Brit. J. Phil. Sci. **67** (2016), 1025–1055

Multi-Level Selection and the Explanatory Value of Mathematical Decompositions

Christopher Clarke

ABSTRACT

Do multi-level selection explanations of the evolution of social traits deepen the understanding provided by single-level explanations? Central to multi-level explanations is a mathematical theorem: the multi-level Price decomposition. I build a framework through which to understand the explanatory role of such non-empirical decompositions in scientific practice. Applying this general framework to the present case places two tasks on the agenda. The first task is to distinguish the various ways by which one might suppress within-collective variation in fitness, or indeed between-collective variation in fitness. I distinguish five such ways: increasing retaliatory capacity; homogenizing assortment; collapsing either fitness structure or character distribution to a mean value; and boosting fitness uniformly within collectives. I then evaluate the biological interest of each of these hypothetical interventions. The second task is to discover whether one of the right-hand terms of the Price decomposition measures the effect of any of these interventions. On this basis I argue that the multi-level Price decomposition has explanatory value primarily when the sharing-out of collective resources is 'subtractable'. Thus its value is more circumscribed than its champions Sober and Wilson ([1998]) suppose.

doi:10.1093/bjps/axv008

Advance Access published on March 29, 2015

1 Single-Level and Multi-Level Selection

One of the key variables in evolutionary theory is character–fitness covariance: the degree to which those organisms that possess a given character are statistically more likely to be fitter than those organisms that don't possess the character. Take for example a lion's inclination to hunt socially rather than on its own. Suppose that the fitness of each lion in a population is given by Table 1. So it's determined by whether or not that lion has this inclination to hunt socially, and by whether or not the lions that it interacts with have this inclination. Making some simple assumptions, one can calculate that the covariance between character and fitness in this case is $f_0(1-f_0)(4f_0-1)$, where f_0 is the proportion of the lion population who are presently social hunters.[1] Consider the case in which the population is evenly divided at present between social hunters and lone hunters; in other words, $f_0 = \frac{1}{2}$. In these circumstances it follows that there is a positive covariance between social hunting and fitness of $\frac{1}{4}$. This fact about covariance is key because it can provide a simple explanation of why the frequency of social hunters increased from the present generation of lions to the next generation: Lions inclined to hunt socially were—in the circumstances above—more likely to be fitter and this caused such lions to have relatively more offspring, most of whom inherited this inclination. And so the frequency of social hunters increased.

For reasons that will soon become clear, I will call such explanations 'single-level selection' explanations. Such explanations are underwritten by the Robertson–Price identity. This equation describes how the covariance of character and fitness determines the increased prevalence of a character in a population (Robertson [1966]; Price [1970]). This equation follows deductively from some common simplifying assumptions: that there is no migration into or out of the population; that the character in question is heritable and inherited without transmission bias; and that there are no stochastic effects at work (Price [1972]; Sober [1984]; Okasha [2006]). In the wake of Darwin's *On the Origin of Species*, single-level selection explanations have become so commonplace in evolutionary biology as to be unremarkable:

> It would be advantageous to the Melipona [bee], if she were to make her cells closer together, and more regular in every way than at present; for then, as we have seen, the spherical surfaces would wholly disappear, and would all be replaced by plane surfaces; and the Melipona would make a comb as perfect as that of the hive-bee [. . .] Thus, as I believe, the most wonderful of all known instincts, that of the hive-bee, can be explained by natural selection [. . .] (Darwin [2008], pp. 174–75)

[1] Assume that lions form pairs completely at random.

Table 1. Example fitness matrix

	Who interacts with social hunters	Who interacts with lone hunters
Fitness of social hunter	4	0
Fitness of lone hunter	1	1

Moving from explanations of concrete biological cases over to abstract mathematical models, this 'single-level' emphasis upon character–fitness covariance remains commonplace.[2] For example, in textbook treatments of evolutionary theory one sees fitness matrices (such as Table 1) being used to identify the circumstances under which this character–fitness covariance will be positive, negative, or zero (McElreath and Boyd [2007], p. 203). In the lion hunting case, for example, this depends upon the initial frequency, f_0, of social hunters in the population. Indeed one could describe the search in evolutionary game theory for so-called evolutionary stable states or strategies roughly as the search for the conditions under which character–fitness covariance is zero; $f_0 = 0$, or $\frac{1}{4}$, or 1 in this example.[3]

This illustrates how the covariance of character with fitness across the whole population is a central explanatory variable. Now, in its multi-level form,[4] the so-called Price equation decomposes this central variable into the sum of two other variables (Okasha [2006]). To put it briefly, one of these variables is supposed to relate in some sense to selection at the level of individual lions and the other to selection at the level of groups of lions. I will say much more about these two variables in Section 3. For now it will suffice to say that both these variables are statistical functions of the distribution of character and fitness among lion groups.

Consider, for example, those cases in which selection for social hunting at the level of lion groups outweighed selection against social hunting at the level of individual lions. (Again, much more on this in Section 3.) In such cases the multi-level Price decomposition suggests a controversial explanation for the increase in the prevalence of social hunters from one generation to the next: group-level selection for social hunting outweighed individual-level selection against social hunting. As a consequence, explanations that employ these two

[2] McElreath and Boyd ([2007], Section 5.1) call the use of single-level explanations the 'personal fitness approach' to evolution.

[3] Note that this is a necessary but not sufficient condition for a distribution of characters across a population to constitute an evolutionarily stable distribution.

[4] The multi-level Price equation is a variation of the Price equation (Price [1972]), which itself is a more general form of the Robertson–Price identity (Robertson [1966]; Price [1970]).

variables from the multi-level Price decomposition are often called 'multi-level selection' explanations.

The main focus of this article will be the contrast between multi-level selection explanations and single-level selection explanations. This will leave no time to say anything about the explanations afforded by selfish-gene theory (Dawkins [1976], [1982]) or inclusive fitness theory (Hamilton [1964]; Frank [1998]). Moreover, considerations of space prevent me from discussing the alternative form of multi-level selection theory based on contextual analysis (Heisler and Damuth [1987]; Goodnight *et al*. [1992]) rather than the multi-level Price decomposition.

In contrasting the multi-level explanatory framework with the single-level framework I do not mean to imply that these frameworks offer competing explanations. (As I define the concept, two explanations of the same case compete exactly when it is highly implausible, if not impossible, that they both be correct; take for instance the explanation that the CIA shot Kennedy and the explanation that Soviet agents shot Kennedy.) In fact I'm happy to accept the so-called pluralist idea that multi-level explanations and single-level explanations—and for that matter selfish-gene and inclusive-fitness explanations—often posit the same process (Kerr and Godfrey-Smith [2002]) and so each framework can plausibly provide a correct explanation of the same case.

Instead, by contrasting multi-level explanations with single-level explanations, what I aim to do is address the issue of explanatory depth. For example, an explanation of why a car accelerated that specifies the car's mechanics or the psychology of its driver provides a deeper explanation than merely citing the fact that the accelerator pedal was pressed. This shows how one explanation can be deeper than another without competing with it. On the one hand, Sober and Wilson ([1998]) think that explanations of the evolution of social characters that employ the multi-level Price decomposition are deeper than single-level explanations. But, on the other hand, there are those who disagree: Maynard-Smith disagrees because he finds multi-level explanations altogether dubious[5]; whereas Dugatkin and Reeve ([1994], p. 121, 124) disagree because they think multi-level explanations are fully equivalent to single-level explanations.[6]

The distinctive strategy of this article will be to separate this issue of explanatory depth from the other issues in the levels of selection literature with which it is entangled. In addressing it, I will draw instead upon the general

[5] See (Okasha [2005], pp. 1000, 1004) for references and discussion of the complexities of Maynard Smith's views.

[6] Things are not quite as clear cut as this; see (Dugatkin and Reeve [1994], p. 123). What is clear is that much confusion has been generated in contrasting multi-level selection explanations with their 'individualist' rivals, without making clear what rivals one has in mind.

philosophical literature on explanation. Thus I will not discuss what it means for selection to 'act at a particular level' such as that of the group (Lloyd [1986], [2000]; Okasha [2006]), nor what it takes for something such as a group of organisms to count as a 'biological individual' (Clarke [forthcoming]), nor whether groups can be vehicles in Dawkin's ([1982]) sense or interactors in Hull's ([1981]) sense. Indeed, one could perhaps think that there is no fact of the matter about such questions,[7] questions concerning vehicles or interactors say, but still think that there is a fact of the matter about the topic of this article, namely, the depth of the multi-level selection framework.

This focus on the explanatory depth of the multi-level Price decomposition will also raise wider philosophical questions. For the decomposition is a mathematical theorem: its truth is not contingent on what the world happens to be like; and one doesn't need any scientific evidence to know that it is true. Consequently, one might wonder how such non-empirical propositions could play a genuine role in scientific explanation (Pincock [2007]; Baker [2009]; Batterman [2010]). As Lange and Rosenberg ([2011], p. 593) point out in response to Sober ([2011]), it is 'difficult to see how [propositions in evolutionary theory that are knowable *a priori*] could figure in causal explanations'.

So I will look beyond the philosophy of biology literature to explore how non-empirical decompositions such as the multi-level Price theorem can play an explanatory role. The suggestion will be—to put it somewhat laconically—that such decompositions highlight those constitutive relationships that help glue different factors in our explanatory reasoning together. Applying this suggestion to the multi-level Price decomposition shows that this decomposition has explanatory value, I will argue, primarily in cases in which the sharing-out of resources is 'subtractable'. Thus the range of cases across which the decomposition provides deep explanations is more circumscribed than its champions suppose.

2 Three Conditions on Explanatory Decompositions

What does one need to know in order to explain a phenomenon? In the philosophical literature, a very popular suggestion is that one needs to know what would happen under certain 'hypothetical alterations' to the system in question. Would the phenomenon still have occurred if certain things had gone differently (Lewis [1986]; Woodward [2003])?[8] To explain why the economy shrank in 2008, for example, it helps to know that the size of the economy

[7] See (Sterelny [1996]; Okasha [2004a]; Sarkar [2008]) for discussion of this sort of pluralism.

[8] Lewis ([1986]) doesn't put it in quite these terms. He says that to explain is to cite a cause; but for Lewis to cite a cause is just to say what could have gone differently such that the phenomenon wouldn't have occurred.

would have been greater if banks had been more tightly regulated. So I am going to follow Lewis and Woodward in assuming that to explain is to answer important what-if-things-had-been-different questions. Accordingly, the depth of an explanation is in proportion, roughly speaking, to the number of what-if questions it allows one to answer concerning important hypothetical alterations to the system in question. This measure of explanatory depth is by no means uncontroversial, but I will wait until Section 8 to examine it in further detail.

One qualification: to explain why the economy shrank it does not help to know that the size of the economy would have been greater if extraterrestrials had landed from outer space and donated a billion barrels of oil to the treasury. The what-if question about bank regulation is therefore different to the question about extraterrestrial oil donation in that answering the former has explanatory value, but answering the latter does not. I will assume that the standard account of such differences is correct: we just happen to be more interested in hypothetical alterations to bank regulation than in far-fetched questions about extraterrestrial oil donations.[9] The importance of a hypothetical alteration depends in this respect upon our personal interests; and thereby so does explanatory depth according to my measure.[10] (Accordingly, the notion of what is interesting to biologists will play a central role later in this article.)

I will now use the Lewis–Woodward approach to explanation in order to build a toy model of how a non-empirical decomposition can play a modest role in explanation. Consider the following decomposition: The number of guests booked into a hotel is equal to the number of guests who are on holiday to ski plus the number of guests who are not on holiday to ski. This decomposition is non-empirical, guaranteed by the logical truth that everyone is either a skier or a non-skier. Compare this decomposition, for example, to a second decomposition of the guests into those with blond hair and whose name begins with 'K', on the one hand, and those who do not have both attributes, on the other hand. The question I want to ask is this: When will a non-empirical decomposition (for example, the first decomposition) be more explanatorily valuable than any of the infinitely many other non-empirical decompositions that one might think of (such as the second decomposition)? To explore this question, let us consider how the first decomposition fits into the following story.

(i) This winter has been unusually warm and so the average depth of snow on the Brixental ski slopes has been half a meter, in contrast to last winter's three metres. As a result (a) there are 200 skiers booked into the Brixental

[9] But see Hart and Honore ([1965]) for an alternative account.

[10] For an account that emphasizes interests but not what-if questions see van Fraassen ([1977]) and Achinstein ([1983]).

hotel, in contrast to last winter's 900; and (b) like last year, there were 100 non-skiers also booked into the Brixental hotel. Most of these non-skiers were there for the annual Wittgenstein conference. So applying our decomposition to (a) and (b), we see can see that (ii) the hotel has had under 500 guests rather than over 500 as they did last winter. As a result the hotel has gone bankrupt.

Note that the low number of guests on its own provides a simple explanation of the bankruptcy. And this explanation is made deeper by adding the point about the lack of snow. But to have a really satisfying explanation of the bankruptcy, one needs also to be able to answer what-if questions of the following form: (*Z*) what if x meters of snow had fallen, and other factors like the Wittgenstein conference had been arranged in such-and-such a way?[11]

To answer such what-if questions, one will typically reason as follows: 'In this hypothetical what-if scenario, there would be g_a skiing guests on account of the snow; and there would be g_b non-skiing guests on account of the other factors such as the Wittgenstein conference. According to our decomposition this constitutes there being g guests in total. There would therefore be under 500 guests, and so the hotel would be bankrupt. Alternatively: there would be over 500 guests, and so the hotel would not be bankrupt.

Let's be fully explicit about how this works. To know our decomposition is to know a constitutive relationship, *X*: g is constituted by g_a and g_b. And knowing this constitutive decomposition *X* is in practice how we come to know the following causal determination relationships, *Y*: An interesting first factor (snowfall) combines with other factors (such as the Wittgenstein conference) to determine a second factor (total guests), which in turn determines the to-be-explained phenomenon (bankruptcy). And knowing these causal determination relationships, *Y*, in turn allows us to answer some important what-if questions, *Z*. Thus this knowledge deepens our simple undecomposed explanation (of the bankruptcy in terms of the total number of guests alone). In short, our decomposition highlights a constitutive relationship that helps us to glue together the relevant factors in our explanatory reasoning.

In principle, of course, one could know these causal determination relationships, *Y*, without knowing the constitutive decomposition, *X*. So the explanatory role of our decomposition is what one might call an 'ancillary' one. It is dispensable in principle, but not in practice.

It will be important for later to abstract three crucial aspects from this toy example concerning the guests at the Brixental hotel:

> Independence aspect: The value of a term on the right hand of the decomposition (g_b non-skiing guests) is independent of the first factor

[11] Arranged, for example, such that there were g_b non-skiing guests.

(snowfall). In other words its value is preserved by some hypothetical alteration to that first factor (eliminating snowfall). Observe that this aspect of the Brixental case is crucial in that, without it, knowledge of the constitutive relationship X would be of no real help in calculating the causal dependencies Y. Later on, I will repeatedly draw upon the observation that this independence aspect of the Brixental case is equivalent to the following condition.

Independence aspect (alternate rendering): The effect upon the value of the left-hand term (g total guests) of this alteration (eliminating snowfall) is measured by the attendant change to the value of a right-hand term in the decomposition (g_a skiing guests). After all, the other right-hand term (g_b) is a residual term that measures the effect of other factors only (such as the Wittgenstein conference).

Interestingness aspect: This hypothetical alteration (eliminating snowfall) is interesting. This aspect of the Brixental case is crucial in that, without it, the what-if question Z would not be an important one. Hence answering this question would be of no explanatory value according to the Lewis–Woodward thesis about explanation; just as in my extraterrestrial oil donation example.

Knowledge aspect: One knows how the value of the left-hand term (g total guests) determines the to-be-explained phenomenon (bankruptcy) in the circumstances. This aspect of the Brixental case is crucial in that, without it, one could not use causal decomposition X to answer what-if question Z.

My conclusion is this: The Lewis–Woodward approach to explanation issues in three criteria that are in general individually necessary and jointly sufficient for a non-empirical decomposition to provide explanatory value in the above manner. That is to say, to issue in an explanation of greater depth than an explanation (of the bankruptcy) in terms of only the left-hand term of a decomposition (the total number of guests).

I note in passing that the decomposition involving guests with blonde hair and names beginning with 'K' would in normal circumstances fail both the independence criterion and the interestingness criterion. Some non-empirical decompositions are evidently more explanatorily valuable than others.

I emphasize that the above are criteria only for the explanatory value of non-empirical decompositions, not empirical ones. To extend them to the case of empirical decompositions would be mistaken. For example the ideal gas law $\ln(P) = \ln(V) + \ln(T)$ has clear explanatory value. But it fails my independence criterion: When a gas is heated in an expandable chamber both the value of the $\ln(V)$ term and of the $\ln(T)$ term are altered as a result. So my first

criterion is not necessary as regards the explanatory depth of empirical decompositions, as opposed to non-empirical ones. Conversely, the length of Edward Heath's premiership is equal to the length of Romano Prodi's premiership plus the length of John F. Kennedy's. This equation may well meet all my criteria, but it is too accidental to have any explanatory value.[12] So my three criteria are also not jointly sufficient as regards the explanatory value of empirical decompositions, as opposed to non-empirical ones.

At any rate, the explanatory role played by the toy decomposition involving hotel guests, I will suggest, is the same explanatory role that many non-empirical decompositions play in the actual practice of science; and in particular, in the multi-level Price decomposition in evolutionary biology.

3 The Multi-Level Price Decomposition

To spell out the multi-level Price decomposition, let me introduce some standard formalism. Consider a population of individuals, be it a population of genes, cells, organisms, or social groups; although the most intuitive case is when one takes individuals to be individual organisms. Take an arbitrary individual, i. Let ω_i denote that individual's (relative) fitness.[13] Let z_i denote the degree to which individual i possesses a particular character in which one is interested. This character of interest will conventionally be a 'pro-social' character such as a lion's being inclined to hunt cooperatively or a vampire bat's being inclined to donate blood to other vampire bats who are in need. The multi-level Price decomposition states that[14]:

$$\mathrm{Cov}(\omega, z) = \mathrm{Cov}[\mathrm{Exp}_g(\omega), \mathrm{Exp}_g(z)] + \mathrm{Exp}[\mathrm{Cov}_g(\omega, z)] \quad (1)$$

What do these three terms mean? The left-hand term $\mathrm{Cov}(\omega, z)$ denotes the covariance of character with fitness across the whole population: To what extent do individuals who score high on character z tend statistically to be fitter than individuals in the population who score low on z? For example are group hunters fitter on average than other lions?

Now imagine that our population of individuals is partitioned into collectives; so each individual is a member of exactly one collective. (I will leave it entirely open what it is for an individual to be a member of a collective.) So $\mathrm{Cov}_g(\omega, z)$ denotes the covariance of character with fitness within collective g, rather than across the whole population: To what extent do individuals in

[12] It certainly isn't invariant under interventions (Woodward [2003]). In contrast, note that non-empirical decompositions are by definition maximally invariant under interventions.

[13] Relative fitness is defined to be an individual's absolute fitness divided by the mean fitness of all individuals in the population. I shall henceforth use 'fitness' to mean relative fitness.

[14] See (Price [1972]; Hamilton [1975]) for a seminal formulation. See (Okasha [2006]) for a very clear commentary.

collective g who score high on character z tend statistically to be fitter than those in the same collective who score low on character z? Thus the third term of the decomposition, $\mathrm{Exp}[\mathrm{Cov}_g(\omega, z)]$, is an average of this measure across the whole population: On average, do group hunters tend statistically to be fitter than those in the same collective who hunt alone?

Finally, the second term: $\mathrm{Exp}_g(\omega)$ is the average fitness of the members of collective g. Let's call this the collective's fitness. Similarly $\mathrm{Exp}_g(z)$ is the average character of the members of collective g. Let's call this the collective's character.[15] So the second term of the multi-level Price decomposition, $\mathrm{Cov}[\mathrm{Exp}_g(\omega), \mathrm{Exp}_g(z)]$, is the covariance between these two variables: to what extent do collectives that score high on character z tend statistically to be fitter than collectives that score low on character z?[16] Putting this less technically and more intuitively: the second term of the decomposition measures the association between collectives of (collective) fitness with (collective) character, whereas the third term measures the association of (individual) fitness with (individual) character within collectives. Importantly, the multi-level Price decomposition is a mathematical theorem, guaranteed by the logic of covariance and of expectation.

It is worth noting at this point that my third criterion for a mathematical decomposition to have explanatory value—the knowledge criterion—just requires that we know how the value of the left-hand term determines our to-be-explained phenomenon in the circumstances. And one does in this case. For one knows the Robertson–Price identity discussed in Section 1, which formally underwrites the intuition that the fitter character z is, so to speak, the more it will increase in frequency. So one knows how the value of the left-hand term (the degree of character–fitness covariance in the whole population) determines our to-be-explained phenomenon, the evolution of character z. The knowledge criterion is satisfied. Consequently, this article will focus on the circumstances under which the multi-level Price decomposition satisfies the independence and the interestingness criteria.

4 The Biological Interest Problem for Sober and Wilson

One suggested explanatory role for the multi-level Price decomposition emphasizes the factor of within-collective variation (Sober and Wilson [1998]). And by this I strongly suspect that Sober and Wilson mean variation

[15] Thus I am focusing on what Damuth and Heisler ([1988]) call multi-level selection type one, rather than type two.

[16] Strictly speaking, the summation Cov[] is over individuals in the population not collectives. So strictly speaking: to what extent do individuals that are part of collectives that score high on character z tend to be members of fit collectives?

in fitness rather than variation in character.[17] Sober and Wilson's key claim is that the third term of the decomposition measures the effect of within-collective variation ([1998], pp. 32–3, 73–5). (Sober and Wilson also claim that the second term of the multi-level Price decomposition measures the effect of between-collective variation. I will set the examination of this claim aside until Section 7.)

The general framework developed in Section 2 shows why Sober and Wilson's key claim bears upon the explanatory value of the multi-level Price decomposition. For this key claim is more or less an application of my independence criterion for explanatory value. Imagine eliminating within-collective variation in fitness. Let ε denote the attendant effect upon character–fitness covariance across the whole population—that is, the effect on the value of the left-hand term of the multi-level Price decomposition. Independence criterion (the alternate rendering): This effect, ε, is measured by the attendant change in the value of a right-hand term in the decomposition, for example, the third term. So Sober and Wilson's key claim is more or less an application of the first of my three criteria for the multi-level Price decomposition to have explanatory value.

Unfortunately, Sober and Wilson do not provide an argument for this key claim. What follows is the most plausible way of developing such an argument in my view.

Take a population of individuals in an environment and consider the 'fitness structure' generated by that environment. This fitness structure is the mapping that specifies how an individual's fitness is determined by her character and by the characters of the individuals with whom she interacts. Take for illustration the function $\omega_i = 2\mathrm{Exp}_g(z) - \frac{1}{2}z_i$. Now consider a hypothetical alteration to this fitness structure such that each individual in any given collective, g, will now enjoy the same fitness as the other individuals in collective g. More precisely, the fitness an individual is to enjoy under this alteration is identical to the mean fitness—prior to this alteration—of the individuals in her collective. Sticking with the above illustration, ω_i becomes equal to $2\mathrm{Exp}_g(z) - \frac{1}{2}\mathrm{Exp}_g(z)$. In other words, it's equal to $\frac{3}{2}\mathrm{Exp}_g(z)$. Call such alterations 'structural collapse to the mean' (SCM) alterations. This alteration is one straightforward way of eliminating any within-collective variation in individual fitness.

Note, however, that the SCM alteration preserves the mean fitness of the members of each collective, and thus preserves collective fitness. But

[17] See (Sober and Wilson [1998], pp. 54, 66–7, 80–91, 115, 139) for textual evidence; indeed, see (Sober [1984]). At any rate my criticism of Sober and Wilson's idea as reconstructed in Sections 6 and 7 will work just as well if you substitute 'fitness' for 'character' and 'character' for 'fitness'. This is because covariance is symmetric: $\mathrm{Cov}(\omega, z) = \mathrm{Cov}(z, \omega)$. So the mathematical reasoning in my criticism will hold even if Sober and Wilson mean 'variation in character' rather than 'variation in fitness'.

individual character is also preserved; so collective character is preserved. Thus the SCM alteration preserves the covariance of collective fitness with collective character. In other words, SCM alterations preserves the value of the second term of the multi-level Price decomposition. And this is equivalent to saying that the independence criterion for explanatory value, on its original rendering, is satisfied here.[18] Incidentally, let ε denote the effect of SCM alterations upon character–fitness covariance across the whole population—that is, its effect on the value of the left-hand term in the decomposition. SCM alterations having preserved the value of the second term, it follows that this effect, ε, is measured by the attendant change in the value of the third term in the decomposition. And this is equivalent to saying that the independence criterion for explanatory value, on the alternate rendering, is satisfied here.

Having established that my first criterion for explanatory value is satisfied with respect to hypothetical SCM alterations, can we now establish my second criterion, the interestingness criterion? Is the SCM alterations elimination of within-collective variation in fitness especially interesting to biologists? I will now argue that are some cases in which the answer is no.

Recall the example in which $\omega_i = 2\mathrm{Exp}_g(z) - \frac{1}{2}z_i$, which we can rewrite as $2\mathrm{Exp}_g(z) - z_i - \frac{1}{2}(-z_i)$. Let's imagine that this describes the fitness structure for the *Polistes fuscatus* wasp in a given environment. Wasps with high z scores are hard workers. And wasps enjoy fitness benefits when they are in a collective whose members are hard working; hence the $2\mathrm{Exp}_g(z)$ term. But working hard requires a costly expenditure of energy; hence the $-z_i$ term. But those lazy wasps who do not work hard run the risk of being stung by the queen, and indeed the risk of other forms of retaliation from the queen (Gamboa *et al.* [1990]); hence the $-\frac{1}{2}(-z_i)$ term.

In the case of the *Polistes* wasp there is indeed a highly interesting way of altering the fitness structure that eliminates within-collective variation in fitness. One imagines an increase in retaliatory capacity: queens are better able to identify the lazy workers, or the queens increase the severity of the punishment for those who are so identified. In particular, it will be interesting to know what would happen were the $\frac{1}{2}$ coefficient—the retaliation parameter, so to speak—to be altered such that each individual in a collective enjoys the same fitness, within-collective variation thus being eliminated. One can calculate that the answer is that the coefficient becomes 1 and that ω_i becomes $2\mathrm{Exp}_g(z)$.

It is crucial to note, however, that this highly interesting hypothetical alteration to fitness structure is distinct from the SCM alteration I considered above. After all, recall that the SCM alteration has it instead that ω_i becomes

[18] Moreover, one can easily show that SCM alterations change the value of the third term to zero. So the magnitude of this attendant change in the third term is given by the unaltered third term itself.

equal to $\frac{3}{2}\text{Exp}_g(z)$, not to $2\text{Exp}_g(z)$. In contrast, there is nothing of especial biological interest, I contend, in the SCM alteration applied to our wasp population. Such alterations have no greater interest than hypothetical alterations that eliminate within-collective variation by letting ω_i become $\frac{7}{13}\text{Exp}_g(z)$, or to $\ln\text{Exp}_g(z)$, or that collapse individual fitness to the collective median or the collective mode, and so on.

This illustrates how the SCM alteration is not biologically interesting across every case in general. In other words SCM alteration does not in general satisfy my second criterion for explanatory value. But I've been considering hypothetical SCM alterations in an attempt to develop Sober and Wilson's analysis into an argument that establishes a general explanatory role for the multi-level Price equation. And one can now see that this attempt has failed.

I emphasize that my intention here is not to criticize the application of the multi-level Price theorem to the *Polistes* wasp case. After all, the theorem is just a mathematical truth. Rather, I am urging a more sanguine assessment of its explanatory value in this case. After all, nothing that I've said so far establishes that the decomposition adds any explanatory depth.

There will, of course, be some theorists who will resist my conclusion here by objecting to my relatively narrow conception of what is biologically interesting. I cannot hope to fully persuade such objectors. But I do hope to persuade them of a somewhat more modest point: the SCM alteration in the wasp case is just as interesting as the infinity of other hypothetical alterations to the distribution of fitnesses—such as those that let ω_i become $\frac{7}{13}\text{Exp}_g(z)$, or $\ln\text{Exp}_g(z)$, and so on. It follows that, in the case of the *Polistes* wasp, we have not established that the explanatory value of the multi-level Price decomposition will be any greater than the infinity of other mathematical decompositions of character–fitness covariance. We've not identified any special explanatory value for the decomposition in the case of the *Polistes* wasp.

5 Explanatory Depth Whenever Resources are Subtractable

One question naturally arises from the last section: Can one appeal to SCM alterations in order to establish the explanatory value of the multi-level Price equation in a more limited class of cases, rather than across all cases in general? This section will identify a class of cases in which SCM alterations are biologically interesting. In other words, I identify a class of cases that satisfy my second criterion (interestingness) for explanatory value. These cases are, namely, those cases in which the sharing-out of resources amongst the individuals in a collective is, in the parlance of economics, subtractable. But I've already shown in Section 3 that my third criterion (knowledge) is satisfied by the multi-level Price decomposition. And I've just shown in Section 4 that my first criterion (independence) is satisfied with respect to hypothetical SCM

alterations. So all my three conditions are satisfied here. Thus this section establishes the explanatory value for the multi-level Price decomposition in a limited class of cases, namely, those in which the sharing-out of resources is subtractable.

Before getting down to business, I will need to invest a substantial amount of time carefully illustrating what I mean by subtractability. An excellent illustration of the subtractability of resources in a biological context is found in the literature on social or cooperative foraging (Giraldeau and Caraco [2000]). To see this, note that many social foraging models can be thought of as having two parts. Consider the amount of food that a collective of foragers will gather. The resource acquisition part of the model describes how this amount depends upon the cooperative behaviour of the members of the collective and upon the environment. The resource sharing-out part of the model describes how this amount is divided amongst the individual members of the collective. Now, to talk of resources being genuinely 'shared out' here presupposes the following: there is an 'analytic separation' of the allocation of resources into a mechanism whereby a collective acquires its resources, and a mechanism whereby these resources are shared out amongst the individual members of the collective. So by this stipulation, resource sharing-out is subtractable only if these mechanisms are analytically separable. This is the first of my two individually necessary and jointly sufficient conditions for subtractability.

Let me be clear about analytic separation. I don't intend my definition of analytic separation to turn upon any substantial notion of 'mechanism'. Similarly, I allow that two analytically separable mechanisms may operate simultaneously, that they may interact, and that they may have overlapping parts. Instead, what I mean by 'analytic separation' is that there is a biologically interesting alteration to the manner in which resources are divided out amongst individuals, an alteration that leaves unaltered the manner in which resources are collectively acquired. To make this intuitive, consider for example those 'scroungers' who have 'cheated' by refusing to cooperate during foraging. In many cases it is biologically interesting to ask what would occur if it became more difficult for scroungers to gain access to the food that the collective has foraged. What if, in the extreme, scroungers were excluded from these resources altogether?

My second condition on subtractability is also rather intuitive. Rough and informal version: Whenever one individual consumes a resource, it must reduce the quantity of the resource available for other users to consume. To spell out the second condition formally, I will make the simplifying assumption that one can use a single variable R_g to quantify the resources that a collective, g, has acquired. In a simple foraging case this is just the quantity of food that the collective has foraged. Furthermore, I will assume that R_g is

entirely determined by the 'pro-social' character of each member of collective g, characters which one might represent by the vector $\mathbf{z}_g$. (In a simple foraging case, this pro-social character might measure how much energy the individual in question chooses to invest in the group hunt.) To emphasize this point, I will often write collective resources R_g as $R_g(\mathbf{z}_g)$ highlighting that it is a function of $\mathbf{z}_g$, and indeed of $\mathbf{z}_g$ alone. Now consider the sum total of the fitnesses of the members of a collective, g; in formal terms $\sum_g \omega_i$. The sharing-out of collective resources is subtractable I stipulate only if this total fitness is entirely determined by collective resources $R_g(\mathbf{z}_g)$; more specifically, just in case this total fitness is an increasing function of collective resources. Choose the right scale on which to measure resources and this becomes the requirement that the fitness structure is characterized by:

$$\sum_g \omega_i = R_g(\mathbf{z}_g). \tag{2}$$

Why is this requirement a fitting formalization of the rough and informal condition on subtractability that I gave above? Notice that were any individual to be fitter than they actually are—but collective resources to remain as they actually are—then Equation (2) requires that some other individual or individuals would be less fit than they actually are, and by an equal amount. In the foraging case, holding fixed the amount of food collectively foraged, one individual's gain in fitness is precisely counter-balanced by another's loss.

It is of crucial importance to emphasize that the present requirement—concerning what would happen were collective resources to remain as they actually are—obviously does not entail that collective resources must remain as they actually are. Therefore there will be many subtractable fitness structures for which collective resources vary according to the distribution of individual characters within the collective. In the foraging case, for example, the amount of food foraged $R_g(\mathbf{z}_g)$ can vary depending on how the individuals are inclined to cooperate during the hunt, as measured by $\mathbf{z}_g$. So I emphasize that subtractability of resources does not entail that individuals are playing a zero-sum game that precludes them from cooperating to increase collective resources. A similar point: subtractability does not entail that the fitness structure in play is additive. In other words, it does not entail that the fitness structure be given by $\omega_i = \lambda z_i + \mu \mathrm{Exp}_g(z)$.

In summary, I stipulate that the sharing-out of resources is subtractable just in case (i) one can analytically separate resource allocation into a mechanism of resource acquisition and into a mechanism of resource division, and (ii) Equation (2) characterizes the fitness structure in play.

A second illustration of the subtractability of resources comes from simple diploid genetics models. An AB genotype causes the organism in which it is instantiated to exemplify a corresponding phenotype, and this organism

interacts with the environment and has a number of offspring. And these offspring, by extension, are counted as the offspring of the AB genotype itself. Call this process the acquisition of the AB genotype's reproductive resources. (I'm happy to be fairly liberal about what counts as a resource.) Consider next that during meiosis the A-allele in the AB genotype will be copied to a certain number of gametes and so will enjoy a particular chance of being represented in each of the aforementioned organism's offspring. The same goes for the B-allele. Call this the sharing-out of the AB genotype's reproductive resources amongst its two alleles, A and B. Again, one can analytically separate resource allocation into collective resource acquisition and the sharing-out of these resources between individuals. For it is biologically interesting to ask what would occur if meiosis were to unfold differently: what if segregation distortion (Lyttle [1991]) occurred and the A-allele in the AB genotype enjoyed more than its fifty percent share of reproductive resources (Maynard Smith and Szathmáry [1995], Section 10)? So my first condition for subtractability is satisfied here. Equally, my second condition for subtractability is also satisfied here: holding the AB genotype's resources fixed, an increased chance of the A-allele of being represented amongst the organism's offspring would be precisely counter-balanced by a decreased chance for the B-allele.

Finally, an example in which resources are, in contrast, not shared-out subtractably is that of the *Polistes* wasp. In this case a worker's fitness is sensitive to whether he is stung by the queen. In virtue of this, avoiding being stung by the queen is a key resource. But it would be absurd to attempt to analytically separate the allocation of this sting-avoidance resource into a mechanism whereby the wasp collective acquires sting-avoidance and a mechanism in which sting-avoidance is then shared out amongst individual wasps. So this resource is, by my definition, not shared out. A second example in which resources are not shared-out subtractably is that of beavers building a channel from their dam to the river bank. I concede that one can analytically separate resource acquisition and resource sharing-out here. But one beaver's using this channel does not exclude other beavers from doing likewise. So this sharing-out is not subtractable.

Almost there. I want now to make Equation (2) easier to work with mathematically. Consider the following constraint on the fitness, ω_i, of each individual, i, in collective g:

$$\omega_i = \left(\frac{1}{n} - \alpha\left[z_i - \mathrm{Exp}_g(z)\right]\right) R_g(\mathbf{z}_g). \tag{3}$$

Let me unpack this equation. $\mathrm{Exp}_g(z)$ is just the average character of the members of collective g. So $[z_i - \mathrm{Exp}_g(z)]$ denotes the degree to which our individual i scores especially highly on pro-social character z. In other words,

Table 2. Fitness of each individual in the subtractability case

	Individuals who interact with a Z individual	Individuals who interact with a non-Z individual
Fitness of Z individuals	$\frac{1}{2}R$	$\left(\frac{1}{2}-\frac{1}{2}\alpha\right)R'$
Fitness of non-Z individuals	$\left(\frac{1}{2}+\frac{1}{2}\alpha\right)R'$	$\frac{1}{2}R''$

whenever an individual has a perfectly average character then this becomes zero and the overall expression reduces to $\frac{1}{n}R_g(\mathbf{z}_g)$. Put differently: whenever this is so, this individual's fitness is equal to collective resources $R_g(\mathbf{z}_g)$ divided by the number of members of the collective n. So whenever an individual is perfectly average, she receives her 'fair share' of collective resources.

Similarly, note that whenever an individual scores especially highly for pro-social character z, then the $-\alpha[z_i - \mathrm{Exp}_g(z)]$ term will be negative, assuming α is positive. So she will enjoy a lesser proportion of the collective's resources and thus she will be less fit. Conversely, whenever an individual scores especially low on z—in other words, she has an especially 'anti-social' character—then this expression will be positive. And so she will enjoy a greater proportion of collective resources and thus will be more fit. So the α parameter denotes the degree to which anti-social individuals can command an unfair share of the resources that the collective has acquired. Thus parameter α measures an important feature of the sharing-out of resources between individuals, as opposed to a feature of collective resource acquisition itself. It is a feature of the fitness-structure generated by the environment.

(Table 2 illustrates the fitness structure that Equation (3) requires in a simple case, namely, in the case of two-membered collectives, and in which an individual either has character z fully or not at all. In formal terms, $z=0$ or $z=1$.)

Take the expression in round brackets in Equation (3) and sum it over all individuals in the collective. Since this necessarily sums to one, it is evident that Equation (3) entails Equation (2). But I don't believe that to assume subtractability in the specific form of Equation (3), rather than more generally in the form of Equation (2), amounts to a significant loss in generality.[19] So from now on I will work with Equation (3) as part of my definition of subtractability, rather than with Equation (2).

Having carefully illustrated what I mean by subtractability, one can now get down to business. I will now show that the multi-level Price decomposition has the ancillary role of answering questions about how character z would evolve

[19] Frank's ([1995]) model, however, satisfies Equation (2) but not Equation (3).

if anti-socially inclined individuals were not permitted unfair access to subtractable resources.

Suppose that the sharing-out of resources amongst individuals is subtractable. Hence it can be characterized by a parameter α that measures the degree to which the fitness-structure in play permits anti-socially inclined individuals to access more than their fair share of collective resources. So intuitively, and as Equation (3) confirms, altering α to become zero will reduce within-collective variation in fitness to zero. In these circumstances, all individuals will receive an equal share of fitness, namely, $R_g(\mathbf{z}_g)$ divided by n. (One example of this is an alteration of the visual environment such that would-be cheaters can be spotted and thereby prevented from stealing extra resources.) But this hypothetical alteration of α is evidently structural collapse to the mean (SCM) alteration. And I've already shown in Section 4 that all SCM alterations satisfy the independence criterion for explanatory value: the effect, ε, of this SCM alteration will be measured by the attendant change to the value of the third term in the multi-level Price decomposition.[20]

My second criterion for explanatory value (interestingness) requires that this alteration to α be of interest to biologists. Note, however, that the genuine sharing-out of resources—as I've defined it—entails that one can analytically separate resource allocation into the acquisition of resources by the collective and the sharing-out of these resources amongst individuals. This in turn entails—again by my definition—that there is an interesting alteration to the mechanism of sharing-out resources amongst individuals, an alteration that does not alter how these resources were acquired by the collective. Therefore all cases of subtractable sharing-out will be cases in which alterations to α are biologically interesting. So my interestingness criterion for explanatory value is, by definition, satisfied in cases in which resources are genuinely shared out.

Here are two such cases, just to illustrate that such cases plausibly exist. Case one: α measures the degree to which visual environment is such that cheating foragers can go undetected, and therefore can steal resources rather than being excluded from them. Case two: in the population genetics example, α measures the degree of so-called segregation distortion, the extent to which the meiotic environment allows selfish alleles to enjoy more than their

[20] Moreover, one can show that the relationship between the third term of the Price equation and α is a linear one. For observe that it follows from Equation (3) that

$$\mathrm{Cov}_g(\omega, z) = \mathrm{Cov}_g\left(\left[\frac{1}{n} - \alpha z + \alpha \mathrm{Exp}_g(z)\right] R_g, z\right) = \alpha R_g(\mathbf{z}_g)\mathrm{Var}_g(z). \quad (4)$$

But one can substitute this into $\mathrm{Exp}[\mathrm{Cov}_g(\omega, z)]$, the third term of the multi-level Price decomposition, to yield $\mathrm{Exp}[\alpha R_g(\mathbf{z}_g)\mathrm{Var}_g(z)]$. And this yields $\alpha\mathrm{Exp}[R_g(\mathbf{z}_g)\mathrm{Var}_g(z)]$. For, being a feature of the environment, α doesn't vary from collective to collective. So the third term of the Price decomposition depends linearly upon α.

fair share of representation in the offspring organisms. These are just two examples of a biologically interesting α parameter. So my interestingness criterion for explanatory value is satisfied non-trivially.

But I've already shown in Section 3 that the third criterion for explanatory value (knowledge) is in general satisfied by the multi-level Price decomposition. So all three of my criteria are satisfied. Thus this section has established an explanatory role for the multi-level Price decomposition in a limited class of cases, namely, cases in which the sharing-out of resources is subtractable. In such cases, the multi-level Price decomposition deepens single-level explanations of the evolution of character z based on population-level character–fitness covariance alone. To put it intuitively, it has the ancillary role of answering questions about what would happen if anti-socially inclined individuals could no longer gain unfair access to subtractable resources.

Recall that Section 4 showed that appealing to SCM alterations cannot establish everything that Sober and Wilson want to establish. For it cannot establish the explanatory value of the multi-level Price decomposition across all cases in general—for example, the case of retaliation in wasps. Instead, the present section has shown how appealing to SCM alterations establishes the explanatory value of the decomposition in the special case in which the sharing-out of resources is more or less subtractable. Unfortunately, I contend, there are no other obvious cases in which SCM alterations have any biological interest. (See my discussion in Section 4.) So it's likely that appealing to SCM alterations can only establish the explanatory value of the multi-level Price decomposition in cases in which resources are more or less subtractable.

6 Other Alterations to Within-Collective Variation

There are hypothetical alterations other than SCM alterations, however, which eliminate within-collective variation. This naturally raises the following question: can one appeal to any of these other alterations in order to establish a further explanatory role for the multi-level Price decomposition? Perhaps the decomposition does indeed have a general explanatory role, or at the very least a role in some cases in which resources are not subtractably shared-out. As I will illustrate momentarily, however, I can't find any such alterations that obviously satisfy the independence and interestingness criteria for explanatory value simultaneously, even for a limited range of cases. Thus it is likely that appealing to (alterations to) within-collective variation can establish no more than Section 5 did: the multi-level Price decomposition is explanatorily valuable in cases in which resources are more or less subtractable.

This section will support my claim here by examining three alternatives to the SCM alteration: the 'increased retaliatory capacity' alteration, the

Table 3. Character collapse to the mean for two three-membered collectives and with $\omega_i = \frac{1}{3}\sqrt[3]{z_i}$

Original z	Original ω	CCM z	CCM ω
3	1	24	2
24	2	24	2
81	3	24	2
–	–	–	–
81	3	192	4
192	4	192	4
375	5	192	4

'homogenizing assortment' (HA) alteration, and the 'character collapse to the mean' (CCM) alteration, as I will label them.

> Character Collapse to the Mean: Consider a collective of vampire bats composed of a few very fit members and many very unfit ones. Imagine, for example, a five-member collective containing individuals with fitnesses $\omega = 1, 1, 1, 2,$ and 10. Imagine altering the character of every member in the collective, and in turn their fitnesses, such that they are all moderately fit. Imagine in particular that this yields fitnesses of $\omega = 3, 3, 3, 3,$ and 3. Thus by altering character, fitnesses have been collapsed to the collective mean. So within-collective variation in fitness has been eliminated. Note that this CCM alteration differs from the SCM alteration in that it does not alter fitness via altering fitness structure; instead, it does so by altering the frequency of the character in the population.

To see an immediate problem for appealing to CCM alterations, calculate the values of the second term in the multi-level Price decomposition for the example given in Table 3: the term is originally 90 but falls to 84 under the CCM alteration. So CCM doesn't just alter the value of the third term of the decomposition,[21] it also alters the value of the second term. In other words, with respect to the CCM alteration in this case, the independence criterion for explanatory value is not satisfied. Therefore one cannot appeal to the CCM alteration to identify an explanatory role for the multi-level Price decomposition for all cases in general.

But this raises the following question: might appeals to CCM establish the explanatory value of the multi-level Price decomposition in a more limited range of cases, rather than across all cases in general? Take, for instance, cases in which collective character maps one-to-one onto collective fitness. One can

[21] Which it alters to zero; see Section 4.

show that the hypothetical CCM alteration does satisfy my independence criterion for explanatory value in such cases. This is because the CCM alteration will preserve collective fitness. And so, given the one-to-one mapping, it will preserve collective character. And so it will, in turn, preserve the covariance of collective fitness and collective character. In other words, CCM will not alter the second term of the multi-level Price decomposition in this case. So the independence criterion for explanatory value is met.

What about the interestingness criterion, however? I certainly do not want to claim that cases of one-to-one mapping are uninteresting as such. Indeed, this range of cases includes as a subset an important range of cases, namely, those in which individual fitness is 'additive'.[22] Additive cases are those in which fitness is a linear function of individual character and collective character: $\omega_i = \lambda z_i + \mu \mathrm{Exp}_g(z)$. Thus collective character maps one-to-one onto collective fitness: $\mathrm{Exp}_g(\omega) = (\lambda + \mu)\mathrm{Exp}_g(z)$.[23]

Instead, what I want to question is the biological interest of the CCM alteration itself. After all, the problems I identified in Section 4 with respect to the SCM alteration can all be extended to CCM. For there is no range of cases—at least obviously—for which hypothetical collapses to the erstwhile mean are more biologically interesting than collapses to any other value (Section 4). Thus it is unlikely that CCM alterations ever satisfy the interestingness criterion for explanatory value, even in a more limited range of cases.

> Homogenizing Assortment: One biologically interesting alteration is the alteration to the mechanism of 'assortment', the mechanism that determines which individuals in a population join themselves into collectives with which other individuals. For example, one might imagine that the mechanism of assortment is altered such that individuals only interact with individuals of a similar character. In the extreme case, then, assortment will be fully homogenous: within-collective variation in character will be zero. And therefore within-collective variation in fitness will be zero. (Thus the HA alteration differs from the CCM alteration in that it does not alter the overall composition of characters in the population, merely how individuals are assorted into collectives.) It is clear that this HA alteration is, in general, biologically interesting. In other words, it satisfies my second criterion for explanatory value.

Unfortunately, with respect to the HA alteration, my independence criterion for explanatory value is not satisfied, except perhaps in a gerrymandered

[22] See (Birch [2014]) for a discussion of assumptions similar to this additivity assumption but in a slightly different context.

[23] I note, incidentally, that cases of one-to-one mapping exclude any form of synergism. In other words, it precludes individuals coordinating their activities so that the benefit to the collective is greater than the sum of each individual's own efforts.

range of cases. To see this, note that HA only alters how individuals in the whole population are grouped into collectives; it preserves the overall composition of characters in the population. But take the very simple case in which an individual's fitness only depends upon her own character. It follows that HA preserves each individual's fitness here. In summary, it preserves the joint distribution of character and fitness in the overall population.[24] In such cases, therefore, HA does not affect character–fitness covariance across the whole population. In formal terms, the effect of HA on the value of the left-hand term of the multi-level Price decomposition is zero. But the attendant change to the third term will be non-zero.[25] It follows that HA also affects the value of the second term. In other words, with respect to the HA alteration, my independence criterion for explanatory value is not satisfied in this very simple case. And there is no obvious range of more complex cases, I contend, for which one might expect HA not to alter the second term as well as altering the third term, or at least not for any non-gerrymandered range of cases. Therefore, I contend, it is unlikely that HA alterations ever satisfy the independence criterion for explanatory value, even in a more limited range of cases.

> Increasing Retaliatory Capacity: Recall the *Polistes* wasp example in which fitness was given by $2\text{Exp}_g(z) - z_i - \frac{1}{2}(-z_i)$. This is a special case of the more general fitness structure $\omega_i = f(\mathbf{z}_g) - p(-z_i)$, where p is the parameter that measures retaliatory capacity (Section 4). Consider the hypothetical alteration in which this parameter is increased by Δp: Queen wasps can, for example, more easily punish lazy workers, or punish them more severely. One can easily show that this increasing retaliatory capacity (IRC) alteration increases the value of the second term of the multi-level Price decomposition, namely, by $\text{Var}[\text{Exp}_g(z)]\Delta p$. Ruling out the trivial case in which there is no variation in collective character, this expression will be non-zero. In other words, IRC doesn't just alter the value of the third term of the decomposition,[26] but also the value of the second term. So the IRC alteration fails the independence criterion for the explanatory value of the decomposition in all non-trivial cases.

[24] I am most grateful to Cedric Patternotte for spotting, prior to publication, a subtle but egregious error at this point.

[25] Homogenizing assortment will eliminate the variation within any collective. So it will eliminate the character–fitness covariance within any collective. Thus it ensures that the value of the third term of the multi-level Price decomposition, $\text{Exp}[\text{Cov}_g(\omega, z)]$, will become zero. Setting aside the trivial case in which within-collective variation was already zero, this demonstrates that the attendant change to the value of the third term is non-zero.

[26] The attendant change to the third term is, one can show, $\text{Exp}[\text{Var}_g(z)]\Delta p$. And this is only zero when there is no within-collective variation in individual character.

To take stock, this section has considered three alterations to within-collective variation: IRC, HA, and CCM. And I've shown decisively that one cannot appeal to the IRC alterations to identify any explanatory role for the multi-level Price decomposition at all. I have also shown decisively that one cannot appeal to the CCM or HA alterations to identify a general explanatory role for the decomposition in all cases. Moreover, it's unlikely that we can find an explanatory role by appealing to CCM or HA in even a more limited range of cases, excluding gerrymandered ranges of cases. So an appeal to any of these three alterations—CCM, IRC, or HA—to establish any explanatory role for the multi-level Price decomposition is unlikely to be successful. Therefore, the SCM alteration from Sections 4 and 5 is the only alteration of within-collective variation to which one might successfully appeal. The tentative conclusion is that appealing to (alterations to) within-collective variation can establish no more than Section 5 did: the multi-level Price decomposition is explanatorily valuable in cases in which resources are more or less subtractable.

7 Alterations to Between-Collective Variation

Sections 4–6 asked whether appealing to (alterations to) within-collective variation can establish the explanatory value of the multi-level Price decomposition. This was prompted by Sober and Wilson's suggestion that the third term of the decomposition measures the effects of within-collective variation. But Sober and Wilson, I've already noted, also place a lot of weight upon an idea that is symmetrical to this one: the second term of the multi-level Price decomposition measures the effects of between-collective variation. If this symmetrical idea is true, then we have an additional strategy for vindicating the decomposition: appeal to alterations to between-collective variation in fitness. Unfortunately, it turns out that it is very difficult to construct a plausible argument that favours Sober and Wilson's symmetrical idea. The following is my best attempt, but one that ultimately fails.

Take a five-member collective with individual fitnesses of $\omega = 1, 3, 6, 6$, and 9; and thus of average fitness of 5. Consider a hypothetical alteration that changes the character of each member such that their fitness is 'boosted' by one unit, resulting in a five-member collective with fitnesses of $\omega = 2, 4, 7, 7$, 10, and thus of average fitness of 6. Note that it's a mathematical fact that this alteration won't alter within-collective variation in fitness. Consider also a second five-member collective with individual fitnesses of $\omega = 1, 6, 8, 10$, and 10, and thus of average fitness of 7. But this time consider a 'boost' of minus one unit, so that this second collective now also has an average fitness of 6. Thus all collectives are altered to have the same collective fitness, in this case 6, eliminating between-collective variation in collective fitness. Consequently,

Table 4. Uniform boosting for two three-membered collectives and with $\omega_i = \frac{1}{3}\sqrt[3]{z_i}$

Original z	Original ω	Boost z	Boost ω
3	1	24	2
24	2	81	3
81	3	192	4
–	–	–	–
81	3	24	2
192	4	81	3
375	5	192	4

this uniform boosting (UB) alteration reduces to zero any covariance of collective fitness with other factors. Therefore $\mathrm{Cov}[\mathrm{Exp}_g(\omega), \mathrm{Exp}_g(z)]$, the second term of the multi-level Price decomposition, will become zero.

Calculate, however, the values of the third term in the multi-level Price decomposition for the example given in Table 4: the term is originally 62 but falls to 56 under the UB alteration. In other words, with respect to UB alteration, the multi-level Price decomposition doesn't in general satisfy the independence criterion for explanatory value. Moreover, let ε denote the effect of this UB alteration upon character–fitness covariance across the whole population—that is, upon the value of the left-hand term in the multi-level Price decomposition. UB alteration having changed the value of the third term, it follows that this effect ε is not measured by the attendant change to the second term of the multi-level Price decomposition. Sober and Wilson's symmetrical idea does not hold in general for all cases.

One response might be to insist that, nevertheless, the attendant change in the second term measures effect ε in a limited but non-gerrymandered class of cases. Take, for example, those cases in which an individual's fitness is a linear function of that individual's own character alone; put in formal terms $\omega_i = mz_i + c$. Whenever the fitness of each member of a collective is uniformly boosted by k, then each member's character will have been uniformly boosted by $\frac{k}{m}$, given this linear relationship. But the logic of covariance has it that $\mathrm{Cov}_g(\omega + k, z + \frac{k}{m}) = \mathrm{Cov}_g(\omega, z)$. So UB alteration preserves the value of the third term in this case. It follows that this effect, ε, is measured by the attendant change to the second term of the multi-level Price decomposition.

Unfortunately, this class of cases is completely irrelevant for present purposes. For there's an intuitive sense in which there is no selection at all at the level of the collective at all in such cases. After all, in such cases individual fitness is not influenced by the collective. And I have no doubt that Sober and

Wilson would agree with this point. This is because, applying their own definition of 'trait groups' ([1998]), there are no genuine collectives in this special case. And hence there is no genuine collective-level selection.

So the problem remains: consider this effect ε of eliminating between-collective variation via UB alterations, that is, the effect upon the value of the left-hand term of the multi-level Price decomposition. I contend that there is no obvious non-gerrymandered class of relevant cases for which this effect, ε, is measured by the attendant change to the second term of the decomposition. So, with respect to the UB alteration, it is unlikely that there are any cases for which the independence criterion for explanatory value holds. So appeals to UB alteration are unlikely to establish any explanatory value for the multi-level Price decomposition. But there are no other obvious, biologically interesting ways—I contend—to alter between-collective variation. I conclude that appeals to (alterations of) between-collective variation are unlikely to establish any explanatory value for the decomposition.

8 Alternative Approaches to Explanatory Depth

This article has taken for granted that the depth of an explanation is in proportion, roughly speaking, to the number of important what-if questions that it allows one to answer. But why should one accept this? I cannot offer a full defence of this view, although interested evolutionary biologists might consult (Woodward [2003]), which has quickly become a philosophical classic. Instead, this section will briefly examine the prospects for an alternative approach to explanatory depth, one that draws upon alternative accounts of explanation.

The first thing to note is that the philosophical literature contains scarcely any alternatives to the what-if account of explanatory depth. Why, for example, did the patient die? Hempel's deductive nomological approach might say that the following was a correct explanation: the patient ingested a large dose of digitalis, and it's a law that all people who ingest that dose will die soon afterwards (Hempel and Oppenheim [1948]). But Hempel's account is not an account of explanatory depth. For it does not offer us a criterion according to which this explanation counts as less deep than an explanation that includes details about how digitalis is metabolized and how it affects the heart. Hempel's approach is an account of explanatory correctness, not an account of the depth of a correct explanation.

Next consider Kitcher's ([1981], [1989]) unificationist approach to explanation. Kitcher provides a criterion for what one might call explanatory promise: the ability of a candidate explanation to deepen one's understanding of what one already knows. And, famously, Kitcher's approach is a 'winner takes all' account. Indeed, it cannot be modified to admit degrees of

explanatory promise on pain of admitting some embarrassing counter-examples (Woodward [2003], p. 368).[27] So, even if one were willing to equate explanatory promise with explanatory depth, Kitcher's approach doesn't delineate degrees of explanatory depth.

Kitcher's approach should not be confused with the more modest—and thereby more plausible—idea that there are at least two virtues with respect to which an explanatory framework such as the multi-level selection framework can be assessed. The first virtue is what I've called depth, which I've urged is to be cashed out in terms of what-if questions. The second virtue is cashed out in terms of the framework's scope of correct application: the broader the range of cases that can be correctly explained within that framework, the more 'unifying' the framework.[28] But it is evident that anyone tempted by this more modest unificationist idea will have no complaints with the assumptions that this article has made about explanatory depth. All that the modest unificationist insists upon is that one also acknowledge the existence of an additional dimension to explanatory frameworks: unification qua broad scope of correct application.

I'm happy to do so. Admittedly, I've said very little about the relative scope of application of the single-level selection and multi-level selection frameworks. But this is because the answer is trivial: the multi-level selection framework has a narrower scope. After all, it embodies an extra restriction, namely, that one's population be partitioned into collectives. So, for this trivial reason, the present consideration concerning breadth of scope is not probative. It does not provide a sense in which multi-level selection explanations add value over and above single-level selection explanations.

Finally, let's consider the causal approach to explanation. Why have I been talking about the explanatory depth of the multi-level Price decomposition, rather than about, as Okasha ([2004b], [2004c]) does, whether the decomposition is 'causally adequate' or 'causally inadequate'? My main reason is that the notion of a decomposition's being causally adequate is incredibly tricky (Okasha [forthcoming]). That is why I have left the discussion in this article incomplete as a far as causal questions are concerned. But one might worry that, in ignoring causation, the discussion in this article is in danger of being not just incomplete but also unsound. I will now address this worry.

I've taken for granted throughout this article that the depth of an explanation is, roughly speaking, in proportion to the number of important what-if questions that it helps to answer. And I've noted that the importance of a what-if question is in part determined by our personal interests. But

[27] Indeed, see (Woodward [2003], Section 8) for what I take to be decisive counter-examples to the view overall.

[28] Birch ([2014], Section 5) proposes this more modest approach, although he seems to suggest that there is a sensible way of aggregating these two virtues into one overall score.

philosophers who favour the causal approach to explanation might wish to place an additional restriction on what counts as an important what-if question. The causal restriction: a what-if question is only important if the correct answer to it cites a cause of the to-be-explained event. I have no doubt that Lewis ([1986]), Lipton ([1991]), Ruben ([1990]), and Woodward ([2003]), amongst others, would endorse this restriction.[29]

Adding this restriction, however, makes no difference to the soundness of the arguments of this article. Firstly, my criticism of Sober and Wilson in Sections 4 and 6 relied primarily on the fact that certain what-if questions are uninteresting. And so my criticism required only that interestingness be a necessary condition for a what-if question to be important. It did not require that interestingness constitute the only necessary condition on importance. Secondly, my positive point in Section 5 relied primarily on the importance of questions about what would happen were parameter α to be different. What happens to my argument if we add the requirement that α has to be a cause of the evolution of social character z, in order for such questions to count as important? Nothing. For there is no reason to think that α—an interesting feature of the environment that determines how much command anti-social individuals have over resources—cannot be a cause of the evolution of character z. So endorsing a causal approach to explanation does not generate a reason to resist the conclusions of this article.

This concludes my defence of the measure of the depth of an explanation as, roughly, the number of important what-if questions that it helps to answer.

9 Conclusion

Sections 2 and 8 built and defended a general framework through which to understand the explanatory role of non-empirical decompositions such as the multi-level Price decomposition. Such decompositions have the ancillary role of describing the constitutive relationships that help glue different factors in our explanatory reasoning together. And I provided three individually necessary and jointly sufficient criteria for a non-empirical decomposition to play this role.

This motivates a search to find a hypothetical intervention that simultaneously meets my independence criterion and my interestingness criterion. Taking my lead from Sober and Wilson, I assume that any such intervention would either be one that (i) eliminates between-collective variation in fitness, or (ii) eliminates within-collective variation in fitness. And this article considered five interventions in total: (i) uniform boosting alterations (UB from Section 7); (ii) increasing retaliatory capacity alterations (IRC from Sections 4 and 6),

[29] But note that, given Lewis's and Woodward's views of the nature of causation, this restriction is a trivial one: roughly speaking, all answers to (the right sort of) what-if-things-had-been-different questions cite causes.

structural collapse to the mean alterations (SCM from Sections 4 and 5), homogenizing assortment alterations (HA from Section 6), and character collapse to the mean alterations (CCM from Section 6).

Only some of these hypothetical alterations turn out to meet my interestingness criterion: HA and IRC alterations are in general interesting; and SCM alteration is interesting whenever resources are subtractable. (In contrast, CCM alteration is of dubious interest.) Similarly, only some of these hypothetical alterations meet my independence criterion. That is, only some of these alterations have their effects measured by a right-hand term of the multi-level Price decomposition: the SCM alteration in all cases, and the CCM alteration in cases of one-to-one mapping of character to fitness. All the other interventions likely fail this criterion in all cases, excluding gerrymandered ones.

In summary, none of these five alterations meet both criteria simultaneously in all cases. Indeed, there isn't even a more limited range of cases for which the IRC, HA, CCM, or the UB alteration meet both criteria simultaneously. However, in the limited case in which resources are subtractable, the SCM alteration does satisfy both criteria. But I assume that these five alterations are the only ones to which one might obviously appeal in order to establish the explanatory value of the multi-level Price decomposition. My conclusion is that the decomposition has explanatory value, most likely, primarily when collective resources are more or less subtractable. Its value is more circumscribed than its champions Sober and Wilson ([1998]) believe.

Let me put the main thrust of the article in intuitive form. What would happen if environmental conditions made it more difficult for anti-socially inclined individuals to access an unfair proportion of the subtractable resources acquired by their collective? I have argued that the explanatory value of the multi-level Price decomposition is that it helps us to answer such questions, questions about what would happen were the 'policing' of subtractable resources strengthened. But, I have shown, it does not help answer questions about other cases, or concerning other policing mechanisms such as retaliatory punishment or homogenizing assortment alteration.[30]

Acknowledgements

I am grateful to Jonathan Birch, Tim Lewens, Samir Okasha, Kim Sterelny, and two anonymous referees for their generous and helpful comments on the manuscript. This research has received funding from the European Research

[30] This raises the question of how the paradigm policing mechanisms identified in (Buss [1987]; Michod [1999]) fit into my scheme for classifying policing mechanisms and, crucially, whether these mechanisms issue in more or less subtractable resources.

Council under the European Union's Seventh Framework Programme (FP7/2007-2013), ERC Grant agreement no. 284123.

Department of History and Philosophy of Science
University of Cambridge
Cambridge, CB2 3RH, UK
cjc84@cam.ac.uk

References

Achinstein, P. [1983]: *The Nature of Explanation*, New York: Oxford University Press.

Baker, A. [2009]: 'Mathematical Explanations in Science', *British Journal for the Philosophy of Science*, **60**, pp. 611–63.

Batterman, R. W. [2010]: 'On the Explanatory Role of Mathematics in Empirical Science', *British Journal for the Philosophy of Science*, **61**, pp. 1–25.

Birch, J. [2014]: 'Hamilton's Rule and Its Discontents', *British Journal for the Philosophy of Science*, **65**, pp. 381–411.

Buss, L. W. [1987]: *The Evolution of Individuality*, Princeton: Princeton University Press.

Clarke, E. [2013]: 'The Multiple Realizability of Biological Individuals', *Journal of Philosophy*, **110**, pp. 413–35.

Damuth, J. and Heisler, I. L. [1988]: 'Alternative Formulations of Multi-Level Selection', *Biology and Philosophy*, **3**, pp. 407–30.

Darwin, C. [2008]: *On the Origin of Species by Means of Natural Selection*, Oxford: Oxford University Press.

Dawkins, R. [1976]: *The Selfish Gene*, Oxford: Oxford University Press.

Dawkins, R. [1982]: *The Extended Phenotype*, Oxford: Oxford University Press.

Dugatkin, L. A. and Reeve, H. K. [1994]: 'Behavioural Ecology and Levels of Selection: Dissolving the Group Selection Controversy', in P. Slater, J. Rosenblatt, C. Snodown and M. Milinski (*eds*), *Advances in the Study of Behaviour*, Volume 23. San Diego: Academic Press, pp. 102–34.

Frank, S. A. [1995]: 'Mutual Policing and Repression of Competition in the Evolution of Co-operative Groups', *Nature*, **377**, pp. 520–2.

Frank, S. A. [1998]: *Foundations of Social Evolution*, Princeton: Princeton University Press.

Gamboa, G. J., Wacker, T. L., Scope, J. A., Cornell, T. J. and Shellman-Reeve, J. [1990]: 'The Mechanism of Queen Regulation of Foraging by Workers in Paper Wasps (*Polistes fuscatus*, Hymenoptera: Vespidae)', *Ethology*, **85**, pp. 335–43.

Giraldeau, L.-A. and Caraco, T. [2000]: *Social Foraging Theory*, Princeton: Princeton University Press.

Goodnight, C. J., Schwartz, J. M. and Stevens, L. [1992]: 'Contextual Analysis of Models of Group Selection, Soft Selection, Hard Selection, and the Evolution of Altruism', *American Naturalist*, **140**, pp. 743–61.

Hamilton, W. D. [1964]: 'The Genetical Evolution of Social Behaviour', *Journal of Theoretical Biology*, **7**, pp. 1–16.

Hamilton, W. D. [1975]: 'Innate Social Aptitudes in Man: An Approach from Evolutionary Genetics', *Biosocial Anthropology*, New York: Wiley, pp. 133–55.

Hart, H. L. A. and Honore, A. [1965]: *Causation in the Law*, Oxford: Clarendon–OUP. Citations refer to the Second Edition (1985).

Heisler, I. L. and Damuth, J. [1987]: 'A Method for Analyzing Selection in Hierarchically Structured Populations', *American Naturalist*, **130**, pp. 582–602.

Hempel, C. G. and Oppenheim, P. [1948]: 'Studies in the Logic of Explanation', *Philosophy of Science*, **15**, pp. 135–75.

Hull, D. L. [1981]: 'Units of Evolution: A Metaphysical Essay', in U. J. Jensen and R. Harré (*eds*), *The Philosophy of Evolution*, Brighton: Harvester Press, pp. 23–44.

Kerr, B. and Godfrey-Smith, P. [2002]: 'Individualist and Multi-Level Perspectives on Selection in Structured Populations', *Biology and Philosophy*, **17**, pp. 477–517.

Kitcher, P. [1981]: 'Explanatory Unification', *Philosophy of Science*, **48**, pp. 507–31.

Kitcher, P. [1989]: 'Explanatory Unification and the Causal Structure of the World', in P. Kitcher and W. Salmon (*eds*), *Scientific Explanation*, Minneapolis: University of Minnesota Press, pp. 410–505.

Lange, M. and Rosenberg, A. [2011]: 'Can There Be *a priori* Causal Models of Natural Selection?', *Australasian Journal of Philosophy*, **89**, pp. 591–9.

Lewis, D. K. [1986]: 'Causal Explanation', in his *Philosophical Papers*, Volume 2. Oxford: Oxford University Press.

Lipton, P. [1991]: *Inference to the Best Explanation*, London: Routledge.

Lloyd, E. A. [1986]: 'Evaluation of Evidence in Group Selection Debates', *Proceedings of the Biennial Meeting of the Philosophy of Science Association*, pp. 483–93.

Lloyd, E. A. [2000]: 'Groups on Groups: Some Dynamics and Possible Resolution of the Units of Selection Debates in Evolutionary Biology', *Biology and Philosophy*, **15**, pp. 389–401.

Lyttle, T. W. [1991]: 'Segregation Distorters', *Annual Review of Genetics*, **25**, pp. 511–57.

Maynard Smith, J. and Szathmáry, E. [1995]: *The Major Transitions in Evolution*, Oxford: Oxford University Press.

McElreath, R. and Boyd, R. [2007]: *Mathematical Models of Social Evolution: A Guide for the Perplexed*, Chicago: University of Chicago Press.

Michod, R. E. [1999]: *Darwinian Dynamics: Evolutionary Transitions in Fitness and Individuality*, Princeton: Princeton University Press.

Okasha, S. [2004a]: 'The "Averaging Fallacy" and the Levels of Selection', *Biology and Philosophy*, **19**, pp. 167–84.

Okasha, S. [2004b]: 'Multi-Level Selection and the Partitioning of Covariance: A Comparison of Three Approaches', *Evolution*, **58**, pp. 486–94.

Okasha, S. [2004c]: 'Multi-Level Selection, Covariance, and Contextual Analysis', *British Journal for the Philosophy of Science*, **55**, pp. 481–504.

Okasha, S. [2005]: 'Maynard Smith on the Levels of Selection Question', *Biology and Philosophy*, **20**, pp. 989–1010.

Okasha, S. [2006]: *Evolution and the Levels of Selection*, Oxford: Oxford University Press.

Okasha, S. [forthcoming]: 'The Relationship between Kin Selection and Multi-Level Selection', *British Journal for the Philosophy of Science*, doi:10.1093/bjps/axu047.

Pincock, C. [2007]: 'A Role for Mathematics in the Physical Sciences', *Noûs*, **41**, pp. 253–75.

Price, G. R. [1970]: 'Selection and Covariance', *Nature*, **227**, pp. 520–1.

Price, G. R. [1972]: 'Extension of Covariance Selection Mathematics', *Annals of Human Genetics*, **35**, pp. 485–90.

Robertson, A. [1966]: 'A Mathematical Model of the Culling Process in Dairy Cattle', *Animal Production*, **8**, pp. 95–108.

Ruben, D.-H. [1990]: *Explaining Explanation*, London: Routledge.

Sarkar, S. [2008]: 'A Note on Frequency Dependence and the Levels/Units of Selection', *Biology and Philosophy*, **23**, pp. 217–28.

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

Sober, E. [2011]: '*A priori* Causal Models of Natural Selection', *Australasian Journal of Philosophy*, **89**, pp. 571–89.

Sober, E. and Wilson, D. S. [1998]: *Unto Others: The Evolution and Psychology of Unselfish Behavior*, Cambridge, MA: Harvard University Press.

Sterelny, K. [1996]: 'The Return of the Group', *Philosophy of Science*, **63**, pp. 562–84.

van Fraassen, B. C. [1977]: 'The Pragmatics of Explanation', *American Philosophical Quarterly*, **14**, pp. 143–50.

Woodward, J. [2003]: *Making Things Happen: A Theory of Causal Explanation*, Oxford: Oxford University Press.