*Brit. J. Phil. Sci.* **0** (2013), 1–20

# Organisms, Traits, and Population Subdivisions: Two Arguments against the Causal Conception of Fitness?

Grant Ramsey

ABSTRACT

A major debate in the philosophy of biology centers on the question of how we should understand the causal structure of natural selection. This debate is polarized into the causal and statistical positions. The main arguments from the statistical side are that a causal construal of the theory of natural selection's central concept, fitness, either (i) leads to inaccurate predictions about population dynamics, or (ii) leads to an incoherent set of causal commitments. In this essay, I argue that neither the predictive inaccuracy nor the incoherency arguments successfully undermine the causal account of fitness.



## 1 Introduction

A central debate in the philosophy of biology concerns the nature of fitness, natural selection, and drift. One side in this debate holds that organisms have fitness values and that heritable differences in these values (causally) lead to evolution by natural selection. Those arguing for the causal nature of fitness


doi:10.1093/bjps/axs010

and selection generally understand fitness to be an organism's probabilistic propensity to produce offspring (Brandon [1978]; Mills and Beatty [1979]). Let's label the proponents of the causal side the 'propensity theorists' and their interpretation of fitness the 'propensity interpretation of fitness' (PIF).[1] On the other side, we have the 'statistical' camp that argues for a non-causal understanding of fitness (and natural selection and drift) (e.g. Matthen and Ariew [2002], [2009]; Walsh, Lewens, and Ariew [2002]; Ariew and Ernst [2009]; Pigliucci and Kaplan [2006]). For them, 'fitness is a mere statistical, noncausal property of trait types' (Walsh [2010], p. 148)—it quantifies but does not cause evolution.

The statisticalist arguments fall into two broad classes. The first of these classes is the argument that the causal account of fitness is predictively weak—I will therefore label it PW—and that these predictive shortcomings undermine the causal account of fitness. The PW arguments attempt to exhibit this inadequacy by either arguing that a causal interpretation of fitness and selection (PW1) is merely qualitative, not quantitative, or that (PW2) when quantified and used to make predictions, the wrong predictions are made in at least some cases. They thus form a modus tollens argument against the causal account:

> If fitness is causal, then it should be able to make X predictions.
> Fitness construed causally cannot make X predictions.
>
> ---
>
> Therefore, fitness is not causal.

This argument—in which X might stand for something like 'accurate trait dynamics'—appears to be supported by the observation that any precise formal model that attempts to quantify the PIF is limited in scope and prone to counterexamples (Beatty and Finsen [1989]; Sober [2001]). What unites the PW arguments is that they do not take a causal account of fitness and selection to be incoherent, they merely argue that causal accounts of fitness and selection are lacking in one or more ways, not being fully able to accomplish the predictive job at hand. The article that I will focus on below that makes a PW argument is Walsh *et al.* ([2002]), henceforth WLA.

The PW arguments are in contrast to another kind of argument that has recently been tendered by the statisticalists, one that suggests not a mere quantitative deficit of some sort, but a deep problem with the idea that a

[1] This is not to say that the propensity interpretation of fitness is the only way to construe fitness as causal, but it is the chief way that fitness has been understood causally. There is a debate on the causal side whether the causes should be understood as individual-level causes or population-level causes (Millstein [2006]). And there have also been several mathematical and theoretical formulations of this propensity (e.g. Brandon [1978], [1990]; Ramsey [2006]; Abrams [2009]). I do not wish to commit to any of these in particular. Finally, although I focus on organisms in what follows, I do this for the sake of simplicity and am not staking a claim in the levels of selection debate and arguing that organisms are the only fitness-bearing entities.

coherent causal picture of fitness is possible. Arguments of this kind I will label fundamental incoherency (FI) arguments, since they propose that causal interpretations of fitness, selection, and drift are fundamentally incoherent. The FI argument has been explicitly made by Walsh ([2010]) and also suggested by Matthen ([2009]). Matthen focuses on drift when he reflects that:

> [. . .] subpopulations of large populations, being smaller, are subject to stronger drift than the populations of which they are parts. Given that drift is nondirectional, this is odd. How can strong nondirectional (and therefore noncanceling) forces operating on parts of a population give rise to a weak force operating on the whole? ([2009], p. 469).

This 'force' seems odd or even incoherent both because it is purported to operate in these strange ways, and also because the answer to the question 'how strongly is drift operating?' appears to depend on subjective interests—if one is interested in subpopulations, then one will report a higher value than one would if interested in full populations. While Matthen focuses on drift, Walsh makes a similar argument based on the characteristics of fitness and natural selection and suggests that the causal interpretation of these concepts leads to paradoxes. Specifically, two or more subpopulations can have some trait, T1, being more fit than an alternate trait, T2, while the sum of these subpopulations can have T2 being fitter than T1. This reversal of the rank ordering of the fitness values would, Walsh suggests, render a causal account of fitness incoherent.

The FI and PW arguments share common premises. They both take trait fitness to be the central concept in the theory of natural selection; WLA attempts to justify this by showing that individual fitness provides poorer predictions than trait fitness, whereas Walsh simply assumes the WLA conclusion that trait fitness is central and argues that a causal construal of trait fitness leads to an incoherent set of commitments.[2] In what follows, I will attempt to undermine both the FI as well as the PW argument that it is in part based on. I focus on the PW argument in Sections 2 and 3 and then turn to the FI argument for the rest of the article.

## 2 The Importance of Trait Fitness

The PIF proponents are chiefly concerned with explicating the fitness of *individual organisms*, whereas the statistical interpretation (SI) proponents are chiefly concerned with the fitness of *traits*. One of the original architects of the PIF, Brandon ([1978]), defines the fitness of organisms in the following way: The fitness (which he labels *adaptedness*) of '[organism] *O* in

[2] Unless otherwise noted, references to 'Walsh' should henceforth be taken as references to Walsh ([2010]).

[environment] $E$ equals the expected value of its genetic contribution to the next generation' (p. 201). For PIF theorists, natural selection is based on organismic fitness differences. The PIF is an attempt at giving causal substance to organismic fitness and to the extent to which this has been accomplished, it seems that a causal account of evolution by natural selection has been accomplished. For what is evolution by natural selection (at the organismic level) but evolution in response to organismic fitness differences?[3]

The PW argument challenges the idea that evolution by natural selection has its causal basis in organismic fitness. As mentioned in the Section 1, the two main arguments to this effect are that PW1, organismic fitness (unlike trait fitness) is not quantitative and PW2, organismic fitness (unlike trait fitness), even if quantitative, is predictively impoverished. Matthen and Ariew ([2002]), who make argument PW1, put it this way: 'vernacular [i.e. organismic] fitness is merely *comparative*, not *quantitive*, and that principles such as the above [the principle that organisms have characteristics that bring about variation in their ability to survive and reproduce] afford us no way of predicting or explaining the *magnitude* of evolutionary change' (p. 56, italics in original). I am in full agreement with Matthen and Ariew that the principle they cite is not sufficient for explaining the magnitude of evolutionary change. But I don't think that such principles are all there is to organismic fitness. In fact, one need merely to peer into the biological literature to see that organismic fitness can be quantified. Perhaps the best place to go is to life history theory, where features of organismic life histories are analyzed, quantified, and their relationship to organismic fitness is determined (McGraw and Caswell [1996]).

I will not argue further against PW1 but will instead focus on the more interesting and challenging PW2 argument, which is found in WLA.[4] Their tactic is to produce a model that is supposed to act as a counterexample to the claim that organismic fitness differences cause evolution by natural selection. Their model is a bit more complicated than is necessary and I will present a simplified version here. Consider a population of 40 individuals who vary in two traits, strength and boldness. There are four different combinations of this trait and each combination is represented by 10 individuals in the population

[3] The causal–statistical divide does not, however, cleanly map onto the divide over the causal locus of fitness/selection. For example, Sober ([1984]) holds that overall fitness is not causal, but that selection—specifically *selection for*—is causal (see Lewens [2010] for a discussion of Sober's argument). And Millstein ([2006]) takes natural selection to be causal, but to operate at the population level, not the organismic level.

[4] WLA's paper has been previously challenged, though not in the way I do so below. For example, Bouchard and Rosenberg ([2004]) argue that the theory of natural selection requires pairwise fitness comparisons (which undercuts the statisticalist 'central tendency' view of fitness); Stephens ([2004], [2010]) argues that, contra WLA, the force metaphor does in fact work; and Millstein ([2006]) argues that natural selection can be both operating at the population level and causal.

**Table 1.** Fitness values and numbers of individuals for organisms in a hypothetical population

| | Fitness | F1 | F2 |
|---|---|---|---|
| **Strong & Bold** | 1.5 | 10 | 15 |
| **Strong & ∼Bold** | 0.5 | 10 | 5 |
| **∼Strong & Bold** | 0.5 | 10 | 5 |
| **∼Strong & ∼Bold** | 1.5 | 10 | 15 |

(Table 1). The fitness values for the individuals of each type are represented in the 'Fitness' column. These fitness values are not unreasonable: Being strong and bold could be a highly fit phenotype. But being strong but not bold has a lower fitness (the organism has to put the energy into being strong, but cannot reap the benefits of strength without boldness), as does boldness without strength (it being bad to be bold without the strength to back it up). Lacking both boldness and strength is an adaptive strategy since the organism is burdened by neither the expense of strength nor the risk of boldness without strength.

What lessons can one draw from the table? WLA point out that, in such a situation, organismic fitness varies but there is no change in the overall proportion of the traits in the population. In both the first generation, F1, and the second generation, F2, there are 20 strong individuals and 20 bold ones—the fitness of both traits is thus the same in such a situation. From this fact they then form the following argument:

> A1. Organismic fitness differences are not sufficient for explaining trait frequency change.
> A2. By contrast, a trait fitness difference is both necessary and sufficient for explaining trait frequency change.
>
> ---
>
> Therefore, it is trait fitness, not individual fitness, that must be used in explaining trait frequency change.

They then use this conclusion to form the following argument:

> B1. Trait fitness must be used in explaining trait frequency changes.
> B2. Trait fitness explanations 'appeal to a set of statistical properties of populations, viz. the mean (and variance) of fitness between trait types. Explanations of this sort do not avert to forces' ([2002], p. 462).
>
> ---
>
> Therefore, explaining trait frequency changes requires statistical properties of populations, not forces.

Are WLA correct that it is trait fitness that is necessary and sufficient for explaining trait dynamics, while individual fitness is merely necessary (and not

sufficient) for evolution by natural selection? I will call this into question in the next section.

## 3 Trait Fitness is not a Silver Bullet

I agree with WLA that organismic fitness alone does not determine evolution by natural selection. There are many ways that evolution can fail to occur despite organismic fitness differences. Examples of this include a lack of heritability (if organisms differ in fitness with respect to a given trait, there will not be evolution in this trait due to these fitness differences unless the trait exhibits some degree of heritability) and countervailing mutations (even if there are fitness differences among organisms due to heritable traits, if there are regular mutations in the traits, evolution can fail to occur). But such facts do not serve to undermine the argument that organismic fitness is causal. From the first explication of evolution by natural selection (Darwin [1859]), fitness differences have been singled out as one of the causal ingredients for evolution, but not the only one. The other conditions classically required include heritability and variation (Lewontin [1970]). But to these we can add further conditions like the absence of countervailing mutations.

What about the situation depicted in Table 1? The first question to answer is whether evolution has indeed *not* occurred. This depends on how we define evolution and on how we individuate traits. Evolution is often defined as having occurred when there is a change in the frequency of a trait in a population over generational time. But is this really a good definition of evolution? If one instead defined evolution as a change in the frequency of trait *combinations* (the combination of being strong and bold, being strong and not bold, etc.), then not only would the situation represented in Table 1 count as evolution by natural selection, but also it is individual fitness, not trait fitness, that is required to explain it. Individual fitness, under this definition of evolution, is both necessary and sufficient for evolution by natural selection. In contrast, trait fitness differences (which there are none in this case) are not necessary for evolution in this sense.

Furthermore, not only can individual fitness (but not trait fitness) be a necessary part of the explanation of evolution, it also can be necessary for the explanation of stasis, even under the traditional definition of evolution. Consider the case of a population of diploid sexual organisms that can possess trait $a$, trait $A$, or both traits $a$ and $A$, where these traits are alleles at a given locus. Organisms can be of three different kinds, $aa$, $Aa$, and $AA$, where $Aa$ is the fittest. This is thus an instance of heterozygote superiority. If the population is in equilibrium, the traits will be equally fit but the individual organisms will differ in fitness. In such a case, does trait or individual fitness better predict trait dynamics? In order to answer this question, we should ask what

predictions about population-level change are made by trait and individual fitness values. Brandon and Nijhout ([2006]) have explored a similar question about the evolutionary predictions derived from genic versus genotypic selection and argue that an explanation at the level of the gene will be empirically wrong—because in the case of heterozygote superiority the genic selection framework posits no selection at equilibrium, it predicts more movement from equilibrium than one would observe. If this is correct,[5] then for the same reason predictions derived from organismic fitness should diverge from those deriving from trait fitness (again, organismic fitness will predict less movement from equilibrium), and it is organismic fitness that provides the correct predictions.

One might try to address this case of heterozygote superiority by claiming that it is the pairs of alleles (the genotypes), not the single alleles, that are the traits, and that trait evolution is thus evolution in allele combinations, not the traditional change in allele frequency. This would solve the heterozygote superiority case, but at the expense of getting rid of the traditional (change in allele frequency) concept of evolution. And doing this would support my trait individuation argument against WLA earlier in this section. This shows that the necessity and sufficiency of trait fitness or individual fitness for explanations of evolution or stasis is dependent on the particularities of the explanandum and on how we individuate traits and define evolution. It is simply not true of every evolutionary explanandum that 'while variation in individual fitness is necessary for changes in relative trait frequencies it is not sufficient; variation in *trait* fitness is, however, both necessary and sufficient' (WLA [2002], p. 462).

Let's summarize what we have so far: Individual fitness cannot, by itself, form a complete causal explanation for evolution by natural selection. But this is not a problem, since no one should ever have thought that it was capable of this. From the beginning, individual fitness differences were understood as one of the causally necessary (but not sufficient) conditions for evolution by natural selection. Trait fitness is similarly dependent on other factors like the genetic constitution of the population. And trait fitness may provide the wrong predictions about evolution in at least some cases, including those of heterozygote superiority. Individual fitness, on the other hand, makes the correct predictions in some cases where trait fitness fails to do so.

Given these points, it is clear that WLA's rendering of the PW2 argument does not work. Their argument A is unsound since premises A1 and A2 are not in general true: there are some evolutionary explananda for which trait fitness forms the best explanation, and there are other explananda for which individual fitness forms the best explanation. And for similar reasons, the

[5] See Weinberger ([2011]) for a challenge to Brandon and Nijhout's argument.

conclusion of argument A is also false, rendering premise B1 false and argument B unsound.

Although I have focused on WLA's rendering of the PW2 argument, my arguments undercut any attempt at a PW2-type argument. This is true because every PW2 argument concludes that fitness cannot be causal based on the premise that fitness construed causally cannot accurately predict trait frequency changes in all cases. My argument here shows that (i) we should not expect this of organismic fitness even if fitness is construed causally and (ii) trait fitness has similar predictive shortcomings. It follows from this that arguing against the causal account of fitness by arguing that trait fitness is not causal is simply an unsound argument. A general argument against fitness being causal would need to argue that *both* trait and individual fitness are not causal.

With this conclusion in hand, let's now turn to the FI argument.

## 4 The Fundamental Incoherency Argument

The FI argument is most forcefully and explicitly made by Walsh ([2010]), and his rendering of the FI argument will thus be the focus of my critiques. Walsh describes his argument as 'Gillespie meets Simpson'. 'Gillespie' here refers to biologist John Gillespie and the element of Gillespie's ([1974]) work that Walsh employs in his argument is the suggestion that the fitness of a trait decreases with an increase in the variance in the potential reproductive outcomes of individuals with the trait, *ceteris paribus*. The idea is this: organisms of a particular kind can have a variety of potential reproductive outcomes (having 0, 1, 2, 3, etc. offspring). This distribution in the potential reproduction values is subject to a variety of statistical measures, including variance. Gillespie's point is that if we hold the mean of the distribution constant and change the variance, the trait with the higher variance will be less fit. This point is not controversial, and provides support for the rather obvious point that it is generally better (all else being equal) to adopt a conservative, non-risky reproductive strategy.

According to Gillespie, trait fitness can be represented by the equation

$$\omega_i = \mu_i - \sigma_i^2/n \quad (1)$$

where $\omega$ is the fitness of a trait, μ is the arithmetic mean of the distribution in reproductive output, $\sigma^2$ is the distribution's variance, and $n$ is the population size. The degree to which this variance affects fitness, then, is inversely proportional to population size. In fact, the rank ordering of fitness values can reverse through changes in population size. To see this, imagine a population with two types of individuals, those bearing trait 1 (T1) and those bearing trait 2 (T2). Individuals with T1 have a mean reproductive output of 0.99 and

a variance of 0.2, i.e. $\mu_1 = 0.99$ and $\sigma^2{}_1 = 0.2$, while T2 is defined by $\mu_2 = 1.01$ and $\sigma^2{}_2 = 0.4$. Which type is fitter? The answer, given by Equation (1), depends on the population size, $n$. As Walsh points out, T1 will be fitter for $n < 10$, while T2 will be fitter for $n > 10$. When $n = 10$, one of Gillespie's ([1974], p. 604) 'comical situations' arises, where the fitnesses of the two types are equivalent. Walsh labels this point the 'Gillespie Point', GP, and bases his argument on its existence.

Gillespie's insight has been used extensively by the statisticalists. But it has been mainly used to form PW arguments. And because PW arguments hold that there is merely some quantitative difference between the correct value of fitness and the one provided by the causal fitness models, this suggests that some moderate change, some correction factor, could rectify the causal account of fitness by compensating for these deviations. Brandon ([1990]) proposed just such a correction factor. He suggested that the PIF should be discounted for variance in offspring number by amending it with an additional element, one that 'denote[s] some function of the variance in offspring number for a given type, $\sigma^2$, and of the pattern of variation' (p. 20). But if Walsh is correct, no mere correction factor will be sufficient for creating a coherent causal story. He takes the *reversal* of fitness values to be so radical that the there is no correction factor that can save the causal account. Instead, the causal account is incoherent and should be abandoned. Let's now see how Walsh constructs his argument.

In forming his argument, Walsh has us reflect on parts of a population constituted by individuals with either T1 or T2, as just described. He imagines the population to be composed of various subpopulations, each with $n < 10$. We are now presented with what Walsh takes to be a paradox: within the population as a whole T2 is fitter, but within each subpopulation, T1 is fitter. What are the implications of this for the causal interpretation of fitness? Walsh spells them out for us:

> The causal interpretation of fitness enjoins us to read the probabilistic relation between fitness distribution and population change as causal. When the fitness of trait 1 exceeds the fitness of trait 2 [...], there is an ensemble-level causal process—selection—that *causes* trait 1 to grow faster than trait 2. The causal interpretation of fitness, then, is committed to saying that within each subpopulation selection (probabilistically) *causes* trait 1 to increase relative to trait 2. But the aggregate of these subpopulation causes *causes* trait 2 to increase relative to trait 1 [...] An action *C* (selection of trait 1 over trait 2) that raises the probability of some effect *E* (the preponderance of trait 1 over trait 2) in each subpopulation lowers the probability of *E* overall. ([2010], p. 165, italics in original)

Walsh then proceeds to suggest that the 'challenge for the causal view of fitness, then, is to articulate a coherent interpretation in which, within each

subpopulation, fitness distribution causes trait 1 to increase over trait 2; yet in the population overall, fitness distribution causes trait 2 to increase over trait 1' (p. 166). Walsh is not optimistic that a causal account of fitness can meet this challenge:

> The upshot is that in our Gillespie model, interpreting the conditional probabilities in the calculus of causes induces an inconsistent set of causal commitments. There is no causal interpretation of the fitness distributions that does justice to their explanatory role and is consistent with the Sure Thing Principle. Consequently, interpreting fitness distributions as causes leads to an incoherent set of causal commitments ([2010], p. 168).

The Sure Thing Principle (STP) referred to by Walsh is Pearl's ([2000]) rendering of this principle:

> An action $C$ that increases the probability of event $E$ in each subpopulation increases the probability of $E$ in the population as a whole, provided that the action does not change the distribution of the subpopulations. (p. 181)

Walsh takes the STP to be a litmus test for revealing incoherent causal commitments: Any set of causal commitments that violates the STP cannot be coherent and one or more of these commitments must be discarded or modified.

How can one challenge Walsh's argument? Below I will suggest that the effects that one observes at the distinct (population and sub-population) levels are *sui generis* and that one cannot project from one level to the other. The observation that fitness values at one level are reversed relative to fitness values at another level is neither a contradiction nor provides a reason to abandon the causal account of fitness. Northcott ([2010]) made a similar point in an argument against an earlier paper by Walsh ([2007]) when he argued that 'it does not follow that just because selection and drift are causes, their strengths in each subsample of a group need bear any simple relation to their strengths in the group as a whole. It is perfectly noncontradictory for drift, for instance, to be strong in each subgroup while simultaneously being weak in the group as a whole' (p. 461).

Before I begin my argument, I would like to acknowledge another critique of Walsh ([2010]), one by Otsuka *et al.* ([2011]). Otsuka *et al.* challenge the way that Walsh uses the STP and they argue that Walsh's use of $n$ is inappropriate. Their argument against Walsh's use of $n$ is based on the premise that the $n$ that appears in Gillespie's equation refers to (and only to) a (whole) biological population. Support for this premise is given by the fact that in population genetics, effective population size, $N_e$, is a parameter crucial to predicting trait dynamics. And this parameter is given by such things as the density and mobility of the organisms, and the structure of their environment. The

effective population size subdivision is not formed at the whim of the biologist. Furthermore, each individual belongs to one and only one population with an effective size $N_e$. If this is correct, then it is illegitimate to subdivide actual biological populations and apply Gillespie's equation.

Although I am sympathetic to the Otsuka *et al.* argument, I do not think that this is a completely satisfactory rebuttal of Walsh's argument. I concede that if we are trying to explain or predict the evolutionary dynamics of a trait in a population, the population parameter that we need to plug in is the parameter for the entire biological population, not some subpart thereof. This is true because the evolution of a trait is a change in its frequency within a population over generational time. But Walsh could reply to this by acknowledging that although evolutionary biologists generally are interested in explaining population-level evolutionary phenomena, there is no reason that they could not investigate the dynamics of population subparts—and that the investigation of the trait dynamics within the subpopulations needs to incorporate the size of these subdivisions. Furthermore, there can be a reversal of the fitness values assigned to the traits, as described by Walsh above. In what follows, I will grant Walsh his premises, allow him to make the subdivisions, to use the Gillespie equation, and to speak of the fitness of the various traits within the subpopulations and the containing population. It may be that Otsuka *et al.* are correct that this is illegitimate, but I will show that even if one allows this of Walsh, his conclusions do not follow.

## 5 Car Racing and Trait Fitness Reversals

In Section 3, we reflected on the differing roles of trait and individual fitness and concluded that neither is a silver bullet in explaining evolutionary dynamics. Now that we have this distinction between individual and trait fitness in mind, we can see Walsh's argument in a new light. The first thing to consider is whether the fitness reversals that he is concerned with are individual fitness reversals or merely trait fitness reversals. If the basis for a causal account of fitness is individual and not trait fitness, then we must consider whether the fitness reversals that form the basis of the FI argument also apply to individual fitness.

Individual fitness, as understood by the PIF, has its basis in the possible future daughter populations of that individual. The possible future daughter populations of an individual, *O*, are a function of such things as *O*'s genetic constitution, the resources (food, etc.) it has available, the dangers (from parasites, predators, etc.) in its environment, as well as the features of its local conspecifics. If *O* is sexual, then the availability (and properties) of local conspecifics of the opposite sex will be an important determinant of

*O*'s fitness. But in Walsh's population-subpopulation fitness comparison, none of the determinants of individual fitness change. The subpopulations, that is, are created in the mind of the observer—the actual population is not physically carved up into discrete populations (suppressing *O*'s access to mates, say). Given this, then, individual fitness is invariant under the changing perspectives of the FI argument.

How, one might object, can fitness be invariant at the individual level but varying at the trait level? To see how this is possible, consider the analogy of a car race in which cars have to complete many laps in order to win, like American NASCAR stock car racing. In a NASCAR race, there is a probability distribution associated with all the possible outcomes of a race. In a particular race a driver, Dale Earnhardt Jr., say, might have a 0.3 probability of winning the race, a probability of 0.2 of finishing the race in second place, and so on. Earnhardt's individual fitness (as it were), then, is a function of such factors as the psychological state of Earnhardt and the other drivers in that race, the state of his car and the other cars in the race, the weather conditions, the number of laps raced, and so on. Some of these factors will be invariant, like the number of laps raced on a given track, while others vary from moment to moment, like the psychological state of the drivers. Nonetheless, given all of these conditions at the beginning of the race, there is a unique set of probabilities associated with each individual for winning the race. Similarly, at the start of an organism's life, there will be a unique set of probabilities associated with each organism's potential reproductive outcomes.

But despite the fact that Earnhardt's (and the other drivers') fitness is fixed by all of these properties, it is nonetheless true that we can ask a variety of questions about the fitness values of various traits possessed by the cars in the race. For example, we could ask what proportion of red cars we would expect to win the first lap. And we can enquire what proportion of red cars we would expect to win the race (of 500 laps, say). Interestingly, the Gillespie reversal of the FI argument can arise here. If the average lap time of the red cars is higher, but have a higher variance, then we might expect the red cars to do worse than the non-red cars in the first lap, but we would expect them to do better in the entire race. For simplicity, let's assume that half of the cars are red that and the rest are blue. It is in fact consistent to hold that the lap times will likely be lower for more blue cars than red cars in *each lap*, but we would expect that of the first half of the cars to cross the finish line, that the majority will be red. This is an instantiation of the paradox at the center of the FI argument. But this paradox is easily resolved when we consider that even though more blue cars will have lower lap times for each lap, when a red car has a better lap time than a blue car, it generally does so by a larger margin than the blue cars do when they have a better lap time.

This example makes it clear exactly what the source of the paradox is. A tally of the number of red and blue cars who have better than average lap times for each lap is not sensitive to the *margins* by which these cars have won. Because of this, it is illegitimate to project from these single-lap statistics to the whole race. Thus, if trait fitness in the context of single laps is not based on the margins by which the cars win, but trait fitness for the complete race *is* based on these margins (since single-lap margins accumulate and affect race outcomes), then 'trait fitness' is merely being used in a polysemous way. Call a trait's single lap fitness F1, and its full race fitness F2. There is no inconsistency in holding that red's F1 is lower than blue's F1, but that red's F2 is higher than blue's F2. Furthermore, the causal ingredients that form the basis for each car's individual fitness are also able to account for both of the 'trait fitnesses', F1 and F2.

Since there is a loss of information when one calculates the fitness values of F1 (since the per lap win/loss margins are irrelevant), and since this lost information is needed for the calculation of F2, it is not surprising that an average of F1 values does not always produce F2 values. In each subpopulation, the fitness value for T1 might be lower than T2. And it would seem paradoxical that one could add up all of these subpopulations to obtain a containing population with a fitness value of T1 higher than T2. But this is merely the result of averaging over the subpopulations and not taking into account other properties, like the margins by which the red cars will tend to win when they do win.

This argument calls into question Walsh's use of *E* in his STP litmus test. Is it true that the red cars will be more successful overall, but that the blue cars will be more successful in each lap? Yes, but this is only the case for one tally of the 'success' of the cars. The single-lap margins matter for the whole race outcomes, *E*, but not the single lap, *E*. One can thus not project from an *E* at one scale (single lap) to that of another scale (the whole race) and vise versa. Similarly for organisms, the sub-population *E* and overall *E* are *sui generis* effects and should not be lumped together as a generic *E*. This, by itself, may seem sufficient for undermining Walsh's FI conclusions. But for those not yet convinced, there are further reasons to question Walsh's use of the STP and the conclusions he draws from it.

## 6 Population Subdivisions and Evolution

In characterizing the FI argument, Walsh asserts that '[i]t is legitimate for biologists to investigate the dynamics of whole populations and their subpopulations, howsoever the latter are demarcated' (p. 165). But is this true for the explanandum that Walsh focuses on (the evolution of T1 and T2)?

Consider the question of whether (and to what extent) evolution is occurring in a population. Defining evolution as a change over generational time in the representation of traits, we can ask of a population and of the subpopulations of which it is composed whether evolution is occurring. To simplify things, let's assume that the individuals in the population are asexual and true breeding, and that there is no migration. To answer the question of whether evolution has occurred in a particular population, a biologist counts the number of T1 and T2 individuals in the population at two different times and sees whether the proportion of the traits has changed. If it has, evolution has occurred. For example, if each T1 individual produced two offspring and each T2 individual produced one offspring, then we would infer that evolution has occurred, that T1 is fitter than T2, and that T1 will come to dominate the population, if it has not already done so.

But consider the possibility of a rival biologist who was observing the same population. This biologist has a large lab and is able to have her students carefully track the reproductive fate of each individual, keeping careful track of any changes in trait frequency from parent to offspring. Because, as we have assumed, each individual breeds true, the lab will have concluded that because there is no change in the frequency of traits across generations for each individual, no evolution is occurring. No evolution is occurring, that is, at the level of individuals and their offspring. Such a fine grain of analysis is not necessary to reach this conclusion, however. The conclusion that no evolution is occurring would be reached for any homogeneous subpopulation. Dividing the population into two subpopulations, one composed solely of T1 and the other of T2 individuals, would also lead to the inference that no evolution is occurring in the subpopulations.

If we combine the conclusion of this lab with the conclusion of the first lab, then it seems that we have a contradiction: the evolution of T1 over T2 is both occurring and not occurring in the population. This 'paradox' is resolved when we track evolution at each level and are careful not to make illegitimate projections from one level to another. The best way to avoid illegitimate projections is instead of labeling the evolution of T1 over T2 (in general) as $E$, we could instead track the populations and subpopulations with labels corresponding to the different levels. For example, we could label the evolutionary effect within a homogeneous subpopulation $E_H$ and the difference in the traits for the whole population $E_W$. Is it a paradox that $E_H$ can be *no evolution* while $E_W$ is a large, positive change in the proportion of T1? No, these are *sui generis* effects; $E_H$ and $E_W$ are just different effects at different scales. It simply does not make sense to make inferences about $E_H$ given $E_W$ alone (and vice versa). It would thus be a mistake for the second lab to infer that 'no evolution is occurring in the population'. The second lab has collected

the appropriate data to infer $E_H$, but they lack the data for $E_W$. Furthermore, the same causal resources are used in explaining both $E_H$ and $E_W$. One can explain $E_H$ with a subset of causal factors necessary to explain $E_W$, since one can account for $E_H$ with heritability alone, while one needs both heritability and fitness differences to account for $E_W$.

This shows that the evolutionary explanandum can change depending on the scale of observation and on how the population is divided with respect to the trait. The explanandum can also change depending on how the population is divided up with respect to other features, like sex. Consider, for example, two subpopulations, one composed of all (and only all) females in the population, and the other subpopulation being all (and only all) males in the population. Now let's assume that there is no parthenogenesis (that all offspring are the result of a male and a female) and, for simplicity, let's assume that there are discrete generations and that T1 and T2 are not sex-linked. What is the nature of the subpopulations in the second generation (the daughter subpopulations)? If the daughter population of a subpopulation consists of all (and only all) of the offspring of that subpopulation, then the daughter population in this case will consist of the entire population since, barring parthenogenesis, all of the individuals in the population are a result of both of the prior generation's subpopulations. The containing population in the following generation will thus be coextensive with each of the daughter subpopulations. Therefore, the expectations for the proportion of T1 and T2 in the subpopulations would be the same and thus cannot differ among the subpopulations, even if they differ in size.

This section has shown that the evolutionary explanandum (the preponderance of one trait over another) is scale-relative, and it is expected that the rate of evolution of one trait over the other at one level of description cannot be projected to other levels of description. Combining this conclusion with the conclusion of the previous section, it follows that we should be very cautious in moving between subpopulations and populations. Not only does the explanandum change (the magnitude of evolution changes radically with changes of scale), but the explanans changes as well (as the NASCAR example showed us, explanations at different scales require different causal resources). Given these complexities, this makes one wonder what the referent of '*E*' is in Walsh's adoption of the STP.

## 7 The STP and the Argument for the Incoherency of the Causal Account of Fitness

The STP is exemplified for Pearl by the case of the administration of a drug, *C*, and its effect on recovery, *E*. The idea is this: If a drug increases the probability of *E* in the each subpopulation then, unless it changes the distribution

of the subpopulations, it must also increase the probability of $E$ in the population as a whole. Simpson's paradox situations can occur if, for example, the population is divided into males and females and one of the sexes is much more likely to take the drug (Pearl [2000], p. 175). In such a case, the drug can decrease the probability of $E$ for males and for females but, paradoxically, can increase the probability of $E$ overall.

How does Walsh see fitness, selection, and evolution fitting this template? The cause, $C$, is selection and the effect, $E$, is the change in trait proportions. The paradox arises since a selection pressure (selection of T1 over T2 in the whole population) can increase the probability of the preponderance of T1 over T2 in the population as a whole, while increasing the probability of the preponderance of T2 over T1 in each subpopulation. Walsh points out that this is unlike the case of the drug, in which the drug has an effect on subpopulation size. No matter how we divide up the populations, Walsh claims, the reversal persists.[6] Therefore, because there is no change in the subpopulation distributions brought about by $C$, a causal construal of fitness and selection violates the STP. Fitness and selection should instead be understood (merely) statistically, not causally.

To see whether Walsh is correct, we need to examine the nature of Walsh's $E$ and $C$ to see if they fit the STP template. One major difference between the case of recovery and the case of reproductive success is that the latter, but not the former, comes in degrees: in this model, individuals either recover or they do not, but organisms can have a few or a large number of offspring. Similarly, cars can win a lap or a race by a wide or a narrow margin. Since recovery does not come in degrees, it is legitimate to use the STP criterion in the way just described. Unless $C$ causes some change in the subpopulations, then the probability of $E$ cannot increase in each subpopulation but decrease overall. And since in the model recovery does not come in degrees, there is no difficulty referring to it generically in subpopulations and the containing population.

But if recovery came in degrees, we would have to be much more careful. There is no longer such a thing as 'recovery', full stop. There is instead X% recovery, Y% percent recovery, and so on. And if this is the case, then a drug, $C$, can have a variety of effects on recovery. It could alter the variance in the degree of recovery, it could change the mean recovery, and so on. Imagine, for example, that we want to test a drug's effectiveness to see if patients tend to recover better with or without the drug. The following results occur (where average recovery $= \mu$, variance in recovery $= \sigma^2$):

Without drug: $\mu = 49\%$, $\sigma^2 = 5\%$

[6] Though we saw in the previous section that this is not the case.

With drug: $\mu = 51\%$, $\sigma^2 = 25\%$

Given these numbers and Equation 1, it follows that in samples of $<10$ individuals, the drug appears detrimental, while for samples of $>10$, the drug appears beneficial. Thus, if our experiment involved a population of eighty individuals composed of ten groups of eight individuals each, we could get the following seemingly contradictory result: The drug is bad for the patients within each group (since within each group the healthiest four will tend to be composed of a majority of controls). But overall, the drug is good (since the healthiest forty will tend to have a minority of controls). This reversal is exactly the reversal that Walsh is worried about: the *C* brings about worse recovery in small groups and better recovery in large groups. Should this lead us to think that the drug does not therefore causally affect recovery rates? No, it would be incorrect to do so, and for the same reason it would be incorrect to think that selection/fitness is not a cause of evolution.

Does this degree of recovery example violate the STP? In the case in which recovery does not come in degrees, 'probability of recovery' captures well the effect of the drug on recovery, not just one element of this effect. If we know the number of individuals in the population, we can convert back and forth between average recovery rate and number of recovered individuals. But when recovery comes in degrees, we can no longer move from average percent recovery to number of individuals X% recovered, Y% recovered, and so on. In order to do so, we must have more information, like the way in which *C* affects the variance in the percent recovered. If we take the *E* in such cases to include not one element of the effect of *C*, but instead the complete effect of *C* on the recovery mean and variance,[7] then there is no reversal of *E* when one moves between populations and subpopulations. And because *E* is no longer a scalar quantity (instead having multiple dimensions like variance and mean), there *cannot* be such a reversal. Only by thinly describing the effect can we obtain the reversal. Tracking the multidimensional effects of *C* in this drug case, the NASCAR case, or the case of natural selection produces no violation of the STP.

## 8 Conclusions

Given the above difficulties with the Walsh and WLA arguments, what, in the end, is the status of the causal account of fitness and selection? In order to answer this question, let's first tally some of the conclusions of the above discussion. We have seen that (C1) an argument against the claim that fitness is not causal cannot merely consist in an argument that trait fitness is not causal, (C2) the fitness reversals that Walsh points to are *trait fitness* reversals, not *individual fitness* reversals, and (C3) trait fitness reversals can occur in

[7] Variance and other higher moments of the distribution (like skew and kurtosis). For simplicity, I will not discuss the higher moments.

the face of invariant individual fitness values. Furthermore, (C4) because of the way in which evolution is defined, we would expect radical shifts in the values of evolutionary change by merely subdividing the population and observing the subdivisions, and we would also expect that (C5) the way in which the subdivision are made does in fact matter for the fitness reversals. Finally, we have just seen that (C6) the way we describe *E* is important for seeing whether the STP has been violated—tracking only a subset of the effect of *C* can make it appear that the STP is violated, but tracking all of the effects does not.

From (C4)–(C6), it follows that changes in the character of evolution accompanying a change in the size of the population under observation should not be troubling and should not automatically lead us to abandon the idea that fitness/selection is one of the causes of evolution. And (C1)–(C3) leads us to the conclusion that one needs to consider individual fitness, and not just trait fitness, in the analysis of the causal structure of evolution by natural selection. The PIF is a causal account of individual fitness and to undermine it one must undermine the causal status of individual fitness. Specifically, the PIF holds that organisms have dispositional properties and that these properties form a causal basis for evolution by natural selection. We saw that WLA show that individual fitness values alone do not always correctly predict evolutionary outcomes. But this is not worrisome, since the same is true of trait fitness, and since fitness is but one of the factors required for evolution by natural selection.

Walsh has pointed out the interesting case of the Gillespie Point and has drawn from its existence the conclusion that a causal account of evolution by natural selection is incoherent. I have tried to argue that we need to be careful when we make these subdivisions and draw general conclusions from them. Walsh's explanandum, a change in trait frequencies, varies depending on the observer's scale—evolution can fail to occur at one level, while occurring at others. But carefully attending to the level of analysis and not making unwarranted projections beyond it resolves this 'paradox'. With this realization comes the conclusion that events like the preponderance of T1 over T2 are not generic effects, but are situated within the scale of observation. This calls into question Walsh's use of the STP and its reference to 'the probability of event *E* in each subpopulation' and 'the probability of *E* in the population as a whole'. What is the *E* in such cases of evolution? By not lumping effects at different scales together and by instead tracking the effects at these scales without trying to project across them, the paradox at the center of the FI argument does not seem quite so paradoxical. If organismic fitness can accommodate the Gillespie reversal, if the same (causal) facts are able to account for one trait prospering within subpopulations but suffering overall, then the FI argument cuts no ice against a causal account of fitness.

While focusing on two central papers in the statisticalist literature, my criticisms are not confined to these papers. Instead, these criticisms constitute criticisms of the PW and FI arguments in general and, if the statisticalist arguments are confined to PW and FI arguments, my paper helps to undermine the entire statisticalist project. While the statisticalists have done us the service of foregrounding some of the complexities of trait fitness and the trait-individual fitness relationship, they have not given us good reason to abandon the idea that fitness/selection is causal.

## Acknowledgements

I thank Charles Pence for generously reading and commenting on multiple drafts of this article. I also thank Jon Hodge as well as the two anonymous reviewers and the editors of this journal for their helpful advice. Finally, some of the ideas for this article were given in talks for Indiana University's Department of History and Philosophy of Science as well as the Philosophy Institute at the Katholieke Universiteit Leuven. I thank the audience members for their thoughtful suggestions.

*Department of Philosophy*
*100 Malloy Hall*
*University of Notre Dame*
*Notre Dame, IN 46556*
*USA*
*grant.ramsey@nd.edu*

## References

Abrams, M. [2009]: 'The Unity of Fitness', *Philosophy of Science*, **76**, pp. 750–61.

Ariew, A. and Ernst, Z. [2009]: 'What Fitness Can't Be', *Erkenntnis*, **71**, pp. 289–301.

Beatty, J. and Finsen, S. [1989]: 'Rethinking the Propensity Interpretation of Fitness: A Peek inside Pandora's Box', in M. Ruse (*ed.*), *What the Philosophy of Biology Is: Essays for David Hull*, Dordrecht: Kluwer, pp. 17–30.

Bouchard, F. and Rosenberg, A. [2004]: 'Fitness, Probability, and the Principles of Natural Selection', *British Journal for the Philosophy of Science*, **55**, pp. 693–712.

Brandon, R. N. [1990]: *Adaptation and Environment*, Princeton, NJ: Princeton University Press.

Brandon, R. N. [1978]: 'Adaptation and Evolutionary Theory', *Studies in History and Philosophy of Science*, **9**, pp. 181–206.

Brandon, R. N. and Nijhout, H. F. [2006]: 'The Empirical Nonequivalence of Genic and Genotypic Models of Selection: A (Decisive) Refutation of Genic Selectionism and Pluralistic Genic Selectionism', *Philosophy of Science*, **73**, pp. 277–97.

Darwin, C. [1859]: *On the Origin of Species by Means of Natural Selection*, London: Murray.

Gillespie, J. H. [1974]: 'Natural Selection for Within-Generation Variance in Offspring Number', *Genetics*, **76**, pp. 601–6.

Lewens, T. [2010]: 'The Natures of Selection', *The British Journal for the Philosophy of Science*, **61**, p. 313.

Lewontin, R. C. [1970]: 'The Units of Selection', *Annual Review of Ecology and Systematics*, **1**, pp. 1–18.

Matthen, M. [2009]: 'Drift and Statistically Abstractive Explanation', *Philosophy of Science*, **76**, pp. 464–87.

Matthen, M. and Ariew, A. [2009]: 'Selection and Causation', *Philosophy of Science*, **76**, pp. 201–24.

Matthen, M. and Ariew, A. [2002]: 'Two Ways of Thinking about Fitness and Natural Selection', *The Journal of Philosophy*, **99**, pp. 55–83.

McGraw, J. B. and Caswell, H. [1996]: 'Estimation of Individual Fitness from Life-History Data', *American Naturalist*, **147**, pp. 47–64.

Mills, S. K. and Beatty, J. H. [1979]: 'The Propensity Interpretation of Fitness', *Philosophy of Science*, **46**, pp. 263–86.

Millstein, R. L. [2006]: 'Natural Selection as a Population-Level Causal Process', *The British Journal for the Philosophy of Science*, **57**, pp. 627–53.

Northcott, R. [2010]: 'Walsh on Causes and Evolution', *Philosophy of Science*, **77**, pp. 457–67.

Otsuka, J., Turner, T., Allen, C. and Lloyd, E. A. [2011]: 'Why the Causal View of Fitness Survives', *Philosophy of Science*, **78**, pp. 209–24.

Pearl, J. [2000]: *Causality: Models, Reasoning, and Inference*, Cambridge: Cambridge University Press.

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

Ramsey, G. [2006]: 'Block Fitness', *Studies in History and Philosophy of Biological and Biomedical Sciences*, **37**, pp. 484–98.

Sober, E. [2001]: 'The Two Faces of Fitness', in R. S. Singh, C. B. Krimbas, D. P. Paus and J. Beatty (*eds*), *Thinking about Evolution: Historical, Philosophical, and Political Perspectives*, Cambridge: Cambridge University Press, pp. 309–21.

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

Stephens, C. [2010]: 'Forces and Causes in Evolutionary Theory', *Philosophy of Science*, **77**, pp. 716–27.

Stephens, C. [2004]: 'Selection, Drift, and the "Forces" of Evolution', *Philosophy of Science*, **71**, pp. 550–70.

Walsh, D. M. [2010]: 'Not a Sure Thing: Fitness, Probability, and Causation', *Philosophy of Science*, **77**, pp. 147–71.

Walsh, D. M., Lewens, T. and Ariew, A. [2002]: 'The Trials of Life: Natural Selection and Random Drift', *Philosophy of Science*, **69**, pp. 452–73.

Walsh, D. M. [2007]: 'The Pomp of Superfluous Causes: The Interpretation of Evolutionary Theory', *Philosophy of Science*, **74**, pp. 281–303.

Weinberger, N. [2011]: 'Is There an Empirical Disagreement between Genic and Genotypic Selection Models? A Response to Brandon and Nijhout', *Philosophy of Science*, **78**, pp. 225–37.