

Three Paths Through the Levels of Selection


DB Krupp

Department of Interdisciplinary Studies and Department of Psychology
Lakehead University

August 7, 2024

To appear in the *Handbook of Evolutionary Psychology*

Author Note

DB Krupp  <https://orcid.org/0000-0001-6297-7739>

I am indebted to Martin Daly, Tom Dickins, Maryanne Fisher, and Jun Otsuka for their exceptionally thoughtful feedback. This work was funded by the Social Sciences and Humanities Research Council of Canada.

Correspondence should be addressed to DB Krupp, Department of Interdisciplinary Studies, Lakehead University, 500 University Ave., Orillia, Ontario L3V 0B9. E-mail: dbkrupp@saltlab.org

Three Paths Through the Levels of Selection

[A] generation has grown up, especially in America, that scatters the name “group selection” around like confetti. It is littered over all kinds of cases that used to be (and by the rest of us still are) clearly and straightforwardly understood as something else, say kin selection. (Dawkins, 1976/2006, p. 297)

[W]hen everything that was ever called group selection can now be described in terms of inclusive fitness theory, it is time to take stock of the original empirical issues at stake. (D. S. Wilson & Wilson, 2007, p. 337)

At which levels of social organization does natural selection operate: the gene, the individual, or the group? This question has fueled one of the most enduring and acrimonious debates in all of evolutionary biology, and the answer is important because, to fully characterize an adaptation, we need to know the forces that shaped it. If gene-, individual-, and group-level selection are distinct processes, then their products will be distinct as well.

The suggestion that there may be more than one level of selection dates back to Darwin’s own writing, and Fisher, Haldane, and Wright each explored the idea, albeit briefly, in the early days of the Modern Synthesis (Darwin, 1859, 1871; Fisher, 1930; Haldane, 1932; Wright, 1931). But the debate truly came to life sixty years ago, when Wynne-Edwards (1962) proposed that organisms were designed by a process of group-level selection to “homeostatically” regulate reproduction, say, or oust group members “in order to retrieve the correct balance between population-density and resources” (p. 9). Maynard Smith (1964) and Williams (1966) took exception to this idea, publishing sharp critiques of it and, more broadly, of the likelihood that organisms have been designed to act for the good of their groups. It was against this backdrop that Hamilton (1963, 1964) introduced inclusive fitness theory—also known as kin selection—an explanation for the evolution of social behavior that operates at the individual level, and Dawkins (1976/2006) advanced the gene’s-eye view, a perspective that casts the gene “as the nearest thing we have to a fundamental, independent agent of evolution” (p. 44). Still, it wasn’t long before group-level selection arguments began to resurface in more sophisticated forms (Hamilton, 1975; Price, 1972; D. S. Wilson, 1975).

Since then, opinions have become spirited and consensus seems elusive. Indeed, the quotations at the top of this chapter testify to the remarkable polarity of views. Whereas some aver that group-level selection is not a useful concept, and should be discarded in favor of lower-level theories such as kin selection, others assert that kin selection is just a special case of group-level selection, and only a hierarchical approach gives the whole picture.

To grapple with this, scholars have travelled three paths through the levels of selection. There is the discursive path, which takes premises or assumptions about selection and adaptation and constructs logical, but informal, arguments from them. There is the statistical path, which takes formal statistical methods and interprets the levels of selection through them. And there is the causal path, only recently forged, which takes knowledge of causal relationships in biological systems and formalizes this to determine both the level of selection and the appropriate statistical approach to use in its analysis. I suggest it is this

third path that provides a clear way through—allowing us to conclude that, for a given trait, there is a proper level of selection, and it is the one that is causally apt (Birch & Okasha, 2015; Godfrey-Smith & Kerr, 2013; Okasha, 2006, 2016; Sober, 1984).

Before we tour these paths, two brief notes. First, this analysis is greatly inspired by Okasha (2006), and a reader wishing to understand the levels of selection debate in greater depth would do well to turn there. Second, I have devoted quite a bit of space to unpacking the statistical and causal methods of the levels of selection, as evolutionary psychologists are rarely trained in them, but they are needed to understand the debate. The cost, however, is that I do not also have the space to cover some of their wider applications, such as species selection and cultural evolution. Nevertheless, I hope to make the traditional gene-organism-group case general enough that it can be ported to other domains without too much difficulty.

The Discursive Path

Benefits to groups can arise as statistical summations of the effects of individual adaptations.... As a very general rule, with some important exceptions, the fitness of a group will be high as a result of this sort of summation of the adaptations of its members. On the other hand, such simple summations obviously cannot produce collective fitness as high as could be achieved by an adaptive organization of the group itself. (Williams, 1966, pp. 16–17)

Evolutionary theory is heavily mathematized (Otsuka, 2019). But behind every equation is a premise or assumption from which discursive arguments can be built or dismantled. These sorts of verbal accounts have paved the way for many important ideas in evolutionary biology. Darwin, for instance, developed the theory of evolution by natural selection with just a few key premises: organisms vary in their traits; parents pass their traits on to their offspring; resources are limited, leading to competition for survival and reproduction; and some traits are better suited than others to such competition. If these premises are true, then selection necessarily follows.

A closely related premise, also likely true, is that natural selection is a causal process.¹ This claim is encapsulated in the following definition of selection, which I credit to Martin Daly: “the differential reproduction of types as a consequence of their differences.” And it is clearly what Darwin (1859, p. 61) had in mind when he wrote “any variation, however slight and from whatever cause proceeding, if it be in any degree profitable to an individual of any species, in its infinitely complex relations to other organic beings and to external nature, will tend to the preservation of that individual, and will generally be inherited by its offspring.” Put simply, traits are selected for *because* of their “profitable” effects on fitness.

¹ Surprisingly, this position is controversial in certain circles. The glossaries of many popular evolution textbooks contain tortured definitions of natural selection that appear intent on avoiding any mention of the effect of phenotypes on fitness (Gregory, 2009). Moreover, a school of philosophy has recently cropped up arguing that contemporary evolutionary theory only traffics in statistical relationships, not causal ones (reviewed in Otsuka, 2016b, 2019). Attempts to strip the theory of natural selection of its causal assumptions are misguided: selection has predictive and explanatory power precisely because it makes these assumptions (Otsuka, 2019).

The historical integration of Darwinian and Mendelian views made it clear that, with respect to natural selection, genes are a beginning and fitness is an end. Hence, traits undergoing selection will tend to be heritable, which is to say that variance in genes causes variance in traits, at least in part. Likewise, traits undergoing selection will tend to affect the gene's fitness, leading to replication. So if selection is operating on a heritable trait, then it will in some sense be selecting genes and the trait will in some sense be serving the gene's fitness. This argument is the heart of the gene's-eye view of evolution: that genes, as near-immortal "replicators" being reproduced from one generation to the next, constitute the proper unit of selection and deserve our attention, whereas organisms, as mere transitory "vehicles" for the genes, are only a distraction (Ågren, 2021; Dawkins, 1982, 1976/2006).

Nevertheless, the phrase "in some sense" is doing a lot of work in the preceding paragraph. The causal story of selection and adaptation depends on more than its beginning (genes) and its end (fitness), and glossing over these details can lead us to some awkward conclusions. For instance, altruism and spite were historically described as alternative kinds of self-sacrifice, the actor paying a fitness cost to help recipients in the former instance and to harm recipients in the latter instance (Hamilton, 1963, 1970). But it has more recently been argued that, in a population of fixed size, altruism and spite could be considered two names for the same thing, on the grounds that a gene that increases the fitness of some parties necessarily decreases the fitness of others (Lehmann et al., 2006). It does not matter whether the actor also helps or harms a recipient, because other recipients elsewhere will feel opposing effects when the population can neither shrink nor grow. By extension, the same is true of selfish and mutually beneficial behavior, since both entail an increase in the actor's fitness at the expense of others. And if we distill this argument even further, treating the actor as just another recipient, we now find that there is no difference at all between altruism, spite, selfishness, or mutual benefit, because each of these presupposes the same effect—that a gene increases its fitness at the expense of its rivals (Krupp, 2013). The result is an amorphous, inscrutable muddle of "behavior."

All of this is technically true. Yet, any student of social evolution knows that altruism does not look like spite and selfishness does not look like mutual benefit. Adaptations designed to feed relatives are not the same as adaptations designed to poison nonrelatives. Adaptations designed to combat rivals are not the same as adaptations designed to attract mates. Siderophores are not bacteriocins (Griffin et al., 2004; Inglis et al., 2009), horns are not nuptial gifts (Gwynne, 2008; McCullough et al., 2014), and the reason is that adaptations are responses to their direct causes, and they speak to their direct effects (Krupp, 2013; Okasha, 2016; Patel et al., 2020).

Thermostats and Sorting Toys

The importance of direct causes and effects can be seen with the help of a thermostat, which turns on the furnace when the temperature drops too low and the air conditioning when the temperature climbs too high. Temperature is its direct cause, and even though there are many possible causes of temperature change, such as latitude and air pressure, the thermostat reacts only to temperature, not to the causes of temperature. Indeed, we could heat just the thermostat's sensor in an otherwise cold room, and the air

conditioning would still turn on. By the same token, once the thermostat turns on, there are many possible downstream effects. The occupants could become more comfortable or they could overheat, and the process could strain the electrical grid or contribute to climate change. The thermostat is ignorant of all of this. Whatever the initial reason and whatever the eventual effect, the function of the thermostat is to regulate the temperature.

On this logic, understanding the level of selection or adaptation requires more than simply pointing to “genes” as the input and “fitness” as the output of a black box. An adaptation optimized by lower-level selection (e.g. at the gene level) will not look the same as an adaptation optimized by higher-level selection (e.g. at the organism level), and genes and fitness alone cannot help to distinguish them. What makes this so challenging is that there will often be alternative characterizations of an adaptation, and these will demand alternative explanations (Williams, 1966). For example, do humans punish free-riders in social dilemmas because they have evolved group-level adaptations, such as mechanisms of strong reciprocity (Bowles & Gintis, 2011), or because they are applying individual-level adaptations that fortuitously benefit other members of their groups, such as mechanisms of reputation management (Yamagishi et al., 2012)?

As is widely known, correlation is not causation, and so a trait may increase in the population without being the target of selection. To borrow an example from Sober (1984), imagine a toy containing large, white marbles and small, gray marbles (Fig. 1). The toy is divided into two tiers, with all of the marbles initially together in one tier, and the divider separating the tiers has holes that allow only the small marbles to pass through to the second tier. Turn the toy around enough, and you will wind up with the marbles fully sorted across the two tiers, with the larger stuck on one side and the smaller now on the other. The marbles are sorted because the divider filters them according to size. Thus, it is reasonable to say that the marbles in both tiers have been selected *for* size. Note, however, that the marbles also happen to be sorted by color, because size and color are confounded. Still, while there is a sense in which there has been selection *of* color, it would make little sense to claim that there was selection *for* color (Sober, 1984). Sorting by color is merely incidental to sorting by size, as nothing about its direct cause (marble size) pertains to color. There is a risk, then, of confounding direct effects with indirect ones, thereby confusing adaptations with their byproducts (Okasha, 2006; Williams, 1966).

A Semantic Miasma

A further complication to discursive arguments is the “semantic miasma” (Salt, 1979, p. 145) of the discourse, in which foundational terms take on wildly different meanings from one paper to the next. For example, “altruism” implies a net, lifetime fitness cost at the population level to the actor in the inclusive fitness sense of the word, but it does not necessarily imply anything beyond a cost relative to group members in the multilevel selection sense. Thus, a selfish behavior in an inclusive fitness model could be considered altruistic in a multilevel model (West et al., 2007). Similarly, “group fitness” seems to suggest the survival and reproduction of whole groups, but it is nevertheless more commonly understood and modelled as the average survival and reproduction of the individuals that make up these groups (A. J. Arnold & Fristrup, 1982; Heisler & Damuth, 1987). While these two meanings can sometimes coincide, they refer to very different

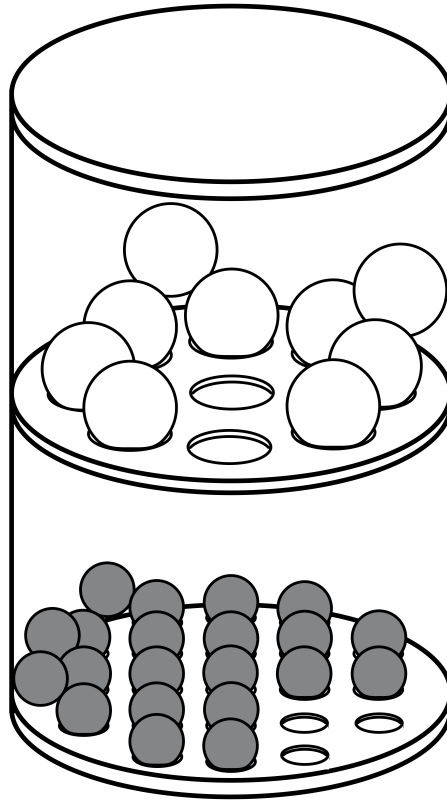


Figure 1

A sorting toy, adapted from Sober (1984). The toy is divided into two tiers, and the divider contains holes that only allow the smaller, gray marbles to pass through. If turned enough, this sorts the larger, white marbles into one tier and the smaller, gray marbles into the other.

things: group fitness is an aggregate property when measured by the average success of individuals within groups in making new individuals, known as collective fitness₁, but it is an emergent property when measured by the success of the groups themselves in making new groups, known as collective fitness₂ (Okasha, 2006). And since conceptions of group fitness form the basis of most multilevel selection models, these two different types of group fitness equate to two different types of multilevel selection—one concerned with the reproduction of individuals (multilevel selection 1) and the other concerned with the reproduction of groups (multilevel selection 2; Heisler & Damuth, 1987; Okasha, 2006).

Emergence is a recurring theme in the levels of selection debate, even beyond definitions of collective fitness. It arises as a natural result of thinking about groups, because an effect is broadly emergent when it depends on the contributions of multiple causes. That is, emergent properties often have a quality of “groupness” about them (Krupp, 2016). But what is meant, exactly, by emergence is hard to pin down (Okasha,

2006, 2014). It is commonly argued that group-level selection occurs when emergent properties are non-additive or frequency-dependent effects of their lower-level causes, though which properties—traits, fitness, or their relation—must be emergent is a matter of some dispute (Goodnight et al., 1992; Lloyd, 1988; Salt, 1979; Smaldino, 2014; Vrba, 1989; Wimsatt, 1980). Others disagree with the non-additivity premise entirely, seeing little reason to discount aggregate effects of groups (Okasha, 2006, 2014; Williams, 1992; D. S. Wilson & Sober, 1994a). In the end, the relevance of emergent properties is unclear.

Because it is easy to get lost in the imprecision of verbal arguments, the discursive path can only lead so far through the levels of selection. There is simply too much ambiguity over the nuances of important concepts and processes to get hold of a convincing solution to the problem. Recognizing this, many travelers have instead placed their faith in the value of statistical models, in the hopes that a more formal approach will offer some clarity.

The Statistical Path

Partitions never change the total effect, and any total effect may be partitioned in various ways. Partitions are simply notational conventions and tools of reasoning. These tools may show logical connections and regularities among otherwise heterogeneous problems. Because alternative partitions are always possible, choice is partly a matter of taste. The possibility of alternatives leads to fruitless debate. Some authors inevitably claim their partition as somehow true; other partitions are labeled false when their goal or method is misunderstood or denigrated. (Frank, 1998, p. 12)

As Maynard Smith's (1964) and Williams' (1966) critiques made the rounds, group-level selection's star began to fade and individual-level theories of social evolution rose to prominence. Inclusive fitness theory, in particular, could handle an ever-expanding range of biological problems, including sex ratio evolution, kin discrimination, parent-offspring conflict, sibling rivalry, and genomic imprinting (West et al., 2008). It produced formal models, intuitive explanations, and surprising results.

But an important development managed to cement inclusive theory and revive group-level selection at the same time (Hamilton, 1996; Harman, 2010). A simple mathematical identity,² published in *Nature* thanks to a plot cooked up by Hamilton, could be used to derive inclusive fitness theory from first principles or to partition selection into a hierarchical structure, making it possible to study evolutionary change at different levels of organization within a single, coherent framework. This is the Price equation (Price, 1970).

The Price equation is a complete and fully general description of evolutionary change in any property, though our focus here will be to track the change of a trait in the population over successive generations. It can be derived from just a handful of ingredients—population size, trait values, reproductive success, and a way to distinguish

² “Simple” is a relative term in the world of mathematics and statistics, but I have tried to make the equations and expressions in this chapter comprehensible to those with an undergraduate education in the social sciences. I hope readers find some value in working through the math here.

individuals—and a bit of algebra. The Price equation can take a number of different forms, but let us begin with a common one, which is derived in Box 1:

$$\Delta\bar{z} = \overbrace{\text{Cov}(w_i, z_i)}^{\text{selection}} + \overbrace{\text{E}(w_i\Delta z_i)}^{\text{transmission bias}}. \quad (1)$$

On the left-hand side of equation (1), $\Delta\bar{z}$ is the change in the population average trait value or phenotype between parental and offspring generations. On the right-hand side, $\text{Cov}(w_i, z_i)$ is the covariance between relative fitness (w_i) and trait value (z_i) of individuals in the parental generation, and $\text{E}(w_i\Delta z_i)$ is the expected value or average change in trait value (Δz_i) between parents and their offspring, weighted by relative fitness. The subscript i denotes an index for each individual in the parental population (Box 1).

Box 1. Deriving the Price Equation.

The Price equation can seem daunting, but it is fairly easy to derive. We will do that here, keeping three caveats in mind. First, things will get messy, but will be tidied up at the end. Second, it helps to make a few simplifying assumptions in the derivation, but they are not strictly necessary. Finally, this derivation takes a few liberties with statistical notation, as is common practice (for helpful discussions, see Birch, 2017; Marshall, 2015).

Consider an asexual population of N individuals, each of whom has been assigned a unique index i . Individual i bears a trait of value z_i , such as its willingness to fight or whether it grows wings. The average trait value in the population, \bar{z} , is simply the sum of all the trait values divided by the number of individuals, or $\bar{z} = (\sum_i z_i)/N$. Another way of expressing \bar{z} is $\text{E}(z_i)$, which can be read as the expected value of z_i ; this will be useful later.

Next, individuals produce q_i offspring and die, so that there is a second population consisting only of the first population's descendants. The average absolute fitness in the parental population, \bar{q} , is the sum of all offspring produced divided by the number of individuals in the parental population, or $\bar{q} = (\sum_i q_i)/N$, and the fitness of i relative to the population average is therefore $w_i = q_i/\bar{q}$. The upshot of using relative fitness is that it will on average be $(\sum_i w_i)/N = 1$ when taken over the population; this will also be useful later. Collecting terms from the parental population, there are now individual and average trait values (z_i and \bar{z}) and absolute fitness, average absolute fitness, and relative fitness (q_i , \bar{q} , and w_i).

To study evolutionary change from parent to offspring populations, it is helpful to keep track of the relation between parental and offspring trait values by assigning a parent's index to its offspring. That is, offspring bear their own trait values but keep their parent's index, connecting the two generations in the analysis. Let the average trait value over all of i 's offspring in the new population be z'_i , where the prime indicates offspring rather than parental values. Notably, offspring trait values may deviate from their parent's for a number of reasons, a phenomenon known as transmission bias. Defining transmission bias as the difference between i 's average offspring trait value and i 's own trait value, $\Delta z_i = z'_i - z_i$, it follows that $z'_i = z_i + \Delta z_i$. Moreover, the average

offspring trait value in the population, \bar{z}' , is the sum of all average offspring trait values weighted by relative parental fitness (because some individuals may produce more offspring than others) and divided by the number of individuals in the parental population, or $\bar{z}' = (\sum_i w_i z'_i)/N$.

Having computed the different terms, it is possible to now work with them algebraically. Evolutionary change between parent and offspring generations can be represented as the change in the average trait value between generations,

$$\Delta\bar{z} = \bar{z}' - \bar{z}. \quad (\text{B1.1})$$

Using the above definitions of \bar{z}' and \bar{z} , equation (B1.1) can be rewritten as

$$\Delta\bar{z} = \frac{\sum_i w_i z'_i}{N} - \frac{\sum_i z_i}{N}. \quad (\text{B1.2})$$

Moreover, it was previously shown that $z'_i = z_i + \Delta z_i$, so equation (B1.2) can be rewritten as

$$\Delta\bar{z} = \frac{\sum_i w_i (z_i + \Delta z_i)}{N} - \frac{\sum_i z_i}{N}. \quad (\text{B1.3})$$

Now everything is written entirely in terms of the parental population. Expanding equation (B1.3) yields

$$\Delta\bar{z} = \frac{\sum_i w_i z_i}{N} + \frac{\sum_i w_i \Delta z_i}{N} - \frac{\sum_i z_i}{N}. \quad (\text{B1.4})$$

Recall that the average relative fitness of the population $(\sum_i w_i)/N = 1$, so the final term on the right-hand side of equation (B1.4) can be multiplied by this without changing its value to give

$$\Delta\bar{z} = \frac{\sum_i w_i z_i}{N} + \frac{\sum_i w_i \Delta z_i}{N} - \frac{\sum_i w_i}{N} \times \frac{\sum_i z_i}{N}. \quad (\text{B1.5})$$

Finally, rearranging the terms of equation (B1.5) gives:

$$\Delta\bar{z} = \frac{\sum_i w_i z_i}{N} - \frac{\sum_i w_i}{N} \times \frac{\sum_i z_i}{N} + \frac{\sum_i w_i \Delta z_i}{N}. \quad (\text{B1.6})$$

At this stage, it is possible to apply a pair of definitions from statistics to obtain the form of equation (1) in the main text. The first definition is that of an expected value: as noted above, the expected value of a variable x is its average value over the population, or $E(x_i) = (\sum_i x_i)/N$. (This could also be written as \bar{x} instead, but $E(x_i)$ is more commonly used in this case.) Each of the terms on the right-hand side of equation (B1.6) are expected values in this sense, so can be rewritten as

$$\Delta\bar{z} = E(w_i z_i) - E(w_i)E(z_i) + E(w_i \Delta z_i). \quad (\text{B1.7})$$

The second definition is that of covariance: the covariance between two variables x_i

and y_i is the difference between the expected value of their product and the product of their expected values, or $\text{Cov}(x_i y_i) = \text{E}(x_i y_i) - \text{E}(x_i)\text{E}(y_i)$. Together, the first and second terms on the right-hand side of equation (B1.7) take this form, and so can be rewritten as

$$\Delta \bar{z} = \text{Cov}(w_i, z_i) + \text{E}(w_i \Delta z_i), \quad (\text{B1.8})$$

which is the same as equation (1) in the main text.

Following Price (1970), equation (1) is usually interpreted as follows: the change in the average trait over generations is the sum of the effects of selection and transmission bias on the trait.³ The effect of selection is thought to be captured by the covariance term, $\text{Cov}(w_i, z_i)$, because selection entails a statistical association between traits and fitness. Likewise, the effect of transmission bias is thought to be captured by the expectation term, $\text{E}(w_i \Delta z_i)$, because it measures the population average, fitness-weighted deviation of offspring trait values from those of their parents, due to mutation or drift, for instance. This is meant to separate selection from other forms of evolutionary change.

As the name suggests, the levels of selection problem is concerned with selection, not transmission bias. Common practice is to assume that there is no such bias, reducing equation (1) to

$$\Delta \bar{z} = \text{Cov}(w_i, z_i). \quad (2)$$

Eliminating the transmission bias term from the Price equation requires some tricky assumptions, but it can nonetheless be useful to do this for the purposes of abstraction (Birch, 2017).

The Price Multilevel Partition

The Price equation is a common starting point for theories of social evolution. It can be used to derive individual-level descriptions of selection, including a general form of Hamilton’s rule, $rb - c > 0$, that uses regression coefficients to define genetic relatedness (r), costs to the focal individual or actor (c), and benefits to partners or recipients (b) (Box 2; Birch, 2017; Gardner et al., 2011; Queller, 1992). It can be used to describe selection taking place at the level of alleles (Gardner & Welch, 2011). And it can be used to partition selection into the lower- and higher-level components of a hierarchical structure, which can then be interpreted within a multilevel selection framework (Hamilton, 1975; Okasha, 2004, 2006; Price, 1972).

To reproduce that framework here, imagine a population subdivided into discrete groups of an equal number of m individuals—a simplifying assumption, but not one that is required—and index the groups by k . Let z_{jk} be the trait value and let w_{jk} be the relative fitness of the j th individual in the k th group. Furthermore, define the average trait value and average relative fitness of the individuals in the k th group as $Z_k = (\sum_j z_{jk})/m$ and

³ After Fisher (1930), there are some forms of the Price equation in which the latter term is referred to as the effect of the “environment,” but this can be confusing (Frank, 2012). In any case, note that the “E” in $\text{E}(w_i \Delta z_i)$ here refers to an expectation, not to the environment.

$W_k = (\sum_j w_{jk})/m$, respectively. Z_k is thus the group trait value and W_k is group fitness of the collective fitness₁ variety.

Box 2. Deriving Hamilton's Rule from the Price Equation.

As the Price equation is a general description of evolutionary change, Queller (1992) was able to derive a general form of Hamilton's rule from it. First, let the trait z_i be the breeding value of the i th individual, denoted p_i . A breeding value can be thought of as the heritable component of i 's phenotype, as predicted by a linear combination of the effects of all relevant alleles. Second, let \hat{p}_i be the breeding value of i 's average partner and let \bar{p} be the average breeding value of the population (Birch, 2017; Gardner et al., 2011). Thus, equation (2) of the main text becomes

$$\Delta\bar{p} = \text{Cov}(w_i, p_i). \quad (\text{B2.1})$$

Next, let us predict relative fitness, w_i , using a least-squares multiple regression equation of the effects of p_i and \hat{p}_i :

$$w_i = \alpha + \beta_1 p_i + \beta_2 \hat{p}_i + \epsilon_{w_i}, \quad (\text{B2.2})$$

where α is baseline fitness (a constant); $\beta_1 = \beta_{w_i p_i \cdot \hat{p}_i}$ is the partial regression coefficient of i 's fitness on breeding value, adjusting for partner breeding value; $\beta_2 = \beta_{w_i \hat{p}_i \cdot p_i}$ is the partial regression coefficient of i 's fitness on partner breeding value, adjusting for i 's breeding value; and ϵ_{w_i} is the residual term.

Substituting equation (B2.2) into equation (B2.1) gives

$$\begin{aligned} \Delta\bar{p} &= \text{Cov}(\alpha + \beta_1 p_i + \beta_2 \hat{p}_i + \epsilon_{w_i}, p_i) \\ &= \text{Cov}(\alpha, p_i) + \beta_1 \text{Cov}(p_i, p_i) + \beta_2 \text{Cov}(\hat{p}_i, p_i) + \text{Cov}(\epsilon_{w_i}, p_i). \end{aligned} \quad (\text{B2.3})$$

Given that α is a constant and p_i cannot covary with ϵ_{w_i} , since residuals are not correlated with predictors in the model, equation (B2.3) reduces to

$$\Delta\bar{p} = \beta_1 \text{Cov}(p_i, p_i) + \beta_2 \text{Cov}(p_i, \hat{p}_i). \quad (\text{B2.4})$$

Now, let us rewrite the right-hand side of equation (B2.4) by pulling out $\text{Cov}(p_i, p_i)$, which is equal to the variance in p_i , or $\text{Var}(p_i)$, to obtain

$$\begin{aligned} \Delta\bar{p} &= \left[\beta_1 + \beta_2 \frac{\text{Cov}(p_i, \hat{p}_i)}{\text{Cov}(p_i, p_i)} \right] \text{Cov}(p_i, p_i) \\ &= \left[\beta_1 + \beta_2 \frac{\text{Cov}(p_i, \hat{p}_i)}{\text{Var}(p_i)} \right] \text{Var}(p_i). \end{aligned} \quad (\text{B2.5})$$

$\text{Cov}(p_i, \hat{p}_i)/\text{Cov}(p_i, p_i) = \text{Cov}(p_i, \hat{p}_i)/\text{Var}(p_i)$ is the standard regression definition of genetic relatedness, which can be denoted simply as r . β_1 is the effect of the i th individual's breeding value on its fitness (holding the effect of partner breeding value constant), and its inverse can be considered the cost to i of its own behavior, or

$\beta_1 = -c$. Likewise, β_2 is the effect of partner breeding value on i 's fitness (holding the effect of i 's breeding value constant), and can thus be considered the benefit to i given by the partner, or $\beta_2 = b$. And since the variance in p_i can never be negative, equation (B2.5) can be used to identify the conditions under which the population average breeding value will increase, $\Delta\bar{p} > 0$:

$$rb - c > 0. \quad (\text{B2.6})$$

This is Hamilton's rule in its "general" form (Birch, 2017; Gardner et al., 2011; Queller, 1992).

Note that equation (B2.6) assumes the neighbor modulated (also known as the direct fitness, personal fitness, or kin selection) interpretation of social interactions rather than the classic inclusive fitness interpretation: instead of taking the perspective of a focal actor whose actions affect its own fitness and the fitness of one or more recipients (the inclusive fitness approach), it takes the perspective of a focal recipient who is affected by a number of actors, including itself (the neighbor-modulated approach; Taylor et al., 2007). Nevertheless, it is easy enough to exchange the benefit received from partners, $b = \beta_2 = \beta_{w_i \hat{p}_i, p_i}$, for a benefit given to recipients, $b = \beta_2 = \beta_{\hat{w}_i p_i, \hat{p}_i}$, which allows for the inclusive fitness interpretation instead (Gardner et al., 2011).

The covariance term of equation (2) concerns all individuals in the population, but if each individual belongs to a discrete group, then it is also true that

$$\text{Cov}(w_i, z_i) = \text{E}[\text{Cov}(w_{jk}, z_{jk})] + \text{Cov}(W_k, Z_k). \quad (3)$$

That is, the covariance between relative fitness and trait values across individuals in the population, $\text{Cov}(w_i, z_i)$, can be decomposed into the expected covariance between individual fitness and individual trait values within groups, $\text{E}[\text{Cov}(w_{jk}, z_{jk})]$, and the covariance between average group fitness and average group trait values, $\text{Cov}(W_k, Z_k)$. Substituting equation (3) into equation (2) gives a partition that appears to separate lower-level and higher-level components of selection:

$$\Delta\bar{z} = \underbrace{\text{E}[\text{Cov}(w_{jk}, z_{jk})]}_{\text{within-group selection}} + \underbrace{\text{Cov}(W_k, Z_k)}_{\text{between-group selection}}. \quad (4)$$

In keeping with Price (1972) and Hamilton (1975), it is customary to interpret the two terms on the right-hand side of equation (4) as the effects of within-group selection and between-group selection, respectively. Thus, lower-level selection is said to act when $\text{E}[\text{Cov}(w_{jk}, z_{jk})] \neq 0$ and higher-level selection is said to act when $\text{Cov}(W_k, Z_k) \neq 0$. Of course, these two forces may work in opposing directions, as in the case of "weak" altruism, wherein cooperators have lower fitness than defectors within the same group, but individuals belonging to groups with more cooperators have higher fitness than individuals belonging to groups with fewer cooperators. Note also that this kind of multilevel partitioning can be extended further up or down the hierarchy.

Since the Price equation can accommodate genic, individual, and multilevel

viewpoints, and since it can reproduce and fortify independently derived results (such as Hamilton’s rule), it is not unreasonable to think that the various theories are only different perspectives on the same events, seen from different points within a hierarchy. There is a conciliatory message in this idea: if we move past the language barrier, perhaps we can see that the gene’s-eye, individual, and multilevel approaches are just different ways of conceptualizing the same thing, and it may even be useful to switch between them (Dugatkin & Reeve, 1994; Marshall, 2011; Panchanathan, 2011). For it is an empirical fact that when the average trait value changes from one perspective, it likewise changes from every perspective, making the various theories descriptively equivalent. But is this equivalence enough?

Unfortunately, it isn’t. While the Price equation certainly appears to partition components of evolutionary change, such as selection and transmission bias or within- and between-group selection, it does not do so cleanly. Notice, first, that no causal statements were made in the derivation of equation (1), other than that parents produced offspring. There was no link made between trait as cause and fitness as effect. Consequently, the covariance term of equation (1) can only speak to selection *of* traits, not selection *for* them, unless we make particular causal assumptions (Okasha & Otsuka, 2020; Sober, 1984). If we wish to understand adaptive design, theories derived directly from the Price equation offer little purchase, because they cannot distinguish cause from correlation.

Consider an example in which the Price multilevel partition ascribes the effect of an asocial, individual-level trait to group-level selection (Heisler & Damuth, 1987; Nunney, 1985; Okasha, 2004; Sober, 1984). In this scenario, individuals bear a trait z_i that only affects their own fitness w_i but, because the population has been subdivided into groups, equation (4) can nevertheless be applied. By chance alone, some groups will contain more individuals with high relative fitness than will other groups, and so the term $\text{Cov}(W_k, Z_k)$ will be nonzero, implying that selection between groups plays a role when it does not. Thus, even though both fitness and traits vary at the group level in the model, they are nothing more than byproducts of lower-level variance existing apart from the group structure.

One solution is to limit the definition of a group to circumstances in which individuals are engaged in social interactions (“trait groups” in the parlance of Sober & Wilson, 1998; D. S. Wilson, 1975). Under this definition, it would be impossible to entertain the idea of group-level selection acting on an asocial trait, because the Price multilevel partition would never be applied to such a case. This move can be helpful, but there remains the larger problem that the Price multilevel partition will always mistake individual-level byproducts for group effects (Okasha, 2004, 2006). For instance, in an effort to show that kin selection is a special case of group selection, D. S. Wilson and Sober (1994b) claim that interactions among genetic relatives cause group-level selection because fitness at the group level varies. Yet this group-level variance is attributable only to the actions of individuals—the groups themselves are not acting in any capacity. These sorts of fitness byproducts are precisely what concerned Williams (1966) when he cautioned us not to confuse individual-level effects with the fortuitous benefits they may entail when measured at the group level, and the Price multilevel partition does not have a clear defense against this.

A second concern with the Price equation is that it does not even cleanly partition

selection and transmission bias in the weaker “selection of” sense. Consider that both terms on the right-hand side of equation (1) contain fitness, w_i . Altering selection by changing the covariance between fitness and trait value is thus likely to alter transmission bias, too. This can be fixed by rewriting the Price equation as

$$\Delta\bar{z} = \text{Cov}(w_i, z'_i) + \text{E}(\Delta z), \quad (5)$$

which removes fitness from the expectation term (Okasha, 2006). However, in doing so, transmission bias has been moved into the covariance term, because offspring trait values z'_i may differ nonrandomly from parental trait values (Godfrey-Smith, 2007; Okasha & Otsuka, 2020). Framed either way, there is no clean separation of evolutionary processes.

Contextual Analysis

An alternative to the Price multilevel partition, popular among empiricists, is contextual analysis, which studies multilevel selection with the help of multiple regression techniques like those used in the derivation of the general form of Hamilton’s rule (Box 2; Goodnight, 2015). Contextual analysis assigns two trait values to each individual: the individual’s own value, z_{jk} , and a “contextual” group trait value, Z_k , which could be additive or non-additive (Heisler & Damuth, 1987; Okasha, 2004, 2006). It does this by predicting individual fitness from both individual and group trait values, holding the other constant:

$$w_{jk} = \alpha + \beta_3 z_{jk} + \beta_4 Z_k + \epsilon_{jk}, \quad (6)$$

where α is baseline fitness; $\beta_3 = \beta_{w_{jk}z_{jk}.Z_k}$ is the partial regression coefficient of individual fitness on individual trait value, adjusting for group trait value; $\beta_4 = \beta_{w_{jk}Z_k.z_{jk}}$ is the partial regression coefficient of individual fitness on group trait value, adjusting for individual trait value; and ϵ_{jk} is the residual term.

Although contextual analysis has its own origins, it can be integrated with the Price equation by substituting equation (6) into equation (2), on the understanding that the individual index i at the population level can be exchanged for the individual and group indices j and k at the group level (Frank, 2012; Gardner, 2017). This gives

$$\begin{aligned} \Delta\bar{z} &= \text{Cov}(\alpha + \beta_3 z_{jk} + \beta_4 Z_k + \epsilon_{jk}, z_{jk}) \\ &= \text{Cov}(\alpha, z_{jk}) + \beta_3 \text{Cov}(z_{jk}, z_{jk}) + \beta_4 \text{Cov}(Z_k, z_{jk}) + \text{Cov}(\epsilon_{jk}, z_{jk}). \end{aligned} \quad (7)$$

The first and last terms of the second line of equation (7) can be eliminated because α is a constant and ϵ_{jk} is uncorrelated with z_{jk} by design. Moreover, the covariance between a variable and itself is simply that variable’s variance, so $\text{Cov}(z_{jk}, z_{jk}) = \text{Var}(z_{jk})$. Likewise, $\text{Cov}(Z_k, z_{jk}) = \text{Var}(Z_k)$, because Z_k is simply the average of z_{jk} (Okasha, 2004). Equation (7) can thus be simplified to

$$\Delta\bar{z} = \underbrace{\beta_3 \text{Var}(z_{jk})}_{\text{within-group selection}} + \underbrace{\beta_4 \text{Var}(Z_k)}_{\text{between-group selection}}. \quad (8)$$

As with the Price multilevel partition, there again appears to be a separation of

components at two levels of social organization: assuming that trait values vary, within-group selection is said to act when $\beta_3 \neq 0$ and between-group selection is said to act when $\beta_4 \neq 0$. Comparing equations (4) and (8) shows how the Price multilevel partition and contextual analysis define the effects of selection in different ways (Okasha, 2004, 2006). The key difference is that the Price multilevel partition identifies any statistical association between group trait values and group fitness [$\text{Cov}(W_k, Z_k)$] as group-level selection whereas contextual analysis only does so after the effect of individual-level trait values have been removed [$\beta_4 \text{Var}(Z_k)$]. This means that when the contextual approach is applied to the case of an asocial trait, as in the example above, $\beta_4 = 0$ because the group trait value adds no new information beyond what is already known from the individual's own trait value, and so the method does not detect any group-level selection at work. It corrects for the individual-level byproduct error made by the Price multilevel partition.

Yet, contextual analysis falls short in other ways. First, it will identify an effect of group-level selection even when groups do not vary in fitness. The prototypical setting for this is a problem of “soft” selection or local competition: individuals compete for fitness within their groups, but each group has the same fitness, typically because the individuals are competing over shares of a fixed resource. Okasha (2006) offers an example by shifting down the hierarchy to a problem of meiotic drive, where the “individuals” are alleles and the “groups” are organisms. Consider a diploid population bearing copies of two alleles, A and B, at a single locus. For clarity, let us focus on absolute rather than relative fitness measures and change our notation, dropping the i , j , and k subscripts in favor of subscripts representing the organism's genotype (AA, AB, or BB) and the allele's genic value (A or B). All organisms have the same absolute fitness, denoted $Q_{AA} = Q_{AB} = Q_{BB}$, because resources are fixed. Moreover, segregation among heterozygotes is distorted in favor of the A allele, or $q_A > q_B$. Intuitively, selection at the organism level cannot act under these circumstances, as every organism performs the same. Nevertheless, genic fitness depends on an allele's trait value within their group—the A allele has increased fitness when paired with a B allele—which means that $\beta_4 \neq 0$ (Goodnight et al., 1992; Okasha, 2006). Thus, contextual analysis detects the action of selection at the organism level in a situation that does not seem to bear one of the prerequisites of selection: variance in fitness (Okasha, 2006).⁴

Second, contextual analysis can fail to detect the effect of higher-level selection when it is acting, as Okasha (2006) shows in a modification of the preceding example, in which both alleles and organisms face fitness consequences that exactly cancel out. Take again the above single-locus model, but now allow organismal absolute fitness to vary, such that $Q_{AA} = 16$, $Q_{AB} = 12$, and $Q_{BB} = 8$, while at the same time setting the fitness of the A allele to $q_A = 8$ and the fitness of the B allele to $q_B = 4$. There are thus two paths to

⁴ Goodnight (2015) has rejected the premise that there is no variance in group fitness under soft selection—again, a context whereby fitness varies at lower levels of organization but is nevertheless constant at higher levels. Rather, he argued that because $\beta_4 \neq 0$, higher-level selection must be acting to exactly cancel the effect of lower-level selection on the higher level. This may be true in some circumstances, but as we have seen from the description of the example above, what explains the lack of organismal fitness variance in this case is fixed resources; selection at the organism level plays no causal part in the process here. Thus, inferring the existence of a causal path from the results of a statistical model is a risky maneuver.

fitness: at the lower level, A alleles outperform B alleles; and at the higher level, organisms bearing A alleles outperform organisms bearing B alleles. An A allele partnered with another A allele produces the same number of copies as its partner, whereas an A allele partnered with a B allele produces 4 more copies than its partner, and therefore has a fitness advantage at the genic level. Conversely, an organism with an AB genotype produces 4 fewer offspring than an organism with an AA genotype, thereby completely offsetting the gene-level advantage of As paired with Bs. Despite this causal dependence of fitness on “group” context, contextual analysis would assign all of the change in trait values to selection at the gene level ($\beta_3 \neq 0$) and none of it to selection at the organism level ($\beta_4 = 0$), because an A allele has the same total fitness irrespective of the identity of its partner, and the same is true of a B allele (Okasha, 2006, 2016). From the perspective of contextual analysis, this cancellation effect makes it seem as if the fitness of the A allele does not depend on the group to which it belongs, but this is wrong: organisms in this example do in fact vary in fitness in a way that is consequential for the alleles they bear.

Okasha (2006, 2016) argues that the difference between these two problem cases—the first concerning competition among alleles and the second concerning competition among organisms—is, at its core, a causal question about which level of fitness is directly affected by the trait. In the first example, the genic trait directly affects the fitness of the alleles within the organism, and the organism’s fitness is unaffected; in the second example, the organismal trait additionally affects the fitness of the organisms directly, and the alleles come along for the ride. This explains why there is no single, universal statistical approach that will always square with our intuitions about selection and adaptation: for each of the many available approaches, there is always a case in which it will mistake a direct causal effect for an indirect one. Instead, what is needed is an extra-statistical method that formalizes preexisting knowledge of causal relations and thereby specifies a corresponding analytical approach (Okasha, 2016; see also Krupp, 2016; Logue & Krupp, 2016; Otsuka, 2019). That method is the graphical causal model, or causal graph.

The Causal Path

When the total evolutionary change is written using a statistical partition, as above, a natural explication of this notion suggests itself: the statistical associations between variables should reflect direct causal influences in the world. (Okasha, 2016, p. 443)

It is often said that the different approaches to social evolution are formally equivalent. This, however, implies that the only formalism that matters to the levels of selection debate is the statistical kind. While it is true that every approach can be connected to the Price equation and the outcome will always be the same, this only makes the various approaches descriptively equivalent (Otsuka, 2019). This sort of equivalence isn’t very satisfying because we know they are not causally so, and it is not obvious why we would choose different partitions unless they were meant to suggest different causal processes (Birch, 2017; Okasha, 2006, 2016). Helpfully, there is a mathematical formalism

that can be used to identify causal non-equivalence: the causal graph.⁵

Causal claims require causal assumptions (Otsuka, 2016a, 2019), and causal graphs permit this by visualizing the presumptive causal relationships generated by the variables in the model. They got their start with Wright (1920), who drew the first known path diagrams to specify genetic, environmental, and developmental effects on the expression of coat colors in guinea pigs. He then used these diagrams to show that the strength of causal effects could be inferred from them, essentially demonstrating a mathematical correspondence between causes in the diagram and correlations in the data. This allows users to estimate causal effects—structural parameters, as familiarly found in structural equation models—from statistical associations, on the assumption of linearity (Pearl & Mackenzie, 2018; Wright, 1920).

Today, the prevailing graphical tool in the causal inference literature is the causal directed acyclic graph (DAG), which we will use here. A DAG is a mathematical object that encodes qualitative causal assumptions, and so can inform theoretical and empirical models (Elwert, 2013). It is a nonparametric generalization of the path diagram, allowing it to express causal relations of any form—not just linear ones. Using prior knowledge of a biological system, it is possible to lay out the causal assumptions first and then use these assumptions to determine an appropriate method of analysis. This should reduce the chances of applying an inappropriate statistical model and making the wrong inferences. As Hernán and Robins (2020, p. 71) put it, “draw your assumptions before your conclusions.”

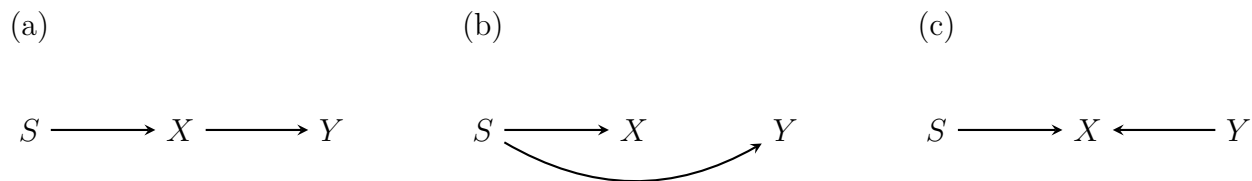
Draw Your Assumptions

I have been somewhat liberal in my use of “cause” up to this point. To make it more explicit, I mean it in the sense that, for a putative cause-effect relationship, had we made a precise change to the cause, the effect would be different. This is a counterfactual definition of causality: that the probability of an outcome differs between the scenario that actually occurred and a counterfactual scenario in which an element has been modified (Hernán & Robins, 2020; Rubin, 1974). For example, “had I taken the bus rather than walked to campus, I would have been on time for class” is a counterfactual claim that implies a causal effect of mode of transit on arrival time. Nothing about this implies that the cause of interest is the only cause, of course; I may also be late for class because I forgot to set an alarm and slept in.

To encode causal assumptions in a DAG, nodes or vertices represent random variables and any single arrow or directed edge running between a pair of nodes represents a direct causal effect, where the tail leaves the cause and the head enters the effect. For instance, in Figure 2a, the nodes represent variables S , X , and Y , and the arrows represent the statements “ S directly causes X ” ($S \rightarrow X$) and “ X directly causes Y ” ($X \rightarrow Y$). Note that this graph also shows an indirect effect of S on Y through its effect on X ($S \rightarrow X \rightarrow Y$); hence, X mediates the relationship between S and Y .

DAGs have rules (Elwert, 2013; Hernán & Robins, 2020; Kunicki et al., 2023; Pearl et al., 2016; Rohrer, 2018). First, as implied by the terms “directed” and “acyclic”,

⁵ As an alternative to causal graphs, the potential outcomes framework could also be used (Imbens & Rubin, 2015). However, it is hard to argue with the value of graphical methods to lay out the core levels of selection problem of transparently specifying trait-fitness relationships.

**Figure 2**

Three elementary directed acyclic graph structures. A DAG can be composed of: (a) shorter and longer chains, such as $S \rightarrow X$, $X \rightarrow Y$, and $S \rightarrow X \rightarrow Y$, which transmit causal information; (b) forks, such as $X \leftarrow S \rightarrow Y$, which transmit spurious associations between confounded variables; and (c) colliders, such as $S \rightarrow X \leftarrow Y$, which block associations between separated variables.

causality only moves forward from cause to effect, in the direction of the arrow, and variables cannot cause themselves, whether directly or indirectly. Second, DAGs should include all common causes of the other variables in the DAG. For example, if there is a variable U that causes both S and X in any of the graphs in Figure 2, then it should be included in those graphs—even if that variable is unknown or has not been measured. Third, an arrow does not imply any specific function, sign, or strength of effect. Fourth, and relatedly, the existence of two or more arrows coming into a single variable (e.g. $S \rightarrow X \leftarrow Y$ in Fig. 2c) does not specify how the different causes interact. Finally, the absence of an arrow connecting two variables is a stronger assumption than the presence of an arrow: a missing arrow signifies a high degree of confidence on the researcher’s part that there is no direct causal relationship between these variables. In Figure 2b, then, the absence of an arrow between X and Y conveys the message that there is no causal relationship between these variables.

There are three ways in which variables can be configured in a DAG: chains, forks, and colliders (Fig. 2; Elwert, 2013; Kunicki et al., 2023; Pearl et al., 2016; Rohrer, 2018). A chain is a causal path in which all effects flow in the same direction, as in $S \rightarrow X$ and $S \rightarrow X \rightarrow Y$ (Fig. 2a). A fork is a noncausal structure that confounds the relationship between two variables as a function of a third variable—their common cause—as in $X \leftarrow S \rightarrow Y$, where S generates a statistical association between X and Y even though neither variable causes the other (Fig. 2b). Finally, a collider or inverted fork is a structure that “blocks” or “screens off” the association between two variables that independently cause a third variable, as in $S \rightarrow X \leftarrow Y$ (Fig. 2c). In this case, we should not expect any association between S and Y , because they are independent of each other, even though they both affect X . Crucially, adjusting for, controlling, or conditioning on a variable can dramatically affect the statistical associations in each of these configurations: adjusting for a mediating variable in a chain (e.g. X in Fig. 2a) can block the association between the variables on either side of the mediator; adjusting for a common cause in a fork (e.g. S in Fig. 2b) can block the spurious association between the variables affected by that cause; and adjusting for a collider (e.g. X in Fig. 2c) can create a spurious association between the variables causing the collider.

With its formalization of cause and effect, a DAG can be used to distinguish

selection *for* a trait from selection *of* a trait. For example, recall Sober’s (1984) example of the two-tiered sorting toy containing larger white and smaller gray marbles (Fig. 1). For the sake of simplicity, imagine that, during the design process, the toy company’s engineers were each given a divider with holes of fixed size and then decided the size and color of the marbles. These decisions weren’t random: the engineers may have matched the marble colors to the marble sizes, and the marbles were sized to either be larger or smaller than the holes. For the marbles to sort between tiers after the toy is built, the toy must also be turned over (perhaps a few times). The DAG in Figure 3 formally represents the causal assumptions of the system just described: the engineer causes marble color and marble size; marble size, hole size, and turning of the toy interact to cause sorting of the marbles.

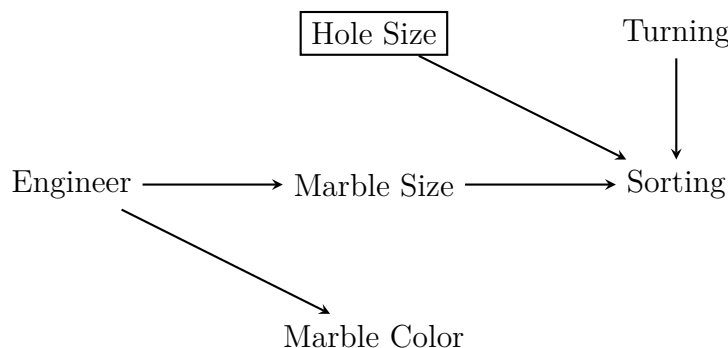


Figure 3

A directed acyclic graph of Sober’s (1984) marble sorting toy. Hole size is fixed, and is represented so here by enclosing it in a box.

Figure 3 makes it clear that the direct effects $\text{Marble Size} \rightarrow \text{Sorting}$ and $\text{Turning} \rightarrow \text{Sorting}$ reflect selection for marble size and turning, respectively: sorting happens because the marbles are small enough to pass through the divider, and because turning the toy increases the chances that a marble will roll into a hole. There is no selection on hole size because it is fixed, which is denoted in the graph by enclosing it in a box. At the same time, the fork $\text{Marble Color} \leftarrow \text{Engineer} \rightarrow \text{Marble Size} \rightarrow \text{Sorting}$ that generates the association between marble color and sorting reflects only selection of, not for, color. Color has no effect whatsoever on sorting in the DAG, but marbles will nevertheless be sorted by color because the engineer confounded color with size, and size causes sorting.

But what about the indirect effect of the engineer on sorting? The chain $\text{Engineer} \rightarrow \text{Marble Size} \rightarrow \text{Sorting}$ clearly shows that the engineer is a cause of the marbles being sorted over the tiers of the toy. Yet, if we already know the size of a given marble, then we also know whether it will pass through the divider, irrespective of which engineer produced the marble. We saw this with the example of the thermostat in a previous section, as well: the thermostat turns on in response to the temperature at the sensor; it is ignorant of the causes of temperature change. More generally, if we know all of the direct causes of a variable X , then its indirect causes add no new information. This is a causal variant of the Markov property: adjusting for all direct causes of X renders all indirect causes of X irrelevant.⁶ In terms of the toy, then, selection is not acting on the engineers, but on

⁶ Somewhat more strictly, the causal Markov property states that if we adjust for or condition on all

marble size. Likewise, in terms of natural selection, what matters is the point of interface between trait and fitness—the direct cause-effect relationship. Selection is not acting on distal causes of the trait, but on the trait itself (Brandon, 1982; Gould, 1980; Krupp, 2013; Mayr, 1963; Okasha, 2016).

Identifying the Levels of Selection

Okasha (2016) has used the causal graph method to formalize the argument that the level of selection is determined by the level of fitness directly affected by the trait. We will follow this method here, with two changes described below. For convenience, we can call the level of fitness that is directly affected by the trait the “primary” fitness effect. If a trait directly affects individual fitness, then individual fitness is primary and selection is therefore operating at the individual level. Conversely, if a trait directly affects group fitness, then group fitness is primary and selection is operating at the group level. Dropping the subscripts of the previous section for ease of reading, we can make this argument precise with the language of a DAG: $z \rightarrow w$ is a case of individual-level selection, because the individual trait directly affects individual fitness, and $Z \rightarrow W$ is a case of group-level selection, because the group trait directly affects group fitness (Barclay & Krupp, 2016; Krupp, 2016; Okasha, 2016).⁷ This marks our first point of departure from Okasha (2016): although the graphical notation of primary fitness effects remains the same, Okasha used path diagrams whereas we will use DAGs.

How can we know when a particular level of fitness is primary? In essence, the distinction amounts to whether, with respect to the trait under selection, the lower-level units have routes to survive and reproduce *within* or *outside* of the higher-level unit versus routes to survive and reproduce *through* the higher-level unit’s own survival and reproduction (Okasha, 2016). If the trait under selection is meiotic drive, alleles can distort their frequencies within their bearers, and not merely because of the actions of their bearers. However, if the trait under selection is the ability to find food, alleles receive fitness by virtue of the success or failure of the organisms bearing them to forage. This matters because, along any causal path where group fitness is primary, effects are experienced in exactly the same way by all lower-level units in the group; their fitness interests are simultaneously aligned. To expand on an analogy given by Barclay and Krupp (2016), the fitness of each member of a ship’s crew can vary within the ship—some sailors might thrive, say, whereas others might die—but their fitnesses will also depend, in precisely the same way, on whether the ship itself makes it back to harbor or goes down at sea.

“parents” of X —that is, the direct causes of X —then X is independent of all other variables in the graph except its effects or “descendants” (Hausman & Woodward, 1999). This idea is helpfully captured by the phrase “given the present, the future is independent of the past.”

⁷ In most cases, z will entail w and Z will entail W , but it is possible that there are cases in which individual traits directly cause group fitness ($z \rightarrow W$) and cases in which group traits directly cause individual fitness ($Z \rightarrow w$). In support of Okasha’s (2016) argument that the levels of selection depend on the level of fitness—rather than on the level of the trait—it should be pointed out that the optimal individual trait value maximizes group fitness when $z \rightarrow W$, and so benefits the group, whereas the optimal group trait value maximizes individual fitness when $Z \rightarrow w$, and so benefits individuals.

In a practical sense, the proposition above hews closely enough to Dawkins' (1976/2006) "vehicle" concept that I will also make use of the term, though I mean something more precise: higher-level selection occurs because a vehicle inextricably binds the causes and effects of the group members together along at least one path between the lower-level cause and the lower-level effect. It is thus the fitness of the vehicle that is primary along this path. The concept does not require complexity of design. It makes no difference whether the vehicle is a well-integrated organism or a mass of undifferentiated cells; vehicular complexity is something that evolves. What does matter is that the lower-level units are united in cause and effect.⁸

This marks our second point of departure from Okasha (2016): rather than assign traits and fitness to the group, such as a collective of alleles, they are assigned to the vehicle that contains the group. In the previous section, a group trait was defined as the average trait value of the individuals in the group and group fitness was defined as the average fitness of the individuals in the group. The latter definition is of the collective fitness₁ variety, being an aggregate measure of the ability of lower-level units to survive and make more lower-level units (Heisler & Damuth, 1987; Okasha, 2006). By the same token, we might say that the former definition is of the collective trait₁ variety, being an aggregate measure of the trait value of the lower-level units. Collective trait₁ and collective fitness₁ measures work naturally with the statistical approaches that are the stock-in-trade of the levels of selection debate. From a causal perspective, however, they are nothing but trouble. The problem is that, so defined, group traits and group fitness are only statistical properties, not inherently "real" objects in their own right; as we have already seen, a group of individuals can be assigned an average group trait and an average group fitness even if they are a group in name only. All we know of these properties is that they are arithmetic composites of, and fully determined by, individual traits and individual fitness (their "parents" in the graph; Berrie et al., in press), and so they may be entirely ephemeral. Without more information, it is unclear whether they can truly be the cause of anything.⁹

In contrast, a vehicle is here defined causally, and so is a very real thing. The fitness of a vehicle seems to be of the collective fitness₂ variety, as it is a measure of the united group's ability to survive and make new groups (Heisler & Damuth, 1987; Okasha, 2006).

⁸ There is a certain irony in making vehicular fitness the centerpiece of the levels of selection problem, as Dawkins (1994, p. 617) has stated "I coined the 'vehicle' not to praise it but to bury it." He went on to say that "vehicles often turn out to be the objects that we recognise as organisms, but this did not have to be so. It is not part of the definition of a vehicle. There did not have to be any vehicles at all. Darwinism can work on replicators whose phenotypic effects (interactors) are too diffuse, too multilevelled, too incoherent to deserve the accolade of vehicle." This is almost certainly true, and if it were the case that no vehicle had ever evolved, then there would indeed be no higher level of selection than the gene. But organisms have evolved, and other vehicles have, too.

⁹ A deterministic variable such as an average group trait is a particularly sharp example of the part-whole "supervenience" problem, as it is known to philosophers (Kim, 1998). Put simply, it could be argued that since lower-level properties (individual traits and fitness) determine higher-level properties (average group traits and fitness), causation must reside at the lower level. However, as Okasha (2006) points out, the question at hand is not whether the whole can be explained by its parts; it is whether there is *selection* on the whole rather than on the parts. I submit that we cannot answer this question unless we already know whether the whole actually exists in the vehicular sense, and therefore whether it can act as a causal force.

Likewise, we could say that a vehicle's trait value may be of the collective trait₂ variety, as it is again a measure of the united group. These are only tentative characterizations, however: we have not been given the causal details to know whether, when the vehicle reproduces, it makes new groups or simply new individuals; if the latter, neither collective fitness₁ nor collective fitness₂ apply. In any case, assigning traits and fitness to the vehicle makes them stricter forms of collective traits and collective fitness, but moving out of the statistical realm and into the causal realm compels us to reevaluate the biological meaning of variables such as group traits and group fitness. The average trait value and average fitness of a group of alleles are qualitatively different sorts of things than the trait value and fitness of the vehicle bearing those alleles, even when there is a quantitative correspondence between the two. For instance, in the context of alleles and organisms, the average amount of a particular protein produced by a group of alleles would be a collective trait₁ measure and the average number of alleles produced would be a collective fitness₁ measure, whereas the visual acuity of an organism would be a vehicular trait and the number of offspring an organism produced would be a measure of vehicular fitness. The vehicle's traits and fitness are causal; group means are not.

With this in mind, let us now make the causal graph method concrete by considering again Okasha's (2006) soft selection problem of the previous section. Partners are pairs of alleles inhabiting a single locus within a diploid organism, and we will fasten attention on a focal allele. Alleles are drawn from the population—possibly non-randomly—and we can denote the probability that the average pair are copies of the focal allele identical by descent as G , also known as the coefficient of consanguinity. Let us assign the focal allele a genic value g and assign its partner allele a genic value \hat{g} . These genic values respectively give rise to trait values z and \hat{z} , which directly affect each allele's ability to take a spot in a fixed number of gametes, w and \hat{w} . Finally, because the number of gametes over which the alleles compete are fixed, organisms do not vary in their fitness. The causal assumptions in this case are encoded in the DAG in Figure 4.

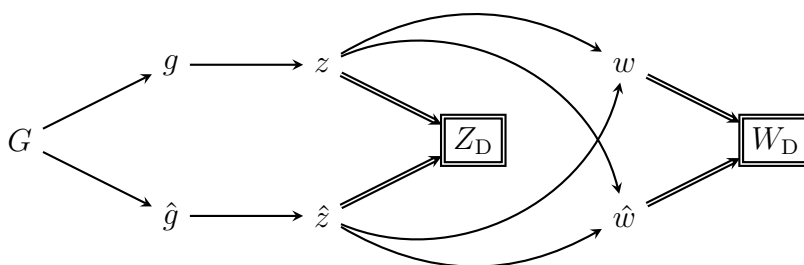


Figure 4

A directed acyclic graph of lower-level selection. Each of the direct effects of traits on fitness occur at the lower level: $z \rightarrow w$, $z \rightarrow \hat{w}$, $\hat{z} \rightarrow \hat{w}$, and $\hat{z} \rightarrow w$. Double-lined arrows denote deterministic relationships and doubly enclosed nodes denote deterministic variables.

This example contains no organismal trait, and organismal fitness is fixed, so there is no reason to include either in the graph. Nevertheless, Figure 4 does include the average group trait value and average group fitness of the previous section, only because they are instructive here. To distinguish these deterministic variables from their probabilistic

counterparts, they are denoted as Z_D and W_D and enclosed in double boxes, with double-lined arrows indicating deterministic relationships (K. F. Arnold et al., 2020). Despite the presence of these variables in the graph, Figure 4 makes it clear that this is a gene-level selection problem, because it is genic fitness (w and \hat{w}) that is primary along every causal path: $z \rightarrow w$, $z \rightarrow \hat{w}$, $\hat{z} \rightarrow \hat{w}$, and $\hat{z} \rightarrow w$. We have not used statistics to tell us this; we already know the level of selection by reading it off the graph, which was built from our understanding of the biological system.

Indeed, inspection of Figure 4 shows exactly why it would be a mistake to infer the level of selection from a statistical model that is indifferent to the causal assumptions built into the graph: even though there is no causal path between Z_D and W_D , they share multiple common causes—including z and \hat{z} —via multiple different forks, and are therefore thoroughly confounded. Contextual analysis introduces bias when measuring the gene-level effect, because Z_D is a collider and adjusting for it in the calculation of β_3 opens up a new non-causal path ($z \rightarrow Z_D \leftarrow \hat{z} \rightarrow w$). Moreover, because \hat{z} confounds the relationship between Z_D and w via the fork $Z_D \leftarrow \hat{z} \rightarrow w$, contextual analysis incorrectly attributes the gene-level effect of \hat{z} on w to the group-level trait Z_D in the calculation of β_4 , which is why it detects group-level selection even when group fitness does not vary. Conversely, the Price multilevel partition only gets this case right because group fitness does not vary, and so W_D is constant. Had it instead been allowed to vary, the Price multilevel partition would incorrectly attribute lower-level effects to higher-level selection, as noted above in Wilson and Sober’s (1994b) argument that variance in fitness among sibling groups shows that kin selection is a form of group-level selection.

Next, consider a problem in which selection acts only at the level of the organism. Partners are again pairs of alleles with genic values g and \hat{g} at a single locus, but the trait values they generate, z and \hat{z} , combine to create an organismal trait, Z_V . This in turn directly affects the organism’s fitness, W_V . Here, the V subscript indicates that Z_V and W_V are vehicular properties—in this case, the organism’s trait and the organism’s survival and reproduction—and the latter causes genic fitnesses w and \hat{w} as a result of organismal reproduction. The DAG in Figure 5 encodes these causal assumptions, showing that organismal fitness is primary along the only causal path to fitness ($Z_V \rightarrow W_V$) and establishing this as a problem of organism-level selection. Importantly, a statistical model that takes only a lower-level approach, such as the gene’s-eye view, would be unable to recognize the effect of higher-level selection at work here, because it is not informed by the causal assumptions identified in the graph.

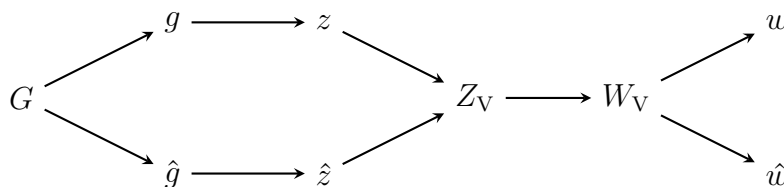


Figure 5

A directed acyclic graph of higher-level selection. The only direct effect of a trait on fitness occurs at the higher level: $Z_V \rightarrow W_V$.

Now consider Okasha’s (2006) example of the previous section in which selection operates at both genic and organismal levels. Let alleles produce traits that bias segregation within the organism and let these traits further produce an organismal trait that directly affects organismal fitness. These two paths combine the lower- and higher-level causal processes of Figures 4 and 5 to give the DAG in Figure 6, which shows this to be a problem of multilevel selection: genic fitness is primary along four causal paths ($z \rightarrow w$, $z \rightarrow \hat{w}$, $\hat{z} \rightarrow \hat{w}$, and $\hat{z} \rightarrow w$) and organismal fitness is primary along a fifth ($Z_V \rightarrow W_V$). Such a problem demands a multilevel statistical model that properly accounts for each of the different causal and non-causal paths. However, neither the Price multilevel partition nor contextual analysis would be appropriate, for the reasons already discussed.

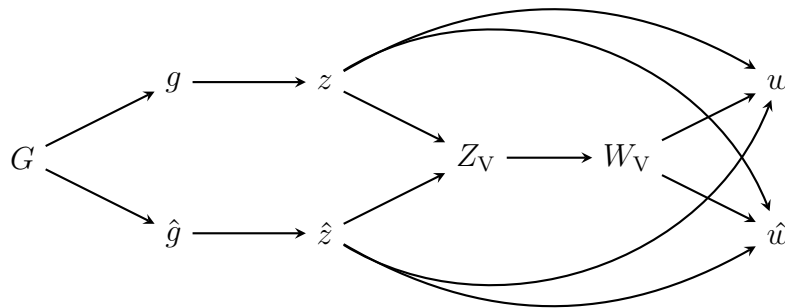


Figure 6

A directed acyclic graph of multilevel selection. Direct effects of traits on fitness occur at both lower and higher levels: $z \rightarrow w$, $z \rightarrow \hat{w}$, $\hat{z} \rightarrow \hat{w}$, $\hat{z} \rightarrow w$, and $Z_V \rightarrow W_V$.

Finally, consider two additional scenarios that conform to the causal structure of the DAG in Figure 4, and which can help to address the relevance of “emergent” group properties in determining the level of selection (Krupp, 2016; Logue & Krupp, 2016). In the first scenario, let us return once again to the same soft selection case that initially inspired Figure 4, but now assume that alleles produce one of two trait values, A and B, which distort segregation. If both alleles produce the same trait value (i.e. both A or both B), then both receive a payoff of 1; however, if the alleles produce different trait values, then the A allele receives a payoff of 2 and the B allele receives a payoff of 0. This kind of interaction is additive but frequency dependent, whereby payoffs depend on how common a given trait value is in the population. Nevertheless, nothing about the DAG in Figure 4 needs to change to accommodate this particular payoff structure: w and \hat{w} remain direct effects of the interaction between z and \hat{z} , and selection is therefore acting at the lower level.

In the second scenario, let us now shift the frame up from alleles to organisms and remove the soft selection constraint on the deterministic fitness W_D of groups. Here, partners both producing A trait values each receive a payoff of 2, partners both producing B trait values each receive a payoff of 0, and, in mixed pairs, the A partner receives -1 while the B partner receives 2. Hence, the payoffs are synergistic and frequency-dependent, and W_D now varies. And yet, selection still only acts at the lower level: nothing about this scenario requires any change to the causal assumptions laid out in Figure 4, and so

organismal fitness remains primary. Thus, despite the existence of emergent properties in both scenarios, a lower-level model is causally apt whereas a higher-level model is not.

The Place Where the Paths Meet

No causal assumption in, no evolutionary prediction out. (Otsuka, 2019, p. 45)

Travelers have typically made their way through the levels of selection either by the discursive or the statistical path. The former is built on foundational claims, including the proposition that natural selection is a causal process. Most can agree on this much, but none of the arguments that follow can be made verbally precise enough to come to a definitive answer. The latter path recognizes precision as its greatest strength and reveals a kind of formal unity among the different perspectives, in which every statistical partition can be derived from or integrated with the Price equation. Yet, if these partitions are meant to have any meaning, then each fails under certain conditions. The result is a set of extremely general mathematical formulations that describe, but do not predict or explain, evolutionary change.

More recently, a third path has opened up. Marrying causality to formality, this path begins by using prior knowledge of the system under study to encode qualitative assumptions about causal processes in a graph. The next steps are to use the graph to determine the direct effects of traits on fitness. And the path ends at a place where the other two paths meet. In one direction are many of the causal intuitions of the discursive path, now made precise. In the other direction are many of the techniques of the statistical path, now suitably tailored to identify the causal effects of the problem case. With an understanding of the rules of a DAG, both intuition and technique can be read right off the graph.

For instance, the graphs in Figures 4–6 relate the conditions needed to identify causal effects and the consequences of applying the wrong statistical model. It does not make sense to measure the effect of a group trait on group fitness in the case of Figure 4, because no such effect exists, and the deterministic variables Z_D and W_D are only spuriously connected—this DAG is a visual guide to the byproduct problem Williams (1966) warned of. Conversely, it does make sense to study this effect in the cases of Figures 5 and 6, because there is a causal path between the group trait and group fitness in these models. Moreover, if the aim is to measure the effect of the lower-level trait z on lower-level fitness w —a reasonable empirical question, even if z is not always the target of selection—it would be better to adjust for \hat{z} in each of these three cases than to adjust for the higher-level trait Z_D or Z_V . In the DAG of Figure 4, this is because Z_D is a collider, so adjusting for it can open up non-causal paths whereas adjusting for \hat{z} does not. In the DAG of Figure 5, this is because Z_V fully mediates the effect of z on w , so adjusting for Z_V blocks the effect of z whereas adjusting for \hat{z} does not. And in the DAG of Figure 6, this is because adjusting for Z_V both blocks a legitimate path and opens an illegitimate one.

The graphs in Figures 4–6 also each relate a basic causal chain of development and selection whereby genes cause traits and traits cause fitness. These same chains show that selection acts on traits: for example, adjusting for the lower-level trait z in the chain $g \rightarrow z \rightarrow w$ renders the indirect effects of genes on fitness moot, and adjusting for the higher-level trait Z_V in the chain $g \rightarrow z \rightarrow Z_V \rightarrow W_V$ renders the indirect effects of both

lower-level traits and genes on fitness moot. This has several consequences. First, as has been suggested before, selection only “sees” traits, not genes (e.g. Brandon, 1982; Gould, 1980). Genes are of course part of the larger causal story, but they are not enough to provide a complete account of adaptation (Krupp, 2013). Second, as Okasha (2016) argues, the direct effect of traits on fitness implies that selection at different levels of social organization can occur. Specifically, selection at level X comes about when traits directly cause fitness at level X (e.g. $z \rightarrow w$ or $Z_V \rightarrow W_V$). This is how we identify the levels of selection.

Replicators and Vehicles Revisited

It is possible to shed more light on the greater evolutionary process by connecting the parent generation to the offspring generation in a graph. For instance, if we let Ω be the frequency of copies of the focal allele in the parental population, let ω' and $\hat{\omega}'$ be, respectively, the number of copies of the focal allele derived from the focal allele and its partner in the offspring population, and let Ω' be the frequency of copies of the focal allele in the offspring population, then we can modify the higher-level selection graph of Figure 5 to produce the graph in Figure 7. This expanded DAG shows the causal process of evolutionary change of the lower-level units over one generation, from Ω to Ω' . Within this larger process, we can also see subprocesses of genetic relatedness or assortment ($g \leftarrow G \rightarrow \hat{g}$), trait development ($g \rightarrow z \rightarrow Z_V \leftarrow \hat{z} \leftarrow \hat{g}$), selection ($Z_V \rightarrow W_V$), and replication ($g \rightarrow \omega' \leftarrow w$ and $\hat{g} \rightarrow \hat{\omega}' \leftarrow \hat{w}$), which is an interaction between the genic value of the template being used to make copies and the number of copies being made.

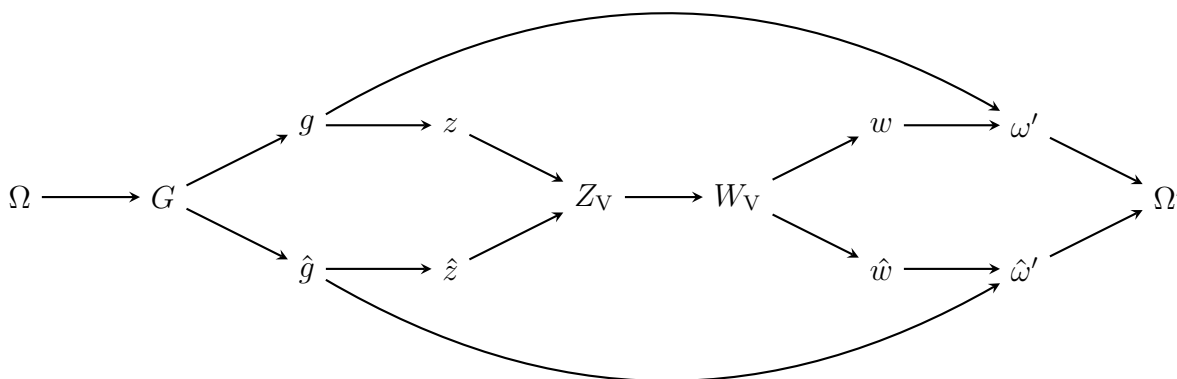


Figure 7

A directed acyclic graph of evolutionary change via higher-level selection. Although there is an effect of genic value on the frequency of copies of the focal allele in the offspring population ($g \rightarrow \omega'$ and $\hat{g} \rightarrow \hat{\omega}'$), selection nevertheless occurs only at the higher level ($Z_V \rightarrow W_V$) in this graph.

Although Dawkins’ (1982, 1976/2006) “replicator” and “vehicle” concepts rarely feature in formal models of evolutionary change, the causal structure of Figure 7 puts them on solid footing, though their meanings are refined. To adapt a definition from Hull (1981, p. 33), a replicator is something that passes on its value directly in replication. In the

current context, the replicator is a gene that passes on its genic value, because it serves as a template for the production of copies; this is the effect of g on ω' and of \hat{g} on $\hat{\omega}'$. In principle, however, a replicator can be anything that directly causes the value of the replicates to match its own value, implying the possibility of entities like epigenetic replicators and cultural replicators (Bonduriansky & Day, 2018). A replicator is not required for selection, though it may be important in the process of adaptation (Godfrey-Smith, 2009). Moreover, if a replicator is involved in the evolutionary process, it does not need to be responsible for the number of replicates being produced (Hull, 1981). To define a replicator, all that matters is that it directly causes the replicate to assume its value.

Conversely, a vehicle is something that brings about a unity of cause and a unity of effect by tethering lower-level units together: all of the units it contains are parties to the trait, even if they are not themselves causes of that trait; and all are equally affected by its consequences, on average, when it directly operates on fitness. This is not quite Dawkins' (1982) own definition of a vehicle, nor is it Hull's (1981) definition of an "interactor," but it is very close. The key difference is that when a vehicle acts on the world, it effectively compels its lower-level constituents to act together (a trait vehicle, Z_V) and to gain or lose fitness together (a fitness vehicle, W_V). Defined in this way, a vehicle is more than a group: it has its own traits and its own fitness, and by virtue of its existence, it represses competition among its contents via at least one causal route, aligning their fitness interests (Frank, 2003). These distinctions are vital to the causal position. If the groups under study do not themselves have traits, survive, or reproduce in any actual sense, then it stands to reason that it is the individual group members that are directly acting, surviving, and reproducing, and the DAG should be drawn as such. Hence, a group trait cannot be a cause and group fitness cannot be an effect—let alone a primary one—if they are not vehicular properties as defined above.

This puts the collective trait₁ and fitness₁ concepts in a tough spot, as they are rarely causally apt, and this may sometimes be true of collective trait₂ and fitness₂ as well. But that does not mean we cannot follow the success of individual genes. As the DAG in Figure 7 shows, it is simple enough to track an individual allele both before and after the vehicle bearing it has acted on fitness. Take the social amoeba *Dictyostelium discoideum* which, when starved, aggregates in large numbers to form a multicellular "slug" that travels to a new location and develops into a fruiting body composed of spores and a stalk. The slug is a vehicle that benefits its cellular cargo equally in several ways: its motility takes them to better environments, its sheath protects them from predation, and its stalk keeps the spores off the ground and helps with their dispersal (Medina et al., 2019). Each of these benefits entails identical expected fitness gains for the individual cells, but they are first accrued by the slug, whose traits affect its own survival and reproduction. While it is true that a slug does not descend from a clear slug lineage (Okasha, 2006), the causal approach does not require one. If a slug's survival and reproduction does not qualify as either collective fitness₁ or collective fitness₂, because groups of amoebae do not produce groups of amoebae and slugs do not produce slugs, then perhaps we are instead dealing with what might be called collective fitness_{1.5}, the survival and reproduction of higher-level units that make new lower-level ones. From a causal perspective, the collective fitness terminology does not really matter, because the goal is not to shoehorn a causal model into

an already-existing statistical model, but to generate a new statistical model that respects the assumptions of the causal model.

The vehicle and its traits are emergent properties of lower-level parts, and there have been numerous attempts to operationalize the levels of selection as depending on emergence of one kind or another (e.g. Gould & Lloyd, 1999; Smaldino, 2014; Vrba, 1989). As we have seen, however, emergence does not inherently change the causal structure of a DAG, because DAGs are nonparametric representations of cause-effect relations; frequency-dependent and synergistic payoffs can be accommodated without modifying the graph. Likewise, what are often characterized as emergent effects can be represented either by a higher-level trait borne by a vehicle (as in $Z_V \rightarrow W_V$) or as the interaction of lower-level traits borne by separate lower-level units (as in $z \rightarrow w \leftarrow \hat{z}$). The choice of representation is not a matter of taste, but again depends on whether there is a vehicle, as defined with respect to cause and effect, and whether the trait directly causes fitness at that level. In other words, what makes a vehicle special is not that it is emergent, but that it is unifying.

Still, the causal basis of replicators and vehicles is not a vindication of the broader agenda of the gene’s-eye view. The replicator is rarely a direct cause of selection, and cannot be thought of as the “true” object of selection without mistaking the agential metaphor of the selfish gene for something more literal. And thanks to its coordination of fitness interests, the vehicle turns out not to be disposable, as Dawkins (1994) would have it, but indispensable to the levels of selection problem: whenever lower-level units obtain fitness through the survival and reproduction of the vehicle, it is the vehicle’s traits that are selected for and the vehicle’s fitness that is primary. This makes the level of adaptation dependent on the level of selection (Gardner & Grafen, 2009; Okasha & Paternotte, 2012; Williams, 1966).

Specifically, lower-level adaptation occurs when there is selection on lower-level traits, leading to their optimization, as is the case along the paths $z \rightarrow w$ and $\hat{z} \rightarrow \hat{w}$ in Figure 4. These traits should appear designed to serve the interests of the lower-level unit that caused it. Similarly, higher-level adaptation occurs when there is selection on higher-level traits, leading to their optimization, as is the case along the path $Z_V \rightarrow W_V$ in Figure 5. These traits should appear designed to serve the interests of the vehicle and therefore all of the lower-level units contained within. However, adaptation under multilevel selection, as depicted by Figure 6, is less straightforward. If the lower-level optimum differs from the corresponding higher-level optimum, there is no reason to expect adaptation at either level (Gardner & Grafen, 2009; Okasha & Paternotte, 2012). Even if both levels pull the population in the same direction at first, the lower- and higher-level trait values will eventually evolve to be a compromise between the optima at the two levels. However, if the difference between them is reasonably small or if the effect of selection at one level is much weaker than the other—both subjective judgments, to be sure—then adaptation can arise at either level (Gardner & Grafen, 2009).

Draw Your Conclusions

Answers to questions about the levels of selection cannot be found in the data alone, as is widely assumed. They will always start with the causal assumptions. With the

causal graph approach, the levels of selection problem thus becomes an exercise in drawing. And when this is done correctly, it is possible to read both the levels of selection and their associated analyses straight from the graph.

With respect to selection, the assumptions that matter most pertain to the point of interface between traits and fitness. The trick is to identify which level of fitness is primary along each causal path (Okasha, 2016): selection at a given level occurs when traits directly cause fitness at that level. It is not an especially complicated task to identify traits and fitness at the gene level. A genic trait is a direct product, like a protein, and genic fitness is typically the number of copies of the gene being produced. Gene-level selection takes place, then, when a gene product directly increases the number of copies of the gene. At higher levels, however, things can get harder. Most would have no trouble accepting selection at the level of the organism, but what about selection at the level of chromosomes, cells, groups of organisms, or species?

Arguably, our comfort with the idea of selection at the level of the organism reflects causal intuitions of what an organism is: a causal force that interacts with the world and has its own fitness. From a genetic perspective, an organism is a vehicle, in that its lower-level passengers together bear the same cause and feel the same effect along a causal pathway. That is, the vehicle imposes an expected “cause and effect homogeneity” on the lower-level units, even if those units can also take other paths to fitness while still riding in the vehicle. By extension, to decide whether there is selection at a given level, one must first be able to locate the vehicle as defined here—if there is one at all—and determine its involvement in the causal process. If such a vehicle does exist, and it does bear associated traits that cause fitness, then it should be added to the graph; otherwise, it can be left out. The vehicle is therefore not a basic element of the evolutionary process, but selection cannot operate at higher levels without it.

Vehicles don’t need to be complex things, as previously noted. A yeast floc is little more than a clump of cells that adhere to one another when exposed to environmental stress—protecting the interior cells at the expense of the exterior ones (Smukalla et al., 2008). A floc’s size and shape are some of its traits and the proportion of surviving cells is a measure of its fitness. But vehicles don’t need to be living things, either. Cages have been used in experiments as vehicles for groups of chickens, where stock for subsequent generations are chosen on the basis of total egg production (Muir, 1996). Hence, there may be many ways to impose vehicular structure on individuals, including cultural innovations.

Others have also pointed to the higher-level unit as being critical to the levels of selection problem, for similar reasons. Hull (1981, p. 33) conceives of an interactor as a holistic cause of fitness, “directly interacting as cohesive wholes”. Conversely, D. S. Wilson and Sober (1994b, p. 591) conceive of it as a holistic effect, stating that the “essence of the vehicle concept is *shared fate*”, though they define this statistically (using a collective fitness₁ measure) rather than causally. But the causal graph formalization offers a more precise rationale. It suggests that it is unity of cause and unity of effect, along at least one causal path, that defines the vehicle for our purposes. This cannot be determined through inspection of the data, but by making causal assumptions based on an understanding of the system.

In sum, committing to the causal graph approach forces us to think critically about the assumptions we must make to answer empirical questions about natural selection. In so

doing, it makes a number of things clear. First, “assumption-free” models may be useful organizing frameworks, but they cannot be truly general explanations. The gene’s-eye view, inclusive fitness, and multilevel selection are not competing approaches, nor are they different ways of looking at the same thing. They are different frameworks appropriate to describing different causal processes, occurring at different levels of selection. Second, the average trait values and fitness of a group of lower-level units are not, in and of themselves, causal forces. They are deterministic variables, useful for statistical modelling but not for causal modelling. Third, replicators and vehicles appear in the evolutionary process, but their effects are not entirely as advertised. Replicators are often far removed from the trait-fitness interface, challenging the notion that they are the real targets of selection, and higher-level vehicles unite their lower-level units in both cause and effect, challenging the notion that vehicles are peripheral to selection. Finally, the levels of selection depend on the primary level of fitness, as Okasha (2016) argues, because it is at this level that trait values are selected and therefore optimized. Adaptation at different levels reflects different causal forces, and therefore requires different explanations.

References

- Ågren, J. A. (2021). *The gene's-eye view of evolution*. Oxford University Press.
<https://doi.org/10.1093/oso/9780198862260.001.0001>
- Arnold, A. J., & Fristrup, K. (1982). The theory of evolution by natural selection: A hierarchical expansion. *Paleobiology*, 8(2), 113–129.
<https://doi.org/10.1017/S0094837300004462>
- Arnold, K. F., Berrie, L., Tennant, P. W. G., & Gilthorpe, M. S. (2020). A causal inference perspective on the analysis of compositional data. *International Journal of Epidemiology*, 49(4), 1307–1313. <https://doi.org/10.1093/ije/dyaa021>
- Barclay, P., & Krupp, D. B. (2016). The burden of proof for a cultural group selection account. *Behavioral and Brain Sciences*, 39, e33.
- Berrie, L., Arnold, K. F., Tomova, G. D., Gilthorpe, M. S., & Tennant, P. W. G. (in press). Depicting deterministic variables within directed acyclic graphs (DAGs): An aid for identifying and interpreting causal effects involving tautological associations, compositional data, and composite variables. *American Journal of Epidemiology*.
<https://doi.org/10.48550/arXiv.2211.13201>
- Birch, J. (2017). *The philosophy of social evolution*. Oxford University Press.
<https://doi.org/10.1093/oso/9780198733058.001.0001>
- Birch, J., & Okasha, S. (2015). Kin selection and its critics. *BioScience*, 65(1), 22–32.
<https://doi.org/10.1093/biosci/biu196>
- Bonduriansky, R., & Day, T. (2018). *Extended heredity*. Princeton University Press.
- Bowles, S., & Gintis, H. (2011). *A cooperative species: Human reciprocity and its evolution*. Princeton University Press.
- Brandon, R. (1982). The levels of selection. *Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 1982(1), 315–323.
<https://doi.org/10.1086/psaprocbienmeetp.1982.1.192676>
- Darwin, C. (1859). *On the origin of species*. J. Murray.
- Darwin, C. (1871). *The descent of man, and selection in relation to sex*. D. Appleton and Company.
- Dawkins, R. (1982). *The extended phenotype*. Oxford University Press.
- Dawkins, R. (1994). Burying the vehicle. *Behavioral and Brain Sciences*, 17(4), 616–617.
<https://doi.org/10.1017/S0140525X00036207>
- Dawkins, R. (2006). *The selfish gene: 30th anniversary edition*. Oxford University Press. (Original work published 1976)
- Dugatkin, L., & Reeve, H. K. (1994). Behavioral ecology and levels of selection: Dissolving the group selection controversy. In P. J. B. Slater, J. S. Rosenblatt, C. T. Snowdon, & M. Milinski (Eds.), *Advances in the Study of Behavior* (pp. 101–133). Academic Press.
- Elwert, F. (2013). Graphical causal models. In S. L. Morgan (Ed.), *Handbook of Causal Analysis for Social Research* (pp. 245–273). Springer Netherlands.
https://doi.org/10.1007/978-94-007-6094-3_13
- Fisher, R. A. (1930). *The genetical theory of natural selection*. Oxford University Press.
- Frank, S. A. (1998). *Foundations of social evolution*. Princeton University Press.

- Frank, S. A. (2003). Repression of competition and the evolution of cooperation. *Evolution*, *57*, 693–705.
- Frank, S. A. (2012). Natural selection. IV. The Price equation. *Journal of Evolutionary Biology*, *25*(6), 1002–1019. <https://doi.org/10.1111/j.1420-9101.2012.02498.x>
- Gardner, A. (2017). Group selection. *Reference Module in Life Sciences*. Elsevier. <https://doi.org/10.1016/B978-0-12-809633-8.06485-2>
- Gardner, A., & Grafen, A. (2009). Capturing the superorganism: A formal theory of group adaptation. *Journal of Evolutionary Biology*, *22*, 659–671.
- Gardner, A., & Welch, J. J. (2011). A formal theory of the selfish gene. *Journal of Evolutionary Biology*, *24*, 1801–1813.
- Gardner, A., West, S. A., & Wild, G. (2011). The genetical theory of kin selection. *Journal of Evolutionary Biology*, *24*, 1020–1043. <https://doi.org/10.1111/J.1420-9101.2011.02236.X>
- Godfrey-Smith, P. (2009). *Darwinian populations and natural selection*. Oxford University Press.
- Godfrey-Smith, P. (2007). Conditions for evolution by natural selection. *The Journal of Philosophy*, *104*(10), 489–516. <https://doi.org/10.5840/jphil2007104103>
- Godfrey-Smith, P., & Kerr, B. (2013). Gestalt-switching and the evolutionary transitions. *The British Journal for the Philosophy of Science*, *64*(1), 205–222. <https://doi.org/10.1093/bjps/axr051>
- Goodnight, C. J. (2015). Multilevel selection theory and evidence: A critique of Gardner, 2015. *Journal of Evolutionary Biology*, *28*(9), 1734–1746. <https://doi.org/10.1111/jeb.12685>
- Goodnight, C. J., Schwartz, J. M., & Stevens, L. (1992). Contextual analysis of models of group selection, soft selection, hard selection, and the evolution of altruism. *American Naturalist*, *140*, 743–761.
- Gould, S. J. (1980). *The panda's thumb*. W. W. Norton & Company.
- Gould, S. J., & Lloyd, E. A. (1999). Individuality and adaptation across levels of selection: How shall we name and generalize the unit of Darwinism? *Proceedings of the National Academy of Sciences*, *96*(21), 11904–11909. <https://doi.org/10.1073/pnas.96.21.11904>
- Gregory, T. R. (2009). Understanding natural selection: Essential concepts and common misconceptions. *Evolution: Education and Outreach*, *2*, 156. <https://doi.org/10.1007/s12052-009-0128-1>
- Griffin, A. S., West, S. A., & Buckling, A. (2004). Cooperation and competition in pathogenic bacteria. *Nature*, *430*, 1024–1027.
- Gwynne, D. T. (2008). Sexual conflict over nuptial gifts in insects. *Annual Review of Entomology*, *53*(1), 83–101. <https://doi.org/10.1146/annurev.ento.53.103106.093423>
- Haldane, J. B. S. (1932). *The causes of evolution*. Longmans, Green and Co.
- Hamilton, W. D. (1963). The evolution of altruistic behavior. *The American Naturalist*, *97*(896), 354–356. <https://doi.org/10.1086/497114>
- Hamilton, W. D. (1964). The genetical evolution of social behaviour (I and II). *Journal of Theoretical Biology*, *7*, 1–52.
- Hamilton, W. D. (1996). *Narrow roads of gene land* (Vol. 1: Evolution of social behaviour). W.H. Freeman/Spektrum.

- Hamilton, W. (1970). Selfish and spiteful behaviour in an evolutionary model. *Nature*, *228*, 1218–1220.
- Hamilton, W. (1975). Innate social aptitudes of Man: An approach from evolutionary genetics. In R. Fox (Ed.), *Biosocial Anthropology* (pp. 133–153). Malaby Press.
- Harman, O. S. (2010). *The price of altruism*. W.W. Norton.
- Hausman, D. M., & Woodward, J. (1999). Independence, invariance and the causal Markov condition. *The British Journal for the Philosophy of Science*, *50*(4), 521–583. <https://doi.org/10.1093/bjps/50.4.521>
- Heisler, I. L., & Damuth, J. (1987). A method for analyzing selection in hierarchically structured populations. *The American Naturalist*, *130*(4), 582–602.
- Hernán, M. A., & Robins, J. M. (2020). *Causal inference: What if*. Chapman & Hall/CRC.
- Hull, D. L. (1981). Units of evolution: A metaphysical essay. In U. J. Jensen & R. Harré (Eds.), *The Philosophy of Evolution* (pp. 23–44). St. Martin's Press.
- Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction*. Cambridge University Press. <https://doi.org/10.1017/CBO9781139025751>
- Inglis, R. F., Gardner, A., Cornelis, P., & Buckling, A. (2009). Spite and virulence in the bacterium *Pseudomonas aeruginosa*. *Proceedings of the National Academy of Sciences*, *106*, 5703–5707. <https://doi.org/10.1073/pnas.0810850106>
- Kim, J. (1998). *Mind in a physical world: An essay on the mind-body problem and mental causation*. MIT Press. <https://doi.org/10.7551/mitpress/4629.001.0001>
- Krupp, D. B. (2013). How to distinguish altruism from spite (and why we should bother). *Journal of Evolutionary Biology*, *26*, 2746–2749. <https://doi.org/10.1111/jeb.12253>
- Krupp, D. B. (2016). Causality and the levels of selection. *Trends in Ecology & Evolution*, *31*, 255–257. <https://doi.org/10.1016/j.tree.2016.01.008>
- Kunicki, Z. J., Smith, M. L., & Murray, E. J. (2023). A primer on structural equation model diagrams and directed acyclic graphs: When and how to use each in psychological and epidemiological research. *Advances in Methods and Practices in Psychological Science*, *6*, 1–14.
- Lehmann, L., Bargum, K., & Reuter, M. (2006). An evolutionary analysis of the relationship between spite and altruism. *Journal of Evolutionary Biology*, *19*, 1507–1516.
- Lloyd, E. A. (1988). *The structure and confirmation of evolutionary theory*. Greenwood Press.
- Logue, D. M., & Krupp, D. B. (2016). Duetting as a collective behavior. *Frontiers in Ecology and Evolution*, *4*, Article 7. <https://doi.org/10.3389/fevo.2016.00007>
- Marshall, J. A. R. (2011). Group selection and kin selection: Formally equivalent approaches. *Trends in Ecology & Evolution*, *26*(7), 325–332. <https://doi.org/10.1016/j.tree.2011.04.008>
- Marshall, J. A. R. (2015). *Social evolution and inclusive fitness theory*. Princeton University Press.
- Maynard Smith, J. (1964). Group selection and kin selection. *Nature*, *201*, 1145–1147.
- Mayr, E. (1963). *Animal species and evolution*. Cambridge : Belknap Press of Harvard University Press.

- McCullough, E. L., Tobalske, B. W., & Emlen, D. J. (2014). Structural adaptations to diverse fighting styles in sexually selected weapons. *Proceedings of the National Academy of Sciences*, *111*, 14484–14488. <https://doi.org/10.1073/pnas.1409585111>
- Medina, J. M., Shreenidhi, P. M., Larsen, T. J., Queller, D. C., & Strassmann, J. E. (2019). Cooperation and conflict in the social amoeba *Dictyostelium discoideum*. *The International Journal of Developmental Biology*, *63*(8-9-10), 371–382. <https://doi.org/10.1387/ijdb.190158jm>
- Muir, W. M. (1996). Group selection for adaptation to multiple-hen cages: Selection program and direct responses. *Poultry Science*, *75*(4), 447–458. <https://doi.org/10.3382/ps.0750447>
- Nunney, L. (1985). Group selection, altruism, and structured-deme models. *The American Naturalist*. <https://doi.org/10.1086/284410>
- Okasha, S. (2004). Multilevel selection and the partitioning of covariance: A comparison of three approaches. *Evolution*, *58*(3), 486–494. <https://doi.org/10.1111/j.0014-3820.2004.tb01672.x>
- Okasha, S. (2006). *Evolution and the levels of selection*. Oxford University Press.
- Okasha, S. (2014). Emergent group traits, reproduction, and levels of selection. *Behavioral and Brain Sciences*, *37*, 268–269. <https://doi.org/10.1017/S0140525X13002963>
- Okasha, S. (2016). The relation between kin and multilevel selection: An approach using causal graphs. *British Journal for the Philosophy of Science*, *67*, 435–470. <https://doi.org/10.1093/bjps/axu047>
- Okasha, S., & Otsuka, J. (2020). The Price equation and the causal analysis of evolutionary change. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1797), 20190365. <https://doi.org/10.1098/rstb.2019.0365>
- Okasha, S., & Paternotte, C. (2012). Group adaptation, formal darwinism and contextual analysis. *Journal of Evolutionary Biology*, *25*(6), 1127–1139. <https://doi.org/10.1111/j.1420-9101.2012.02501.x>
- Otsuka, J. (2016a). Causal foundations of evolutionary genetics. *The British Journal for the Philosophy of Science*, *67*(1), 247–269. <https://doi.org/10.1093/bjps/axu039>
- Otsuka, J. (2016b). A critical review of the statisticalist debate. *Biology & Philosophy*, *31*(4), 459–482. <https://doi.org/10.1007/s10539-016-9528-0>
- Otsuka, J. (2019). *The role of mathematics in evolutionary theory*. Cambridge University Press.
- Panchanathan, K. (2011). George Price, the Price equation, and cultural group selection. *Evolution and Human Behavior*, *32*(5), 368–371. <https://doi.org/10.1016/j.evolhumbehav.2011.04.001>
- Patel, M., West, S. A., & Biernaskie, J. M. (2020). Kin discrimination, negative relatedness, and how to distinguish between selfishness and spite. *Evolution Letters*, *4*(1), 65–72. <https://doi.org/10.1002/evl3.150>
- Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal inference in statistics: A primer* (1 edition). Wiley.
- Pearl, J., & Mackenzie, D. (2018). *The Book of Why*. Basic Books.
- Price, G. R. (1970). Selection and covariance. *Nature*, *227*, 520–521.
- Price, G. R. (1972). Extension of covariance selection mathematics. *Annals of Human Genetics*, *35*(4), 485–490. <https://doi.org/10.1111/j.1469-1809.1977.tb01874.x>

- Queller, D. (1992). A general model for kin selection. *Evolution*, *46*, 376–380.
- Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, *1*(1), 27–42. <https://doi.org/10.1177/2515245917745629>
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. <https://doi.org/10.1037/h0037350>
- Salt, G. W. (1979). A comment on the use of the term *emergent properties*. *The American Naturalist*, *113*(1), 145–148. <https://doi.org/10.1086/283370>
- Smaldino, P. E. (2014). The cultural evolution of emergent group-level traits. *Behavioral and Brain Sciences*, *37*(03), 243–254. <https://doi.org/10.1017/S0140525X13001544>
- Smukalla, S., Caldara, M., Pochet, N., Beauvais, A., Guadagnini, S., Yan, C., Vinces, M. D., Jansen, A., Prevost, M. C., Latge, J. P., Fink, G. R., Foster, K. R., & Verstrepen, K. J. (2008). FLO1 is a variable green beard gene that drives biofilm-like cooperation in budding yeast. *Cell*, *135*, 726–737. <https://doi.org/10.1016/J.Cell.2008.09.037>
- Sober, E., & Wilson, D. S. (1998). *Unto others*. Harvard University Press.
- Sober, E. (1984). *The nature of selection: Evolutionary theory in philosophical focus*. University of Chicago Press.
- Taylor, P. D., Wild, G., & Gardner, A. (2007). Direct fitness or inclusive fitness: How shall we model kin selection? *Journal of Evolutionary Biology*, *20*, 301–309. <https://doi.org/10.1111/J.1420-9101.2006.01196.X>
- Vrba, E. S. (1989). Levels of selection and sorting with special reference to the species level. In P. H. Harvey & L. Partridge (Eds.), *Oxford Surveys in Evolutionary Biology* (pp. 111–168). Oxford University Press.
- West, S. A., Griffin, A. S., & Gardner, A. (2007). Social semantics: Altruism, cooperation, mutualism, strong reciprocity and group selection. *Journal of Evolutionary Biology*, *20*, 415–432.
- West, S. A., Griffin, A. S., & Gardner, A. (2008). Social semantics: How useful has group selection been? *Journal of Evolutionary Biology*, *21*, 374–385.
- Williams, G. C. (1992). *Natural selection: Domains, levels, and challenges*. Oxford University Press.
- Williams, G. (1966). *Adaptation and natural selection*. Princeton University Press.
- Wilson, D. S. (1975). A theory of group selection. *Proceedings of the National Academy of Sciences*, *72*(1), 143–146. <https://doi.org/10.1073/pnas.72.1.143>
- Wilson, D. S., & Wilson, E. O. (2007). Rethinking the theoretical foundation of sociobiology. *The Quarterly Review of Biology*, *82*(4), 327–348. <https://doi.org/10.1086/522809>
- Wilson, D. S., & Sober, E. (1994a). Group selection: The theory replaces the bogey man. *Behavioral and Brain Sciences*, *17*(4), 639–654. <https://doi.org/10.1017/S0140525X0003644X>
- Wilson, D. S., & Sober, E. (1994b). Reintroducing group selection to the human behavioral sciences. *Behavioral and Brain Sciences*, *17*(4), 585–608. <https://doi.org/10.1017/S0140525X00036104>

- Wimsatt, W. C. (1980). Reductionistic research strategies and their biases in the units of selection controversy. In T. Nickles (Ed.), *Scientific discovery: Case studies* (pp. 213–259). D. Reidel.
- Wright, S. (1920). The relative importance of heredity and environment in determining the piebald pattern of guinea-pigs. *Proceedings of the National Academy of Sciences*, 6(6), 320–332. <https://doi.org/10.1073/pnas.6.6.320>
- Wright, S. (1931). Evolution in Mendelian populations. *Genetics*, 16(2), 97–159. <https://doi.org/10.1093/genetics/16.2.97>
- Wynne-Edwards, V. (1962). *Animal dispersion in relation to social behaviour*. Oliver and Boyd.
- Yamagishi, T., Horita, Y., Mifune, N., Hashimoto, H., Li, Y., Shinada, M., Miura, A., Inukai, K., Takagishi, H., & Simunovic, D. (2012). Rejection of unfair offers in the ultimatum game is no evidence of strong reciprocity. *Proceedings of the National Academy of Sciences*, 109(50), 20364–20368. <https://doi.org/10.1073/pnas.1212126109>