Can fitness differences be a cause of evolution?

Grant Ramsey[§]

The causal status of fitness and natural selection is increasingly called into doubt in the philosophical literature. For example, Elliott Sober argues that the fitness of individual organisms is holistic; i.e., it is dependent on causally independent factors like census size. Others have argued that fitness differences cannot properly be causes of evolutionary change. In this paper I directly challenge the holistic conclusion, and thereby shed light on the debates over the causal status of fitness. I show that the causalists and statisticalists are—to a large degree—arguing past each other. There is a plurality of fitness concepts; some are legitimately causal, while others seem to be based, at least in part, on purely statistical parameters. But such facts say nothing about whether fitness *in general* is causal or statistical.

1. Introduction

Biological fitness is a foundational concept in the theory of natural selection. Natural selection is often defined in terms of fitness differences as "*any consistent difference in fitness* (i.e., survival and reproduction) *among phenotypically different biological entities*" (Futuyma 1998, 349). And in Lewontin's (1970) classic articulation of the theory of natural selection, he lists fitness differences as one of the necessary conditions for evolution by natural selection to occur. Despite this foundational position of fitness, there remains much debate over the nature of fitness, especially whether fitness differences can truly be said to cause evolutionary change. In recent years these debates have crystalized into two camps: (1) *causalists*, who see fitness differences as being one of the causes of evolutionary change, and (2) *statisticalists*, who deny the causal efficacy of fitness and instead hold that "fitness is a mere statistical, noncausal property of trait types" (Walsh 2010, 148).

The statisticalist/causalist debates do not represent a unified front, but instead constitute a diverse array of arguments. Some have argued that selection is a force (e.g., Sober 1984), while others have challenged the force metaphor (Matthen and Ariew 2002). Some have suggested that the causal conception of fitness is qualitative, not quantitative (Matthen and Ariew 2002), or that if quantified, it provides the wrong results (Walsh, Lewens, and Ariew 2002). And some have placed the causal nexus of selection at the individual level (Bouchard and Rosenberg 2004), while others place it at the population level (Millstein 2006).

The chief position that I will articulate and challenge below is put forward by Sober (2001) in one of the most important philosophical papers on fitness. The conclusion that Sober draws is that the fitness of an

§ Department of Philosophy, University of Notre Dame, 100 Malloy Hall, Notre Dame, IN 46556, USA
✉ E-mail: grant.ramsey@nd.edu

individual organism is dependent on features of its population that are causally independent of the organism's reproductive outcomes. Although Sober is not in the statisticalist camp, his conclusion, as we will see, seems to cause trouble for the causalist. Because of its dependence on the whole population, Sober describes fitness as being "holistic" (320). I will follow Sober in labeling this position "fitness holism."

In what follows, I critique fitness holism and then use the framework developed in this critique to draw out some broader implications for the causalist-statisticalist debates. My critique will be based on a crucial premise that has been accepted by some philosophers, but whose consequences have not been fully appreciated in debates over the causal nature of evolutionary theory:

> (*P1*) the fitness of a trait is not, in general, both the average fitness of individuals bearing the trait and a predictor of trait dynamics

Before beginning my critique of holism, I need to make some preliminary distinctions and lend support to *P1*.

2. Some Preliminaries

In order to say something meaningful about whether fitness is holistic, and whether it can be a cause of evolutionary change, it will be necessary to say something more general about the nature of fitness. To answer a question about the causal status of fitness we need to have some grasp on what fitness is. The difficulty here is that there are numerous debates—and numerous positions within these debates—that are on, or are related to, fitness. My task in this section is to show which of these debates can be set aside for present purposes, and which distinctions must be made for the debates that I will engage.

One debate in the philosophy of biology concerns the level—genes, organisms, populations, species, etc.—at which selection operates. Debates range from monisms (claiming, for example, that all selection is genic), to pluralisms (holding that there are multiple, equivalent ways of describing selection), to a hierarchical view holding that selection acts at one or more level simultaneously and that there is a fact of the matter which level is undergoing selection at any given time (see Okasha 2006 for a primer on these positions). These debates are related to discussions of fitness, since a claim that selection is operating at a particular level is (for some, at least) a claim that an entity at that level is a fitness-bearing entity. In what follows, I will remain as neutral as possible about these debates. Although I restrict my discussion to the fitness of organisms and their traits, this should not be read as a claim that these are the only legitimate fitness-bearing entities. Instead, I hope that the arguments about fitness here can be applied to all fitness-bearing entities, whether they be genes, populations, or entire species.

But given that trait and individual fitness will be the central fitness concepts discussed below, a few words of clarification are needed. I defer discussion of the fitness of individual organisms to Section 2 when discussing the propensity interpretation of fitness. Here I focus on the fitness of traits. The difficulty with trait fitness is that there are multiple internally consistent, yet incompatible, ways that it is commonly understood.

If one were interested in determining the fitness of a trait, it would seem sensible to record the fitness values of all of the individuals bearing that trait and take an average of these quantities. It is perhaps for this reason that philosophers generally understand trait fitness to be nothing but this average. Mills and Beatty (1979) define the fitness of types as "the *average* [individual fitness] of the members of each of the types under consideration" (276). And Sober holds that "the fitness value of a trait is the average of the fitness values of the individuals that have the trait" (2001, 310). Biologists also commonly understand fitness in this way: "[trait] fitness is the mean survival or reproductive success of all individuals having the same phenotype" (Schluter and Nychka 1994, 598).

Although average individual fitness might in some cases be a good *estimate* of the fitness of a trait, a *definition* of the fitness of a trait as the average of the fitnesses of individuals bearing the trait can lead to problems if one also expects trait fitness to be a good predictor of evolutionary dynamics. To see this, consider a more complete range of what 'trait fitness' might mean. The three main ways of understanding

trait fitness are: (*T1*) the fitness impact (on individuals) of the trait; (*T2*) the average fitness values of individuals possessing the trait; or (*T3*) the tendency of the trait to spread through the population. *T1* is often an implicit assumption about trait fitness and is rarely explicitly stated, unlike *T2*, which, as quoted in the previous paragraph, is how trait fitness is often defined. *T3* also is an explicitly defined understanding of trait fitness from the field of population genetics. Kimura and Crow (1970), for example, note that the number of individuals in the next generation (of a discrete model with non-overlapping generations) may be computed by the equation:

$$(2.1) \qquad N_t = w N_{t-1}$$

They go on to say that "[w]e regard *w* as a measure of both survival and reproduction […] We call *w* the Darwinian fitness, or simply the fitness" (5). Equation (2.1) clearly links fitness directly to population dynamics, as suggested by *T3*.

It can be seen easily that these three notions of trait fitness are not equivalent and, furthermore, that one cannot infer high or low trait fitness in one of these senses from high or low values in the others. Consider first the relationship between *T2* and *T3*. If a trait is passed on with less than perfect fidelity, the fitness of the trait *sensu T3* can be lower than the average of the fitnesses of individuals bearing the trait (i.e., *T2*). Similarly, if the trait can be passed on to (or induced in) others in the population through non-genetic means, then the trait fitness *sensu T3* could have a higher value than the average fitness of the individuals. The lack of equivalence of *T2* and *T3* shows that trait fitness cannot be equivalent to both. It is for this reason that *P1* (the fitness of a trait is not, in general, both the average fitness of individuals bearing the trait and a predictor of trait dynamics) is indeed true. Although *T2* and *T3* are coherent concepts, their lack of equivalence implies *P1*. There is thus not a single concept of trait fitness that can be both *T2* and *T3* simultaneously.

More subtly, the value of *T2* is obtained by reducing the diversity of individual fitness values in a population to a single scalar quantity, and it does so via the arithmetic mean. It is not clear if one is attempting to obtain a value that maximally predicts population dynamics that one should choose the arithmetic mean. If we consider the distribution in the fitness values of the individuals bearing the trait, the evolutionary dynamics are going to be a function of various properties of this distribution, such as its variance and skew, to which the arithmetic mean is not sensitive. Other statistics on individual fitness values may be better predictors of population dynamics, such as the geometric mean (Lewontin and Cohen 1969). Thus, one could support the claim that trait fitness is a mathematical function of individual fitness values without holding that this function is an arithmetic mean.

The relationship between *T1* and *T2* depends on how we understand the vague concept of the "fitness impact of the trait." If we take *T1* to be the counterfactual difference in the fitness of a particular individual with and without the trait, then this concept will be clear and unambiguous. But the problem with identifying *T1* and *T2*, then, is that traits will not always have homogeneous effects on an individual's fitness. If a trait doubles the number of offspring for half of the population but kills the other half at birth, then fitness *sensu T2* will be no different from the *T2* fitness of a neutral trait. Thus, the impact of a trait on individuals—fitness *sensu T1*—can change in the face of invariant *T2* fitness. (Incidentally, *T2* will not be equivalent to *T3* for such a case either.) The distinction between these ways of understanding trait fitness will play an important role in what follows. With this in hand, we are now prepared to analyze the argument for holism.

3. Fitness Holism

Sober (2001) concludes that fitness is a "holistic" quantity that, "reflects a property of the containing population—namely, its census size—that may have no effect on the organism's reproductive behavior" (320).[1] If this is true, one of two conclusions should be drawn: (a) individual fitness differences cause evolutionary change (via differences in reproductive outcomes) despite the fact that these fitness differences are at times entirely due to factors that have no causal influence on the reproductive outcomes of

fitness-bearing individuals;[2] or, (b) fitness differences are not in fact a cause of evolution. Neither of these conclusions are attractive—the former is uncanny and the latter is nihilistic—and they should not be accepted without an airtight argument backing them up. The argument offered by Sober for holistic fitness is flawed, but it is flawed in a very interesting and important way: it is implicitly based on the denial of *P1*. Before showing that the holism argument assumes that it is false, let's examine in more detail Sober's argument and the account of fitness that he is challenging.

The main way that philosophers have understood and articulated the causal account of fitness is what is known as the "propensity interpretation of fitness" (Brandon 1978; Mills and Beatty 1979). The propensity interpretation of fitness (henceforth *PIF*) holds that fitness is a probabilistic propensity to produce offspring and that this propensity is one of the causes of evolutionary outcomes. The *PIF* was introduced because of two considerations. First, fitness cannot be identified with any particular kind of trait like fleetness or massiveness, since for one organism being fleeter or more massive may be advantageous, while for others it may be disadvantageous. Second, organisms with equivalent fitness values do not always have identical reproductive fates—a high or low fitness is no guarantee of high or low reproductive success. This second point is important because if fitness were identified with actual reproductive outcomes, there would be no sense in which fitness could causally explain these outcomes, and fitness would thus not be part of the causes of evolutionary change. The propensity interpretation of fitness takes fitness to be based on a distribution of potential outcomes, not on actual outcomes.

If fitness is based on a distribution of potential outcomes instead of actual outcomes, this distribution has a number of important qualities. One of these qualities is the distribution's variance. Populations can (and generally do) include conspecifics of different types that differ in the variance in their distribution of potential reproductive outcomes. This variance within a generation in potential reproductive outcomes is labeled "within generation variance" and it is this variance that leads Sober to the conclusion that fitness is holistic. The holistic conclusion comes from a reflection on the way in which variance influences fitness. Sober draws on the work of biologist John Gillespie who, in a 1974 paper, argued for the following claim: if there is within-generation variance in offspring number, then the fitness of a trait (ω) is a function of properties of the distribution in reproductive success—the arithmetic mean of the distribution (μ) and the distribution's variance (σ^2)—as well as the population size (n). In particular, he proposed that

$$(2.2) \qquad \omega_i = \mu_i - \sigma^2_i / n$$

This equation implies that fitness decreases with increasing variance. But this variance matters less as the population increases in size. As the population size approaches infinity, the effect of variance will vanish. But for any finite (i.e., real) population, the fitness of a trait will be a function of population size.

The dependence of fitness on population size is how Sober is able to derive his radical conclusion that the fitness of individual organisms is a function of the existence of other individuals, even if these individuals are causally independent of one another. For example, consider a population containing two members each of two types of individuals, A and C. As Sober argues, "The two A's and two C's in my example might be four cows standing in the four corners of a large pasture; the two A's have two calves each, whereas each of the C's flips a coin to decide whether she will have one calf or three. The cows are causally isolated from each other, but the fitnesses of the two strategies reflect population size" (2001, 316). Sober's conclusion is that the fitness of each of these individuals is a function of how many others there are in the population (n), regardless of whether they causally interact with one another—individual fitness is thus holistic.

We should pause to consider and reject a quick way of dismissing Sober's argument. His argument, it might seem, is a nonstarter because is has an unrealistic assumption about population structure. That is, Sober makes two seemingly incompatible statements: (1) A and C individuals are in the same population; and, (2) A and C individuals do not causally interact. If one takes a population to be a collection of causally interacting conspecifics, then A and C individuals are either causally interacting population members, or they are not in the same population. Sober's conclusion should be rejected because it is based on incompatible premises.

This rejection of Sober, however, is too fast and misses his point. His unrealistic assumptions about the nature of populations are not necessary for his argument to work. It could be that A and C individuals do interact in some fitness-affecting ways. It is nevertheless true that the number of individuals in the population is a component of the fitness of their traits (assuming a *T3* conception of trait fitness). Sober's toy example was crafted to provide an example of fitness differences being *entirely* due to *n*. But one could easily provide a more realistic example in which fitness is still affected by changes in *n* even though changes in *n* correspond with other changes as well.

For example, if we hold a more realistic premise that the individuals in the population interact, then an increase in the number of individuals will modify such things as the frequency of mating opportunities and the degree of competition over key environmental resources (e.g., food). Such changes are surely the sort of thing that can change the fitness of the organism's traits. But, even if increasing *n* has these effects, it is nevertheless true that the increase in *n* still has an effect on trait fitness in and of itself, independently of these other effects. That is, it is still true that the effect of variance on trait fitness changes with *n* even if changes in *n* track other trait fitness component changes as well. Therefore, the unrealistic nature of Sober's toy argument is not its downfall.

With this caveat, it appears that Sober's argument is convincing. If it were sound, then it would have profound implications for how we understand the causal structure of natural selection. But the argument is not sound because it relies on a false premise about the relationship between the fitness of a trait (or type) and the fitness of individuals bearing that trait (or being of that type). The two key premises that Sober uses in his argument for holism are: (H_1) the fitness of a trait is, in some instances, a function of population size; and, (H_2) some individuals in a population bearing a particular trait may not have any causal interaction with other individuals in the population bearing that trait. I will not contest the truth of these premises (with the above caveat regarding H_2). The conclusion that he is drawing—an individual's fitness is a function of some properties that have no causal effect on that individual—seems to follow easily from the premises. But notice that H_1 is about *traits*, whereas H_2 and the conclusion are about *individual organisms*. For the argument to go through, trait fitness values must be a direct function of individual fitness values. But as I will now argue, if trait fitness is supposed to be tied to trait dynamics, no such function exists.

Recall the distinction made above between the various concepts of trait fitness. Given that *T1– T3* are not equivalent, which of these notions of fitness is at play in the argument for holism? The problem is that the holism argument is employing both *T2* and *T3* without noting that they are not equivalent. Sober explicitly endorses *T2* and for this reason holds that one can move between trait and individual fitness in an unproblematic way, allowing the holistic conclusion to follow from H_1 and H_2. But he is also taking trait fitness to be *T3*. For example, he says that, "[t]he fitness values of traits, along with the number of individuals initially possessing each trait, are supposed to entail the expected frequencies of the traits one or more generations in the future (if selection is the only force influencing evolutionary change)" (2001, 317).

To see how the holistic conclusion appears to follow, consider again Sober's example of four cows in a field, two each of A and C types. Let's flesh out this example and say that the offspring of an A is always an A and that the offspring of a C is always a C. Now let's say that each A has a probability of 0.5 of having zero offspring and a probability of 0.5 of having two offspring. Type C individuals have a probability of 1.0 of having one offspring.[3] What happens to the fitness of the traits of being an A or a C if the population doubles? In other words, what happens to the expected proportions of A and C when we change the population size? In a population of one A and one C individual, the expected A:C trait frequency in the next generation is 1:2.[4] If we double the number of A and C individuals, the ratio changes to 1:1.4. If we continue to increase population size, the ratio will approach the limit of 1:1. This is not controversial and is a direct consequence of within-generation variance in potential reproductive output. It is in this sense that trait fitness depends on *n*.

To achieve the holistic conclusion, one needs to show that individual fitness also depends on *n*. The ω in equation (2.2) is clearly trait fitness, and this is how both Gillespie (1974) and Sober (2001) interpret it. What argument can be made for this equation also applying to the fitness of individuals? One argument would be that trait fitness must be a direct consequence of individual fitness and nothing else, and that because *n* modulates trait fitness, it must also modulate individual fitness. But such an assertion would be question-

begging if there were not stronger, independent reasons for linking *n* and individual fitness. A possible argument could be that we want individual fitness to be maximally predictive of trait dynamics. This is tantamount to arguing that we want individual fitness to maximally track *T3* values. But recall that there is already a rift between *T2* and *T3*. Thus, there also will be a rift between individual fitness and *T3*: features like imperfect trait transmission change trait fitness *T3* without changing individual fitness. This shows that we need idealizations like perfect trait transmission to get from individual fitness to *T3*. Similarly, another idealization that is needed to link individual fitness with trait dynamics is a large population size: trait fitness *T2* is equivalent to *T3* only if the population size is infinite, though finite but large populations can bring about a tight correlation between *T2* and *T3*. Population size, like the imperfect transmission of traits, is just one of the factors that complicates trait dynamics. It should thus be regarded in the same way as imperfect trait transmission, i.e., a factor independent of individual fitness but important for—and can cause predictable, directional changes in—trait dynamics. That is, if the heritability of a trait is not included in individual fitness, then, *a fortiori*, population size should not be included either.[5]

To support this argument, consider a coin flipping analogy in which there are two kinds of coins, A coins and C coins. A game can be played with the coins in which some number of A and C coins are flipped once per round during each of a series of rounds. At the end of each round, the coins are removed and the player is paid back (with A and C coins) based on the outcomes of the coin flips. The game begins with an equal number of A and C coins and the coins are repeatedly flipped until the ratio of A:C meets or falls below 1:2. Now consider the question of how many rounds are expected to occur in the game. In order to answer this question, there are four crucial facts that we need to know: (*i*) the probability of the coins landing on each side; (*ii*) the standard paybacks from the results of each flip; (*iii*) the probability (and kind) of payback errors; and (*iv*) the starting number of coins. Which of these four are elements of the fitness of individual coins and which are part of the fitness of the A and C traits? Applying his stance on organismic fitness, Sober's answer would be that (*i*), (*ii*), and (*iv*) are part of both individual and trait fitness,[6] but that (*iii*) is part of neither. What is the justification for excluding (*iii*) but not (*iv*) from individual fitness? By excluding (*iii*), we are giving up on accurate predictions of trait dynamics and such predictions cannot therefore be the reason for including (*iv*) in individual fitness. Instead, I suggest that individual fitness includes (*i*) and (*ii*) only, and that predicting trait dynamics requires the addition of factors outside of individual fitness, in this case (*iii*) and (*iv*). Support for this way of understanding individual fitness is that by including only (*i*) and (*ii*), it fulfills key desiderata of individual fitness: being distinct from inheritance, being a function of properties of (and only of) the individual and its local environment, and helping to predict future trait dynamics. Because of this, there is a burden of proof for the holist to support the inclusion of (*iv*) in individual fitness. Although there is good reason to support the inclusion of (*iv*) in *T3*, it is question-begging to simply assume that individual fitness also includes (*iv*). And not only is this assumption unwarranted, it is also troubling. As Sober himself points out, the holistic conclusion is radical and quite mysterious—individual organisms have fitness values that are based on a feature (*n*) causally independent of their reproductive outcomes. This is like arguing that the fitness of a particular A or C coin is a function of how many coins have been (or are going to be) flipped, and is therefore "holistic." We should instead just sidestep the holistic conclusion (and the problems entailed by it) by considering the fitness of A and C coins to be a function of (*i*) and (*ii*) only, and that questions about the expected ratios of A:C are answered only by considering additional information, such as (*iii*) and (*iv*).

If *T1*–*T3* are non-equivalent, and if therefore trait fitness values are not (always) a direct function of individual fitness values, we can now see how the argument for holism fails. Recall the premises: (*H1*) the fitness of a trait is, in some instances, a function of population size; and, (*H2*) some individuals in a population bearing a particular trait may not have any causal interaction with other individuals in the population bearing that trait. It is now clear that in spite of the truth of these premises, the conclusion that Sober is drawing—"an organism's fitness is not a propensity that it has" (320)—does not follow any more than the analogous conclusion that because the probability of achieving an A:C ratio of 0.5 in 100 rounds varies with the number of coins flipped, the fitness of an individual coin is based on causally independent factors (like the number of coins being flipped).

The argument for holism takes predictions about the future proportion of a trait in a population, considers this to be trait fitness, assumes that trait fitness is both an average of individual fitness values and a

predictor of future trait proportions (both *T2* and *T3*), and draws the conclusion that individual fitness is holistic. Add *P1*, which denies this assumption, and the chain of reasoning breaks; with the chain of reasoning broken, individual fitness is saved from holism, as is *PIF*. What are the implications of the these conclusions and distinctions for the causalist-statisticalist debates?

4. Broader Implications for the Causalist-Statisticalist Debates

Over the last decade, a series of papers has been published in support of one or more of the following related claims: (1) selection is not a cause of evolution, (2) fitness is not a cause of evolution, or (3) drift is not a cause of evolution (e.g., Matthen and Ariew 2002, 2009; Walsh, Lewens, and Ariew 2002; Ariew and Ernst 2009; Pigliucci and Kaplan 2006; Walsh 2007, 2010). Although it is possible to argue for one of (1)–(3) without holding all of these positions (and some of the statisticalist papers just cited focus on only a subset of them), they generally come as a package. If, for example, one asserts (1) and also holds that natural selection operates through (or is explained by) fitness differences, then one must also hold (2). Thus, a position that these papers explicitly stake out (or tacitly imply) is that fitness should be understood as a statistic, as the rate of change over time in the representation of a trait in a population.[7] This notion of fitness is known as Fisherian fitness after Ronald Fisher, who dubbed this growth rate statistic the Malthusian parameter (Fisher 1930). In what follows I will label the statistical interpretation of fitness *SI* and the causal interpretation *CI*.

The *SI* is a hard pill to swallow—accepting it means giving up on causal explanations of evolutionary change due to fitness differences since, under the statisticalist framework, "fitness is a mere statistical, noncausal property of trait types" (Walsh 2010, 148). Like the holism position, the *SI* should not be accepted without an airtight argument backing it up. Therefore, it is important to explore the question of whether the above arguments and distinctions undercut or support the *SI*. To narrow the scope of this exploration, I will focus on two papers supporting the *SI*: one of the founding papers in the *SI* camp, Matthen and Ariew (2002), as well as a more recent paper by Ariew and Ernst (2009). Any critique of the positions in these papers thus does not constitute a comprehensive attack on the statisticalist position, but does nevertheless call into question some of the core argumentative strategies of the statisticalists.

4.1 Three Desiderata We Should Not Desire

Ariew and Ernst (henceforth, A&E) "identify three desiderata for an account of fitness that propensity theorists accept" (2009, 289). They then form a *modus tollens* argument against the *PIF*: if the *PIF* is correct, then the *PIF* must meet the desiderata. The *PIF* cannot meet the desiderata. Therefore, the *PIF* is not the correct account of fitness. I will argue that the desiderata are not ones that we should expect the *PIF* to meet, and therefore the first premise is false. I begin my critique with their third desideratum and work backwards.

A&E's third desideratum is: "(C) The fitness of a trait must be a function of the properties of the individual members of the population within their local environmental conditions" (291). Interestingly, this is tantamount to saying that trait fitness cannot be holistic. They justify the inclusion of this desideratum by claiming that "because the fitness of a trait is simply the fitness of individuals with that trait, we are also guaranteed that condition (C) is satisfied" (292). Their argument thus takes the premises that individual fitness is local, and that trait fitness is a function of individual fitness values, and draws the conclusion that trait fitness must also be local. The problem, however, is that the notion of trait fitness that A&E are arguing for is that of *T3*. And the argument against individual fitness holism shows that trait fitness, if it is to be predictive of trait dynamics (as suggested in A&E's desideratum A below), must include factors external to individual fitness. And one of these factors is population size. Because population size is (or can be) "non-local," trait fitness *sensu T3* can be non-local. It would be strange if individual fitness were similarly non-local. But, as we saw above, worries about the non-locality and holism for organismic fitness can be set aside.

Thus, if the "function" referred to in (C) is anything like the arithmetic mean employed in *T2*, then this is clearly not a desideratum for the *PIF*. If instead, the point is merely that trait fitness values must be derivable from the properties of individuals bearing the traits as well as their environmental contexts, then it is clearly (though trivially) a desideratum. This is true because if we include features like population size and

structure, we can enumerate all of the ingredients that go into *T3*. It seems, however, that this trivial interpretation of "function" is not what A&E are getting at in (C). If not, then this desideratum is clearly not one that "propensity theorists accept."

Now consider desideratum (B): "[An individual] fitness concept must enable us to compare the degree to which natural selection will favor the spread of one trait over another, alternative trait" (290). We can again see here the leap from individual to trait fitness *T3*—individual fitness, they are asserting, is supposed to give us predictions about trait dynamics. But individual fitness merely provides us with predictions about what will happen to individuals. Whether or not a trait will spread through the population is quite another matter. Having the trait find itself in fitter individuals at a disproportionately high frequency will help the trait spread, but will not guarantee its spread and, furthermore, such a trait is not necessarily fitter than another trait that is associated with less fit individuals (since the trait associated with the less fit could be passed on with higher fidelity, for example). The trait of "having disease *X*" could have a high fitness *sensu T3* in spite of the fact that the disease affects only the least fit in the population. A&E justify (B) by claiming that "[b]ecause the propensity interpretation equates the fitness of an organism with a particular number that can be greater or lower than another, every trait's fitness can be compared with others, satisfying condition (B)" (292). The above discussion clearly shows that while it might be true that one organism's fitness can be compared to another, and that the fitness of alternate traits can be compared, it does not follow that we can derive trait dynamics—or *T3*—from individual fitness alone. Thus, (B) cannot be a desideratum for the *PIF*.

Finally, consider desideratum (A): "[An individual] fitness concept must be able to explain why one trait is expected to be better represented in a population under the influence of natural selection" (290). We could read "be able to explain" as "help to explain" or "suffice to explain." If it is the former, then this is a desideratum of the *PIF*. But under this reading the fact that observations will often deviate significantly from predictions if ancillary data like heritability values are not used does not serve as a counterexample to the *PIF*. If instead we read it as "suffice to explain," then we can easily see that *P1* and the discussion in Section 2 shows us that this is not a desideratum. Again, individual fitness values alone do not give us trait dynamics.

By showing that these desiderata are not ones that a *PIF* proponent should hold, is the *PIF*, and therefore the causal conception of fitness, thereby supported? Instead of reading my critique of A&E as a critique of statisticalism and a support of causalism, I want to make it clear that things are a bit more complex. In fact, *P1* might appear to be a statisticalist, and not a causalist, premise. But as I suggest below (Section 5), the causalists and statisticalists are each correct about a particular conception of fitness. *T3* is modulated by *n*, but this does not mean that the *PIF* is not causal. Similarly, the coherence of a causal conception of individual fitness does not undermine the fact that particular renderings of trait fitness can have non-causal dependencies. Before we move to this larger discussion, let's examine another of the statisticalist arguments.

4.2 The Vices of Vectors

If selection can be understood properly as an evolutionary force, as Sober (1984) classically argued, then there is good reason to think that it can also be understood as a cause of evolutionary change. Thus, one strategy for undermining the causal efficacy of selection/fitness is to show that the force metaphor breaks down. Such an argument would by no means be a definitive argument against the causalist position, but will serve to constrain the resources of the causalists, pressing them to provide alternative, force-free conceptualizations of how selection/fitness can cause evolutionary change. Matthen and Ariew (2002) offer an argument against the force conception and hold that "[t]he disanalogy [between physics and evolutionary biology] is that, while force affords Newtonian mechanics the means to compare and add up the consequences of these diverse causes, fitness does not add up or resolve. This is why population geneticists are forced to estimate fitness by measuring population change" (68). They support this position with claims like the following:

> For example, you might learn that the optimal reproductive strategy with respect to sex determination is to produce male offspring when there are fewer males in the population, and females when there are fewer females. But this only tells you about the relative merits of strategies within a circumscribed set,

with other factors held constant. The analysis does not tell you whether producing offspring of the minority sex is more or less advantageous than other fitness-relevant things you can do; there is, generally, no way of combining the effects of a good strategy in this game, with good or bad strategies in other games (Matthen and Ariew 2002, 67).

In order to analyze and critique M&A's argument, we must first get some sense of what they even mean by vector addition and what concept of fitness they are employing. One clue follows the above quotation: "Suppose a certain species undertakes parental care, is resistant to malaria, and is somewhat weak but very quick. How do these fitness factors add up? We have no idea at all" (67). Here they speak of a "species" but presumably they are not talking about species fitness. It appears that they are getting at this point: if we endow an individual in a species with several traits like resistance to malaria, moderate quickness, etc., we have no idea what the net fitness effect would be. Although these traits may be good for some individuals, certain trait combinations may be deleterious.

First, if M&A are simply arguing that to determine the fitness of an individual organism, one cannot always merely sum the average fitness values of each of the individual's traits, they are correct because there are many cases where there would be systematic errors. For example, for a weak organism the trait of boldness might be fitness sapping, but for a strong organism boldness might be fitness boosting. (And this reinforces the point that *T1* is not, in general, equivalent to *T2* or *T3*.) Though nature abounds with examples like these, there are good reasons to think that the situation for biologists is not as grave as M&A suggest. One might argue (*contra* M&A's premises) that biologists are able to do reasonably well at inferring fitness advantages of suites of traits given detailed data on how the traits affect organisms. Brandon and Ramsey (2007) took this stance against M&A and I will not repeat their arguments. Instead, I would like to grant M&A their premises and see—keeping *P1* and the arguments above in mind—whether their conclusion follows.

The key question is this: if it is true that one cannot take trait fitness values for the traits of an individual organism, sum them, and thereby obtain the fitness value of the organism, does this mean that fitness should be understood in a merely statistical way? The answer is no because vector summation does not work even under the assumption that the causal interpretation (*CI*) is a viable interpretation of fitness. The *CI*, in a nutshell, involves the following. Organisms have fitness values, and these values are about the propensity of producing descendants,[8] and an organism's fitness is causally dependent on (at least some of) the traits it possesses. Natural selection for these traits occurs via this causal dependence, and evolution by natural selection will occur when the selected traits are heritable. This is the bare-bones picture of the *CI*. To flesh things out a bit more and see the implications for vector addition, let's consider in more detail the nature of trait fitness that M&A endorse and what the relationship is between trait and organismic fitness under this view.

A&E and Sober explicitly endorse *T2*, but also hold *T3*; A&E also assume *T1*. It is not a problem to hold that there are many senses of trait fitness and that each of them has a useful domain. But we have seen that their arguments are undermined by asserting that individual fitness must simultaneously underwrite *T1*–*T3*, and that its failure to do so evinces deep problems with individual fitness. How do M&A understand trait fitness? In their words, "predictive fitness is a statistical measure of evolutionary change, the expected rate of increase (normalized relative to others) of a gene, a trait, or an organism's representation in future generations" (56). Fitness for them is thus a statistical measure of evolutionary change, and the measure of choice is an expectation value. This expectation value, projected into the future, is how M&A get the "predictive" in predictive fitness. Thus, if M&A's conception of trait fitness is any of the three conceptions described above, it is a form of *T3*. If this is true, then the important question concerns the relationship that organismic fitness bears to *T3* under the *CI*.

Under the *CI*, there is no way to infer *T3* from organismic fitness alone (even if one knew the fitness values of each organism possessing the trait), and that given *T3* values for all of the traits possessed by the organism, one cannot deduce the organism's fitness. This paradoxically seems to contradict the characterization of the *CI*, which describes organismic fitness as being causally dependent on the traits the organism possesses. The resolution of this paradox involves recognizing two things. First, an organism's fitness is dependent on its traits, not (merely or directly) on the *T3* values of its traits. Second, *T3* concerns

the change in the representation of traits in a population. And traits can spread not just because the organism possessing them has a high fitness, but the traits can achieve their high *T3* values by sacrificing the fitness of the organisms that possess them. An organism possessing many virulent diseases will have more fit traits (*sensu T3*) than an otherwise identical organism lacking these traits. If one balks at calling the possession of a transmissible disease a "trait" (because it lacks a genetic basis or is not inherited), then one can consider examples like the mouse *t* haplotype. Its possession is detrimental to organisms (it is lethal when homozygous), yet it has spread through the mouse species *Mus musculus* via the mechanism of meiotic drive, achieving population frequencies as high as 40% (Morita et al. 1992). This shows how a high *T3* trait can be associated with low organismic fitness. High organismic fitness can also be associated with a low *T3*. An organism might have a high fitness phenotype because of being an ideal height, say, but if height is a polygenic trait (e.g., if it is a function of several genes on different chromosomes), then the trait will have fairly low *T3*, since its spread through the population will be hindered by its polygenic nature. One observation that we can make from this is that organismic fitness is a distinct concept from inheritance, whereas *T3* depends crucially on inheritance.

Another way to analyze M&A's argument is to identify their "fitness factors" like parental care and resistance to malaria with *T1* values. Understood this way, the *T1* values of these traits do not provide individual fitness values because $T1 \neq T2$, and they do not provide trait dynamics because $T1 \neq T3$. We can now see precisely why trait fitness values do not add up like Newtonian vectors, as M&A suggest, and that this in no way implies the truth of the *SI* or the falsity of the *CI*. Does this mean that the statisticalists are wrong and the causalists are correct? As we will see, the *SI* is largely correct for one conception of fitness, but because this is distinct from the conception of fitness that the causalist is focused on, the *SI* and *CI* proponents are, to a large degree, arguing past each other.

5. Toward a Truce in the Fitness Wars

Let us take stock. We have seen that the fitness of organisms is not holistic as suggested by Sober, though some conceptions of trait fitness are indeed holistic. And Ariew and Ernst offered three criteria that are supposed to serve as desiderata for the causal conception of fitness, but the causalist should not regard these as desiderata. Finally, the fact that trait fitness values do not add like vectors serves neither as evidence for or against the *SI*.

What then are the implications of the above arguments for the *SI-CI* debates? Is fitness a cause of evolutionary change or not? The answer is both yes and no. If organismic fitness is understood as a probabilistic propensity to produce offspring, then, so long as propensities are taken to cause their effects, organismic fitness can clearly cause population-level changes in trait frequencies (i.e., organismic fitness can be a cause of evolutionary dynamics). But organismic fitness is limited. If an organism's fitness is based on its propensity to survive and reproduce, then organismic fitness makes predictions about the number of descendants an individual is likely to leave behind, but it does not imply what traits the descendants are likely to bear. Thus, one cannot derive trait dynamics from individual fitness values alone even if provided fitness values for each individual in the population, plus data on which individuals bear which traits. Organismic fitness is thus a cause of evolutionary change, but a limited one. Changes in population structure or size, or changes in mutation rates or heritability values, can modify trait dynamics without modifying the fitness of individual organisms.

In order to see more clearly how individual fitness can be linked with evolution, let's begin with an idealized population in which all the fitness concepts discussed above converge. Consider a population that is effectively infinite in size and is composed of asexual organisms of two types, A and C, which reproduce with perfect fidelity. The two types differ with respect to one trait only and the possession of this trait has a homogeneous effect on the individuals bearing it; that is, being an A and being born in one part of the population instead of another, or at one moment instead of another, does not affect A's fitness. In such a population, trait fitness $T1-T3$ are equivalent; the fitness impact of being an A will be directly linked to the average fitness values of the A's, which will be directly linked to the expected proportion of A's in future generations. Thus, in such a case, individual fitness is directly causally responsible for the changes in the proportion of A's and C's in the population (i.e., evolution).

It is easily seen that as we relax these idealizations toward the states of actual populations, there is no qualitative break; there is no point at which organismic fitness suddenly stops mattering for trait fitness. What occurs is a gradual disassociation of the two: as the population size reduces, population size will increasingly bear on trait dynamics and weaken the link between trait dynamics and individual fitness. Individual fitness is still a cause of evolution in small populations, but such populations introduce a significant intervening factor—census size—that has an effect on trait dynamics. Similarly, if trait transmission occurs without perfect fidelity (e.g., A's sometimes produce C's), then the fitness that individual A's have will still bear on the dynamics of A and C types. It is merely the case that the fitnesses of the individual A's and C's will do a poorer job of predicting future trait dynamics. In the extreme case in which each A and C produce either an A or a C with a probability of 0.5, it is true that individual fitness will not be a good predictor of evolutionary change. But this is hardly surprising, since this would not count as a case of evolution by natural selection.[9] Thus, individual trait differences provide causal explanations for trait dynamics only for systems undergoing evolution by natural selection, which is exactly what we would like individual fitness to be able to accomplish.

The *CI*, or at least one interpretation of it, is therefore defensible. Does this mean that the *SI* is therefore undermined? We have seen that there are a multitude of fitness concepts in circulation, four of which were highlighted (one version of organismic fitness plus *T1–T3*), and in order to make statements like "fitness is merely statistical, not causal," one first needs to make clear which of these concepts is in use. Many of the points that the statisticalists make do apply to *T3*. The values of *T3* do vary with *n* and appear to be "holistic." Thus, one might concede that the statisticalists are largely correct regarding *T3*. But this does not imply that all conceptions of fitness are "merely statistical." A *T3* with some statistical elements can exist alongside a causal conception of organismic fitness.

6. Conclusion

This paper has attempted to show that the fitness of organisms is not holistic, and that organismic fitness can be understood legitimately as one of the causes of evolutionary dynamics. But I also have shown that there is a plurality of fitness concepts and that both *SI* and *CI* arguments, to the extent that they are sound, are sound only for a subset of these concepts. One might assert that the fitness of organisms is causal, but that fitness *T3* is (to some degree, at least) statistical and holistic. There is no contradiction entailed in making these assertions. This implies the overarching conclusion that the *SI* and *CI* proponents are, to a large degree, arguing past each other. Conclusions about the statistical nature of *T3*, for example, simply do not serve as evidence for the statistical or holistic nature of organismic fitness. Thus, my hope is that this paper will help to show which of the debates over the causal status of fitness are vapid, and which represent genuine points of tension.

Literature cited

Ariew, A., and Z. Ernst. 2009. What fitness can't be. Erkenntnis 71: 289–301. doi:10.1007/s10670-009-9183-9

Bouchard, F., and A. Rosenberg. 2004. Fitness, probability, and the principles of natural selection. The British Journal for the Philosophy of Science 55: 693–712. doi:10.1093/bjps/55.4.693

Brandon, R., and G. Ramsey. 2007. What's wrong with the emergentist statistical interpretation of natural selection and random drift? In: The Cambridge Companion to the Philosophy of Biology. Ed. D. Hull, M. Ruse. New York: Cambridge University Press.

Brandon, R.N. 1978. Adaptation and evolutionary theory. Studies in History and Philosophy of Science Part A 9: 181–206. doi:10.1016/0039-3681(78)90005-5

Fisher, R.A. 1930. The Genetical Theory of Natural Selection. Oxford: Oxford University Press.

Futuyma, D.J. 1998. Evolutionary Biology. 3rd ed. Sunderland, MA: Sinauer Associates.

Gillespie, J.H. 1974. Natural selection for within-generation variance in offspring number. Genetics 76: 601–6.

Kimura, M. and J.F. Crow. 1970. An Introduction to Population Genetics Theory. Caldwell, NJ: Blackburn.

Lewens, T. 2010. The natures of selection. The British Journal for the Philosophy of Science 61: 313–3. doi: 10.1093/bjps/axp041

Lewontin, R.C. 1970. The units of selection. Annual Review of Ecology and Systematics 1: 1–18. doi: 10.1146/annurev.es.01.110170.000245

Lewontin, R.C. and D. Cohen. 1969. On population growth in a randomly varying environment. Proceedings of the National Academy of Sciences USA 62:1056–1060. doi:10.1073/pnas.62.4.1056

Matthen, M. and A. Ariew. 2009. Selection and causation. Philosophy of Science 76: 201–24. doi: 10.1086/648102

Matthen, M. and A. Ariew. 2002. Two ways of thinking about fitness and natural selection. The Journal of Philosophy 99: 55–83. doi:10.2307/3655552

Mills, S. K. and J. H. Beatty. 1979. The propensity interpretation of fitness. Philosophy of Science 46: 263–86. doi:10.1086/288865

Millstein, R.L. 2006. Natural selection as a population-level causal process. The British Journal for the Philosophy of Science 57: 627–53. doi:10.1093/bjps/axl025

Morita, T., H. Kubota, K. Murata, M. Nozaki, C. Delarbre, K. Willison, Y. Satta, M. Sakaizumi, N. Takahata, and G. Gachelin. 1992. Evolution of the mouse *t* haplotype: Recent and worldwide introgression to *Mus musculus*. Proceedings of the National Academy of Sciences USA 89: 6851–5. doi: 10.1073/pnas.89.15.6851

Okasha, S. 2006. Evolution and the Levels of Selection. New York: Oxford University Press. doi:10.1093/acprof:oso/9780199267972.001.0001

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

Schluter, D. and D. Nychka. 1994. Exploring fitness surfaces. American Naturalist 143: 597–616. doi: 10.1086/285622

Sober, E. 2001. The two faces of fitness. In: Thinking about Evolution: Historical, Philosophical, and Political Perspectives. Eds. R.S. Singh, C.B. Krimbas, D.P. Paus and J. Beatty. New York: Cambridge University Press.

Sober, E. 1984. The Nature of Selection. Chicago: University of Chicago Press.

Walsh, D.M. 2007. The pomp of superfluous causes: The interpretation of evolutionary theory. Philosophy of Science 74: 281–303. doi:10.1086/520777

Walsh, D.M. 2010. Not a sure thing: Fitness, probability, and causation. Philosophy of Science 77: 147–71. doi:10.1086/651320

Walsh, D.M., T. Lewens, and A. Ariew. 2002. The trials of life: Natural selection and random drift. Philosophy of Science 69: 452–73. doi:10.1086/342454

Notes

1. Sober is using 'reproductive behavior' broadly, meaning the reproductive outcomes that individuals have (or probably will have), and not something narrower like mating behavior. To avoid confusion, I will use the phrase 'reproductive outcomes.'
2. For the sake of simplicity, I will henceforth speak of "one individual not causally interacting with another." But by this I mean that one individual's reproductive success is in no significant way causally influenced by the presence or features of the other individual. There will of course always be some causal link between the individuals, even if this is but a slight gravitational attraction.
3. Such an example is not realistic, but this is the simplest case that is needed to make this point. More complex cases, as can be easily seen, exhibit the same phenomenon.
4. For how expected trait frequencies are calculated, see Sober (2001, 314).
5. Lewens (2010) makes the case that "[w]e should not expect any good principled answer to the question of which elements of some evolutionary process should count among contributors to fitness, hence we should not expect any principled account of how we should understand the force of selection" (314). I hope to convince the reader otherwise by arguing that there are principled reasons for excluding population size from individual fitness determinations, but not from certain conceptions of trait fitness, like *T3*.

6. To draw a close analogy with Sober's cow example, the A coins could have a 0 on one side and a 2 on the other, the C coins could have a 1 on both sides, and the paybacks for each coin could be two A coins for each "2" and one C coin for each "1."

7. In arguing for a statistical interpretation of fitness, the statisticalists often acknowledge the coherence of the concept of (causal) organismic fitness, but they dismiss the theoretical use of such a concept in the theory of natural selection. Matthen and Ariew (2002), for example, distinguish "informal fitness" (a causal, individual organismic fitness) from "predictive fitness" (a statistical measure of population-level change). Only the latter, they argue, is the theoretically important fitness concept and it is this concept that they discuss and analyze.

8. This is vague but specifying what, exactly, this means (e.g., what counts as a descendant) is well beyond the scope of this paper.

9. Since one of the necessary conditions for evolution by natural selection is the inheritance of traits (Darwin 1859; Lewontin 1970).