. In a favourable scenario for MAX-D, they do: every gene that is favoured by selection in the short term is also a step along the way to the optimal phenotype. Selection steadily accumulates improvements, leading to ever closer approximations of perfection. But there are well-known models in which changes in gene frequency show no pattern of steady accumulation: one gene spreads, then another, then another, each outcompeting its immediate rivals, but with no cumulative effect on the closeness to optimality (or otherwise) of their bearers. Rock-paper-scissors

games, with their notorious cyclical dynamics, provide one example (Sinervo & Lively, 1996; Kerr *et al*., 2002). MAX-D, in effect, amounts to the assertion that convergence towards fitness-optimal phenotypes occurs reliably in nature, and that cases in which it does not occur are the exception rather than the rule.

Grafen's Links 2 and 3 do not bear either way on this issue. They are no less true in the unfavourable cases than in the favourable ones. In all cases, genes spread when they co-vary positively with fitness. And, in all cases, suboptimality implies room for phenotypes that would co-vary positively with fitness if introduced. But this does not tell us whether or not the long-run dynamics of gene frequency change will exhibit a pattern of steady accumulation of adaptive improvements. The links therefore give us no reason to think that favourable cases (in which such accumulation occurs) are any more likely than unfavourable ones (in which it does not). They are compatible with unfavourable cases being rare, but they are equally compatible with their being extremely common. The frequency of such cases in the real world—and hence the tenability of MAX-D—is a question on which Formal Darwinism has no bearing.

None of the foregoing discussion, I should add, is intended to imply that Grafen's links have no biological significance at all. Like Fisher's fundamental theorem, it may be that Grafen's fail to imply a biologically significant maximization principle and yet still provide important insights in other ways. One proposal, which I have criticized elsewhere, is that the links amount to a formal reconstruction of Darwin's argument in the *Origin of Species* (Grafen, 2014; Birch, 2014). Another proposal (made by Grafen, 2007, and also by Okasha & Paternotte, 2014) is that the links provide formal constraints on an acceptable concept of individual fitness, constraints that turn out to be non-trivially satisfied by Hamilton's (1964) notion of inclusive fitness. My claim, then, is not that Formal Darwinism has no value at all.

My claim is simply that it does not vindicate any of our four varieties of fitness maximization.

VI. THE LIMITS OF PURE THEORY

My aim in this article has been to examine two theoretical developments that initially appear to rescue a form of fitness maximization. In the case of Fisher's fundamental theorem, I argued that while the Price/Ewens 'new' interpretation does leave us with a true mathematical result, it is a result that does not provide support for MAX-A or MAX-B. At most, it supports a modified version of MAX-B ('MAX-B*'), but the biological significance of this principle remains doubtful. The story was similar in the case of Formal Darwinism. Here I argued that the links Grafen derives between the Price equation and an optimization programme do not support MAX-C or MAX-D. At most, they support a modified version of MAX-C ('MAX-C**'), but questions remain as to the biological significance of this principle.

It would be hasty to conclude from this that all four varieties of fitness maximization are false. It would be safer to conclude that they remain unsupported by theoretical population genetics. This, however, may still sound troubling to evolutionary ecologists who assume one or more of these varieties in their day-to-day work.

Is there any way to reconcile evolutionary ecology with population genetics? Although my conclusions so far have been sceptical, I would like to end on a more optimistic note. In Section II, we saw that fitness maximization in the sense of MAX-A and MAX-B occurs only under very special conditions. This looks like bad news for evolutionary ecology, and this has allowed alternative conceptions of fitness maximization (in terms of partial change and/or the 'individual as maximizing agent' analogy) to gain traction.

From a different perspective, however, the news is not so bad. The case for scepticism about fitness maximization goes like this: fitness is maximized only under special conditions, natural populations rarely meet these conditions, so we should not expect fitness to be maximized very often in nature. However, this argument looks different if recast in retrospective, rather than prospective, terms. It becomes: if fitness has in fact been maximized in a given population, then rare, special conditions must have obtained in its evolutionary past.

When formulated retrospectively, the argument is not so much a case for scepticism as the basis for an empirical research programme. We know that, in some cases, evolution by natural selection has led to traits that approximately maximize fitness within a set of feasible alternatives (Dawkins, 1986; Reeve & Sherman, 1993; Orzack & Sober, 2001; Davies *et al*., 2012). There must be some explanation for these traits, and one potential explanation is that, in these particular cases, the conditions required for natural selection to maximize fitness were at least approximately met. In this way, a theoretical understanding of the conditions under which natural selection maximizes fitness can inform empirical inferences about the evolutionary history of populations that currently approximate fitness maxima.

Ewens (2004) makes a suggestion along similar lines in relation to Fisher's fundamental theorem. He contrasts (as we have done) the new interpretation of the FTNS, on which the theorem concerns a mere partial change, with the traditional interpretation, on which it states that, given non-zero additive genetic variance in fitness, the mean fitness of a population will actually increase. This 'Mean Fitness Increase Theorem' (MFIT) may not have been what Fisher had in mind when formulating the FTNS, but it has taken on a life of its own in subsequent literature. Ewens (2004, p.67) proceeds to remark that:

> [I]t is possible that the MFIT, even though it is restricted to random-mating populations and [...] might not hold when fitness depends on a two-locus or more

> generally a multi-locus genotype, nevertheless gives a greater biological insight into the evolutionary process than the FTNS.

As Ewens observes, by studying the biological conditions under which evolutionary processes approximate the conditions required for the maximization of mean fitness, we may arrive at a deeper understanding of the relationship between selection and maximization than that afforded by the FTNS.

A parallel move is available in the case of Formal Darwinism. Grafen's project has so far been concerned with finding links between population genetics and optimization programmes that hold with as few assumptions as possible. Over the past decade, Grafen and colleagues have sought to relax several of the main assumptions in Grafen's (2002) paper, in the hope of uncovering universal truths about the workings of natural selection. Unfortunately, as with the FTNS, the generality of the links arguably comes at the expense of their biological significance. It is possible, however, to imagine a different approach. We could instead begin by specifying an optimization programme and asking: under what biological conditions would natural selection reliably lead to organisms that solve this programme? What are the constraints on the mating scheme, or on the structure of social interaction, or on the mutation rate? How malleable does the genetic architecture have to be? Such results would inform empirical inferences about the circumstances that have enabled optimal phenotypes to evolve, in cases in which they have. A brief remark by Grafen (2008, p. 430) gestures towards further work in this direction:

> Taken all together, these results [i.e. the four links] do not directly state conditions under which a population can be expected to exhibit optimality on the part of each of its individuals, nor even whether that is to be expected. Conclusions of this kind would require further dynamic assumptions, and that may well be a direction for future work.

Theory would still lie at the heart of such a programme, but it would be brought into closer contact with empirical work (something for which Orzack, 2014, has also called, echoing earlier calls from Lewontin, Ginzburg & Tuljapurkar, 1978, and May, 1973). Rather than seeking a purely theoretical vindication for a universal maximization principle—with a view to using such a principle as an unchallenged foundation for empirical projects—we would instead seek tentative, local, conditional maximization principles to use as sources of empirical hypotheses about ancestral mating schemes and inheritance systems, hypotheses which should then be subjected to empirical test (on this point I am very much in accord with Orzack & Sober, 1994*a*, *b*, 1996). This gives theory a crucial role, while embodying a more modest conception of what it can achieve. I cautiously suggest that this approach may provide a more effective way to reconcile evolutionary ecology with population genetics.

VII. CONCLUSIONS

(1) Evolutionary ecologists routinely view natural selection as a process of fitness maximization, but there is a long tradition of scepticism about this idea in population genetics.

(2) It is important to distinguish maximization principles that concern what happens at equilibrium from those that concern the direction of change, and it is important to distinguish those that concern population means from those that concern fitness-maximizing behaviour at the individual level. This leads to a four-part classification.

(3) On the Price/Ewens 'new' interpretation, Fisher's fundamental theorem of natural selection concerns 'partial change' rather than total change. But the partial change it describes should not be interpreted as the partial change attributable to natural selection. Consequently, the fundamental theorem does not support a biologically significant maximization principle.

(4) Grafen's Formal Darwinism project is an attempt to vindicate an 'individual as maximizing agent' analogy by proving formal links between gene frequency change and optimal strategy choice. In broad terms, the aim is to show that (given various assumptions) a population is at a stable equilibrium if and only if all individuals have optimal phenotypes.
(5) The relationship between Grafen's links and fitness maximization is complicated, because the absence of 'scope' and 'potential' for selection is neither necessary nor sufficient for a stable equilibrium. This casts doubt on the biological significance of the links.
(6) It may be a mistake to seek a universal maximization principle describing the action of natural selection. A better approach may be to identify specific biological conditions in which natural selection leads reliably to fitness-maximizing phenotypes. This can then inform empirical inferences about the evolutionary past of populations in which such phenotypes are found.

VIII. ACKNOWLEDGEMENTS

I thank Anthony Edwards, Alan Grafen, Rufus Johnstone, Tim Lewens, Samir Okasha, Cedric Paternotte, John Welch and two anonymous referees for their extensive and very helpful comments on earlier drafts. I also thank an audience at the University of Cambridge, where an early version of this paper was presented.

IX. REFERENCES

ASMUSSEN, M. A., CARTWRIGHT, R. A. & SPENCER, H. G. (2004). Frequency-dependent selection with dominance: a window onto the behavior of the mean fitness. *Genetics* **166**, 499–512.

BATTY, C. J. K., CREWE, P., GRAFEN, A. & GRATWICK, R. (2014). Foundations of a mathematical theory of Darwinism. *Journal of Mathematical Biology* **69**, 295–334.

BIJMA, P. (2010). Fisher's fundamental theorem of inclusive fitness and the change in fitness due to natural selection when conspecifics interact. *Journal of Evolutionary Biology* **23**, 194–206.

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

DAVIES, N. B., KREBS, J. R. & WEST, S. A. (2012). *An introduction to behavioural ecology*. Hoboken, NJ: Wiley-Blackwell.

DAWKINS, R. (1986). *The blind watchmaker: why the evidence of evolution reveals a universe without design*. New York, NY: W. W. Norton and Company.

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: Oxford University Press.

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

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

DOEBELI, M. & HAUERT, C. (2005). Models of cooperation based on the prisoner's dilemma and the snowdrift game. *Ecology Letters* **8**, 748–766.

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 (2nd edition)*. Cambridge: Cambridge University Press.

EDWARDS, A. W. F. (2007). Maximisation principles in evolutionary biology. In *Handbook of the philosophy of science: philosophy of biology* (eds MATTHEN, M. & STEPHENS, C.), pp. 335–347. Amsterdam: North-Holland.

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. & 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 *Adaptationism and optimality* (eds ORZACK, S. H. AND SOBER, E.), pp. 161–190. Cambridge: Cambridge University Press.

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.

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 (2nd edition)*. New York: Springer.

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

EWENS, W. J. (2014). Grafen, the Price equations, fitness maximization, optimisation and the fundamental theorem of natural selection. *Biology and Philosophy* **29**, 197–205.

FISHER, R. A. (1930). *The genetical theory of natural selection (1st edition)*. Oxford: Clarendon Press.

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. (1998). *Foundations of social evolution*. Princeton, NJ: Princeton University Press.

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

GARDNER, A. & GRAFEN, A. (2009). Capturing the superorganism: a formal theory of group adaptation. *Journal of Evolutionary Biology* **22**, 659–671.

GARDNER, A. & WELCH, J. J. (2011). A formal theory of the selfish gene. *Journal of Evolutionary Biology* **24**, 1020–1043.

GERITZ, S. A. H., MESZE, G. & METZ, J. A. J. (1998). Evolutionarily singular strategies and the adaptive growth and branching of the evolutionary tree. *Evolutionary Ecology* **12**, 35–57.

GRAFEN, A. (1984). Natural selection, kin selection and group selection. In *Behavioural ecology (2nd edition)* (eds KREBS, J. R. & DAVIES, N. B.), pp. 62–84. Oxford: Blackwell.

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

GRAFEN, A. (1999). Formal Darwinism, the individual-as-maximising-agent analogy, and bet-hedging. *Proceedings of the Royal Society B: Biological Sciences* **266**, 799–803.

GRAFEN, A. (2000). Developments of the Price equation and natural selection under uncertainty. *Proceedings of the Royal Society B: Biological Sciences* **267**, 1223–1227.

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

GRAFEN, A. (2003). Fisher the evolutionary biologist. *The Statistician* **52**, 319–329.

GRAFEN, A. (2006*a*). A theory of Fisher's reproductive value. *Journal of Mathematical Biology* **53**, 15–60.

GRAFEN, A. (2006*b*). 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. (2008). The simplest formal argument for fitness optimisation. *Journal of Genetics* **87**, 421–433.

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

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 *Evolution and rationality: decisions, co-operation and strategic behavior* (eds OKASHA, S. & BINMORE, K.), pp. 7–22. Cambridge: Cambridge University Press.

HAMMERSTEIN, P. & SELTEN, R. (1994). Game theory and evolutionary biology. In *Handbook of game theory with economic applications, Vol. 2* (eds AUMANN, R. J. & HART, S.), pp. 929–993. Amsterdam: Elsevier.

HEDRICK, P. W. (2011). *Genetics of populations (4th edition)*. Sudbury MA: Jones and Bartlett.

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

KERR, B., RILEY, M. A., FELDMAN, M. W. & BOHANNAN, B. J. M. (2002). Local dispersal promotes biodiversity in a real-life game of rock-paper-scissors. *Nature* **418**, 171–174.

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

LEWONTIN, R. C., GINZBURG, L. & TULJAPURKAR, S. (1978). Heterosis as an explanation for large amounts of genic polymorphism. *Genetics* **88**, 149–170.

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

MARROW, P., JOHNSTONE, R. A. & HURST, L. D. (1996). Riding the evolutionary streetcar: where population genetics and game theory meet. *Trends in Ecology and Evolution* **11**, 445–446.

MAS-COLELL, A., WHINSTON, M. D. & GREEN, J. R. (1995). *Microeconomic theory*. New York: Oxford University Press.

MAY, R. M. (1973). *Stability and complexity in model ecosystems*. Princeton, NJ: Princeton University Press.

MAYNARD SMITH, J. (1982). *Evolution and the theory of games*. Cambridge: Cambridge University Press.

MAYNARD SMITH, J. & PRICE, G. R. (1973). The logic of animal conflict. *Nature* **246**, 15–18.

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

MULHOLLAND H. P. & 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. & PATERNOTTE, C. (2012). Group adaptation, formal Darwinism and contextual analysis. *Journal of Evolutionary Biology* **25**, 1127–1139.

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

OKASHA, S., WEYMARK, J. A. & BOSSERT, W. (2014). Inclusive fitness maximization: an axiomatic approach. *Journal of Theoretical Biology* **350**, 24–31.

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*a*). How (not) to test an optimality model. *Trends in Ecology and Evolution* **9**, 265–267.

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

ORZACK, S. H. & SOBER, E. (1996). How to formulate and test adaptationism. *American Naturalist* **148**, 202–210.

ORZACK, S. H. & SOBER, E. (2001). *Adaptation and optimality*. Cambridge: Cambridge University Press.

PARKER, G. A. & MAYNARD SMITH, J. (1990). Optimality theory in evolutionary biology. *Nature* **348**, 27–33.

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.

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

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

REEVE, H. K. & SHERMAN, P. W. (1993). Adaptation and the goals of evolutionary research. *Quarterly Review of Biology* **68**, 1–32.

RICE, S. H. (2008). A stochastic version of the Price equation reveals the interplay of deterministic and stochastic processes in evolution. *BMC Evolutionary Biology* **8**, 262–279.

ROFF, D. A. (1992). *The evolution of life histories: theory and analysis*. New York: Chapman and Hall.

ROSENBERG, A. & BOUCHARD, F. (2010). Fitness. In *The Stanford encyclopedia of philosophy (Fall 2010 edition)* (ed. ZALTA, E. N.). URL (accessed 27/03/15): <http://plato.stanford.edu/archives/fall2010/entries/fitness/>.

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

SCHEUER, P. A. G. & MANDEL, S. P. H. (1959). An inequality in population genetics. *Heredity* **31**, 519–524.

SIGMUND, K. (1987). A maximum principle for frequency-dependent selection. *Journal of Mathematical Biosciences* **84**, 189–195.

SINERVO, B. & LIVELY, C. M. (1996). The rock-scissors-paper game and the evolution of alternative male strategies. *Nature* **380**, 240–243.

STERELNY, K. & KITCHER, P. (1988). The return of the gene. *Journal of Philosophy* **85**, 339–361.

WEST, S. A. & GARDNER, A. (2013). Adaptation and inclusive fitness. *Current Biology* **23**, R577–R584.

WRIGHT, S. (1932). The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proceedings of the Sixth International Congress of Genetics* **1**, 356–366.

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

Table 1. Four varieties of fitness maximization