KIN SELECTION

A PHILOSOPHICAL ANALYSIS

Jonathan Birch

Clare College,

University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

KIN SELECTION

A PHILOSOPHICAL ANALYSIS

Jonathan Birch

This dissertation examines the conceptual and theoretical foundations of the most general and most widely used framework for understanding social evolution, W. D. Hamilton's theory of kin selection. While the core idea is intuitive enough (when organisms share genes, they sometimes have an evolutionary incentive to help one another), its apparent simplicity masks a host of conceptual subtleties, and the theory has proved a perennial source of controversy in evolutionary biology. To move towards a resolution of these controversies, we need a careful and rigorous analysis of the philosophical foundations of the theory. My aim in this work is to provide such an analysis.

I begin with an examination of the concepts behavioural ecologists employ to describe and classify types of social behaviour. I stress the need to distinguish concepts that are often conflated: for example, we need to distinguish simple cooperation from collaboration in collective tasks, behaviours from strategies, and control from manipulation and coercion. I proceed from here to the formal representation of kin selection via George R. Price's covariance selection mathematics. I address a number of interpretative issues the Price formalism raises, including the vexed question of whether kin selection theory is 'formally equivalent' to multi-level selection theory. In the second half of the dissertation, I assess the uses and limits of Hamilton's rule for the evolution of social behaviour; I provide a precise statement of the conditions under which the rival neighbour-modulated fitness and inclusive fitness approaches in contemporary kin selection theory are equivalent (and describe cases in which they are not); and I criticize recent formal attempts to establish the controversial claim that kin selection leads to organisms behaving as if maximizing their inclusive fitness.

DECLARATION OF ORIGINALITY

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

STATEMENT OF LENGTH

This dissertation, including footnotes and appendices but excluding the bibliography, contains 78,072 words. It does not exceed the word limit of 80,000 words set by the Degree Committee of the Department of History and Philosophy of Science, University of Cambridge.

ACKNOWLEDGEMENTS

I have received a huge amount of support, both personal and institutional, without which this work would never have been possible. First of all, I thank the Arts and Humanities Research Council and Christ's College for funding, and I thank the staff and students of the Department of History and Philosophy of Science, University of Cambridge, for providing such a friendly and stimulating intellectual environment.

I have presented work drawn from or related to this dissertation to audiences at the University of Cambridge, IHPST (Paris), the University of Bristol, the University of Utah, All Souls College (Oxford), the University of Toronto and the University of Stirling, and I thank all who attended for their comments and questions.

For their detailed comments on parts of this dissertation or on related work, I thank Anna Alexandrova, Kevin Brosnan, Tim Button, Hasok Chang, Ellen Clarke, Sarah Coakley, Eugene Earnshaw-Whyte, Andy Gardner, Herbert Gintis, Philippe Huneman, Nick Jardine, Bram Kuijper, Johannes Martens, Bence Nanay, Robert Northcott, Martin Nowak, Samir Okasha, Cedric Paternotte, Denis Walsh, Rob Wilson and Joeri Witteveen. For helpful email exchanges and/or face-to-face discussions on these issues, I thank (in addition to the above) Andrew Bourke, Alan Grafen, Rufus Johnstone, Joel Peck, David Queller, Joan Strassmann and John Welch.

For their outstanding work in copyediting and proofreading the dissertation, I thank Caroline Birch, Peter Birch, Beth Hannon and Caitlin Wylie.

For his invaluable advice and feedback at all stages of the project, I thank my supervisor Tim Lewens.

Most of all, I thank my parents Peter and Marie, my sister Rosie and my wife Caroline. This dissertation is dedicated to them.

Jonathan Birch

January 2013

TABLE OF CONTENTS

ANALYTIC TABLE OF CONTENTS

ONE

Jumping into the River

1.1 Haldane's dangerous idea

Brothers and cousins

As legend has it, the pithiest expression of the concept of kin selection was made long before the term 'kin selection' was coined, and long before the theory itself was devised. When asked if he would dive into a river to rescue a drowning stranger, the geneticist and co-architect of the modern synthesis J. B. S. Haldane is said to have replied: 'No, but I would do it for two brothers or eight cousins' (see, e.g., Maynard Smith 1976a; McElreath and Boyd 2007). The sound bite captures an intuitive and powerful thought: when interacting organisms share genes, the organisms may, in certain circumstances, have an evolutionary incentive to help one another. Moreover, and more profoundly, it suggests that the size of the incentive to help is *directly proportional to the closeness of the organisms' genealogical relationship*.

This simple idea, in one form or another, has for almost fifty years served as a guiding thread for theoretical and empirical work on the evolution of social behaviour. Thanks largely to Richard Dawkins (1976), it has also come to be the public face of social evolution theory: an idea virtually all biologists and biology students take themselves to be familiar with in outline, even if they have never studied it in detail. Perhaps unsurprisingly, however, the idea turns out to be a great deal subtler than it may at first appear to be. It also turns out to be a lot *more general* than it first appears to be, for it potentially sheds light

on the evolution of social behaviour even when interacting organisms are *not* closely related genealogically.

From kin recognition to population structure

To see just one way in which the concept of kin selection might be subtler and more general than it first appears to be, we can look more closely at Haldane's thoughts on the subject. While the story of the 'brothers and cousins' quip may be apocryphal, there is no doubt that Haldane considered such issues in some detail, and with impressive subtlety (Maynard Smith 1976a). Clear evidence of this can be seen in the following passage from his 1955 article, 'Population Genetics':

> [I]t is only in such small populations that natural selection would favour the spread of genes making for certain kinds of altruistic behaviour. Let us suppose that you carry a rare gene which affects your behaviour so that you jump into a river and save a child, but you have one chance in ten of being drowned, while I do not possess the gene, and stand on the bank and watch the child drown. If the child is your own child or your brother or sister, there is an even chance that the child will also have the gene, so five such genes will be saved in children for one lost in an adult. If you save a grandchild or nephew the advantage is only two and a half to one. If you only save a first cousin, the effect is very slight. If you try to save your first cousin once removed the population is more likely to lose this valuable gene than to gain it. But on the two occasions when I have pulled possibly drowning people out of the water (at an infinitesimal risk to myself) I had no time to make such calculations. Paleolithic men did not make them. It is clear that genes making for conduct of this kind would only have a chance of spreading in rather small populations where most of the children were fairly near relatives of the man who risked his life. It is not easy to see how, except in small populations, such genes could have been established. Of course the

> conditions are even better in a community such as a beehive or an ants' nest, whose members are all literally brothers and sisters. (Haldane 1955, 44)

This may or may not be the true origin of the famous 'brothers and cousins' remark. Either way, however, it is plain that the popular sound bite does not do justice to Haldane's considered opinions on the matter. For we see here that Haldane considers *and rejects* the possibility that the evolution of altruism relies on actors making explicit cognitive judgements about their degree of kinship to prospective recipients. This is just as well, since it would be rather implausible to attribute such a cognitive capacity to bees, ants, amoebae, bacteria, or many of the other non-human species in which apparently altruistic behaviour is rife, suggesting that any evolutionary mechanism that *required* active and conscious kin discrimination would be rather limited in scope (cf. Chapter 2). As an alternative, Haldane suggests that what is needed are 'small populations' composed of close relatives, such that actors are likely to interact with close kin *without any need for active discrimination on their part.* His suggestion, in other words, is that ecological factors often do the work of bringing relatives together, so that cognitive judgements about degrees of kinship are unnecessary.

Contemporary kin selection theorists often share Haldane's scepticism about the biological importance of active kin discrimination, and would agree with him about the comparatively greater importance of population structure in bringing relatives together; but they would not follow him in concluding that kin selection thus requires small populations. Note, however, that when Haldane talks of small populations he is specifically envisaging beehives, ants' nests and other similar 'communities' that we would today more commonly describe as groups of related individuals, which compete against other such groups in a larger, overarching population (or 'meta-population'). Arguably, therefore, Haldane's talk of small populations should be read in modern terminology as describing a group-structured population. If this is indeed what he had in mind, Haldane had latched on to a profound insight about social evolution. Kin selection

requires that relatives interact differentially, but differential interaction between relatives does not require active kin recognition on the part of social actors, *for it can also arise through population structure*.[1] This is the key to understanding how kin selection can drive sociality not merely in organisms with advanced cognitive faculties, but also in populations of insects, amoebae, bacteria and perhaps even simple replicating molecules.

1.2 Kin selection broad and narrow

We owe the formal theory of kin selection chiefly to W. D. Hamilton (1963, 1964, 1970, 1971, 1972, 1975), and we owe the term 'kin selection' to Hamilton's early champion, John Maynard Smith (1964). Today, Hamilton's theory lies at the heart of an established and burgeoning research programme, the explanatory domain of which has steadily expanded over recent decades (Bourke 2011; Birch 2012a). From very early on, Hamilton and other theorists realized (as Haldane too appears to have done) that kin recognition is by no means necessary for kin selection: population structure may bring relatives together so that they interact differentially (with selectively significant consequences), irrespective of whether organisms are capable of detecting their kin and adjusting their behaviour in response (Hamilton 1964, 1971, 1972, 1975; Dawkins 1979).

Theorists quickly realized that shared genealogy is technically unnecessary too: for the purposes of (early formulations of) kin selection theory, relatives are simply individuals who are more likely than average to share genes at genomic loci relevant to the social behaviour of interest, and this sort of 'relatedness' (i.e., genetic correlation) could in principle arise by mechanisms other than shared genealogy (Hamilton 1975; Michod and Hamilton 1980; Queller 1985). Dawkins (1976) famously offers the example of a

[1] Group structure is not strictly required; relatives may also interact differentially in a 'neighbour-structured' population (see Chapter 2; see also Maynard Smith 1976; Godfrey-Smith 2006a, 2008; Godfrey-Smith and Kerr 2009).

'greenbeard' mechanism, in which a gene or gene-complex causes its bearers to (a) grow a green beard, (b) recognize other bearers on the basis of their green beards, and (c) differentially assist these individuals. If this greenbeard gene were to arise by mutation in several individuals, it would lead to differential interaction between these individuals, in turn leading to positive relatedness at the greenbeard locus. The example is hypothetical, of course; but strikingly similar effects (mediated not by literal green beards, but by phenotypic markers playing a similar role) have since been discovered empirically (West and Gardner 2010).

Recent work has seen a yet more dramatic broadening of the notion of kin selection, on which it is not even required that the selectively salient correlations captured by coefficients of relatedness are wholly *genetic* in character. Extending relatedness to encompass partly or wholly *phenotypic* correlations allows the theory of kin selection to be extended to new classes of phenomena, including interspecific mutualisms and reciprocal altruism in humans (Fletcher and Zwick 2006; Fletcher and Doebeli 2009; Godfrey-Smith 2009a; see also Chapter 5). All that is essential to a process of kin selection in this broader sense is that relatives interact differentially, where 'relatives' are organisms who resemble each other more than a randomly chosen pair of organisms would do with respect to certain selectively significant properties. What is *not* essential to the process is any particular mechanism for generating differential interactions among relatives, or any particular characterization of the selectively significant properties that count for the purposes of evaluating relatedness. On this conception, the notion of kin selection encompasses, in effect, all processes of evolution by natural selection in which resemblance between interacting organisms matters. This dissertation is concerned with kin selection in this maximally broad sense.

One might object that this makes the notion of kin selection absurdly broad—so broad that it incorporates, in effect, any process of social evolution driven by natural selection. I reply that this is not so absurd: I see nothing seriously problematic about extending the domain

of a theory to encompass all the diverse phenomena it is apt to explain. In my view, the theory of kin selection is a powerful and highly general theory of how natural selection can drive the evolution of social behaviour. Given that the theory may be readily extended to *any* salient relation of resemblance between organisms, little is gained by reserving the term 'kin selection' exclusively for that subset of processes in which social behaviours are favoured in virtue of shared genealogy between social partners, or for that even narrower subset of processes in which social behaviours are favoured in virtue of mechanisms for kin discrimination.[2]

The expanded explanatory domain of broad-sense kin selection includes a dizzying array of social phenomena that cross hugely disparate taxa, from bacteria to baboons, from mitochondria to meerkats, from sperm cells to sperm whales (see Queller 1997, 2000; Bourke 2011; see also Chapter 2). Of course, while the selective processes responsible for these phenomena may all be instances of kin selection in the broadest possible sense, they obviously vary in many superficial respects; and they undoubtedly vary in deep and important respects too. The challenge for the social evolution theorist is to formulate and defend a theory of kin selection broad enough to apply to all these cases and yet informative enough to say something interesting about each of them. The goal of this dissertation is to undertake the conceptual groundwork for this highly ambitious project.

1.3 Five central questions

The predominant place of kin selection in contemporary evolutionary biology is not well reflected in the philosophy of biology. For reasons that probably have to do more with historical contingencies than anything else, philosophers of biology have typically devoted

[2] Of course, when the term 'kin selection' is used in this broad sense, it is arguably something of a misnomer, since broad-sense kin selection does not require kinship in the traditional sense of the word. But the ordinary concept of kinship can also be extended (metaphorically) to similarity relations other than shared genealogy (e.g., 'the heart is akin to a pump'; 'they are kindred spirits').

more attention to the concept of *group* (or *multi-level*) selection (e.g., Wimsatt 1980, 1981; Brandon 1982, 1988; Sober 1984; Lloyd 1998, 2012; Sober and D. Wilson 1994, 1998; Sterelny 1996; Sterelny and Griffiths 1999; Godfrey-Smith and Kerr 2002, forthcoming; Okasha 2001, 2003, 2004a,b,c, 2005a,b, 2006, 2009; R. Wilson 2003, 2005; Glymour 2008; Godfrey-Smith 2009a), to such an extent that a philosophical reader with no familiarity with the biological literature might reasonably jump to the conclusion that group selection, not kin selection, was the prevailing explanatory paradigm for the evolutionary explanation of social behaviour.

One aim of this dissertation is to redress the balance: to give the theory of kin selection the kind of detailed philosophical treatment that previous authors have afforded the theory of multi-level selection. I am undertaking this task not because I take kin selection theory to be overwhelmingly superior to its rivals, as its loudest defenders often (regrettably) take it to be. I think the true relationship between the theories of kin and multi-level selection is rather subtle (not least because, as we have already noted, interactions between relatives often arise due to group structure) and that there is room for them to coexist peacefully in contemporary theory (see Chapter 3; see also Birch 2012b). My motivation is not so adversarial: I am undertaking this task simply because I see a range of conceptual and foundational questions regarding the theory of (broad-sense) kin selection that are worth addressing, but that previous work in the philosophy of biology has not adequately addressed.

The dissertation is structured around five such questions. Taken together, they provide the framework for a comprehensive examination of the conceptual foundations of the theory of (broad-sense) kin selection. I start at the beginning, with an examination of the very ideas of biological cooperation (Chapter 2) and natural selection (Chapter 3). I then move on to methodological and conceptual issues specific to the analysis of kin selection in the broadest sense of the term. I examine the scope and limits of Hamiltion's rule for the evolution of social behaviour (Chapter 4), I address the question of whether the two

alternative conceptions of social fitness in contemporary theory are formally equivalent (Chapter 5), and I consider whether we have reason to believe that kin selection leads to organisms behaving as if maximizing their inclusive fitness (Chapter 6). Here, I will briefly introduce each of these issues.

What is biological cooperation?

One might be sceptical of the very idea of biological cooperation, if the notion is intended to apply not merely to human beings and other intelligent mammals, but also to insects and cells. After all, cooperation in the ordinary sense of the word seems to denote a sophisticated *cognitive* achievement: the product of thinking agents working in pursuit of common goals. Could it really be anything other than a loose and anthropomorphic metaphor to talk of cooperation among insects or among cells? In Chapter 2, 'Cooperation, Collaboration and Control', I argue that biological cooperation is an unmistakable natural phenomenon, and that a genuine capacity for intentional action is not a prerequisite for its manifestation. This naturally leads to the question of how biological cooperation *is* to be characterized, if not in terms of intentional agency. The usual approach is to gloss cooperation as a special type of 'fitness transaction' between actors and recipients. I defend this approach, but I also discuss a number of conceptual subtleties it raises. From here, I proceed to an expanded conceptual taxonomy of biological cooperation. I propose and defend similarly naturalistic accounts of task-based cooperation, control, enforcement and manipulation, and I show how these categories can be used to classify social behaviours along several different axes.

How should we formalize the concept of natural selection?

Kin selection is a type of natural selection, and it would be foolhardy to attempt to make sense of the former without first examining the latter. In Chapter 3, 'Selection, Transmission and the Price Formalism', I grapple with a cluster of issues concerning the formal representation of natural selection that recent philosophical work in this area has

raised (see especially Godfrey-Smith 2006a, 2007, 2008, 2009; Kerr and Godfrey-Smith 2009; Okasha 2004a, b, 2006, 2011; Waters 2011).

I first distinguish phenotypic and genetic formulations of the Price equation for evolutionary change (Price 1970, 1972a), and argue that genetic formulations (especially a quantitative-genetic formulation in terms of breeding values) are particularly useful for the purposes of social evolution theory. I then move on to the vexed issue of how to separate the effects of natural selection from the effects of other evolutionary processes within the Price formalism. I argue that we need to distinguish three effects of natural selection on the population mean for some character of interest: a primary effect (covariance between the character and fitness), a secondary effect (covariance between fitness and transmission biases with respect to the character) and a tertiary effect (changes in the average effects of alleles). The three effects influence different terms of the Price equation. It follows that to neglect any of the terms is to risk neglecting at least one of these effects.

Finally, I turn to the question of how the sorting of organisms into equivalence classes affects the analysis of natural selection. Drawing on work by Peter Godfrey-Smith (2006, 2008), I formulate a general framework for thinking about different types of equivalence class, and I use this general framework to relate three special cases: trait-groups, genotypic classes and developmental classes. I argue that these types of equivalence class are useful for different reasons: they all entitle us to neglect something for the purposes of analysis, but they differ with respect to what they entitle us to neglect. When a population can be subdivided into trait-groups, we can discount as irrelevant interactions that cut across group boundaries; when populations can be subdivided into genotypic classes, we can discount as irrelevant variance in fitness within classes; and when populations can be subdivided into well-chosen developmental classes, we can discount as irrelevant variance in fitness between classes. I close by relating the foregoing discussion to two further issues: the relationship between the Price formalism and evolutionary nominalism (*sensu*

Godfrey-Smith 2009a), and the relationship between kin- and group-selectionist approaches to the analysis of social evolution.

What are the uses and limits of Hamilton's rule?

The most important bridge from the abstract world of population genetics to the real world of behavioural ecology is Hamilton's rule, a deceptively simple statement of the conditions under which we can expect a social behaviour to be favoured by natural selection. The rule states, broadly speaking, that a social behaviour will be favoured by natural selection if and only if $rb - c > 0$, where b represents the benefit the behaviour confers on the recipient, c represents the cost it imposes on the actor, and r represents the relatedness between actors and recipients.

Recent years have seen considerable debate about the validity and value of the rule: some authors argue that it only holds under restrictive assumptions (e.g., Nowak et al. 2010; van Veelen 2009); others argue that it is too simple to capture the causal structure of most real processes of social evolution (e.g., Queller 2011); while others have sought to defend the traditional version of the rule from these critiques (e.g., Gardner et al. 2011). Much of the debate has centred on whether the rule still applies (and if it does, whether it is still useful) when related organisms interact synergistically. A synergistic effect is a joint effect of multiple social behaviours, the value of which in fitness terms differs from a mere sum of the effects each of these behaviours would have had in the absence of the others; in short, synergistic effects are 'more (or less) than the sum of their parts'.[3] The applicability of Hamilton's rule in the presence of such effects has long been a bone of contention in kin selection theory (see, e.g., Queller 1985; Grafen 1985b), and the issue continues to polarize theorists.

[3] Synergistic effects are sometimes referred to as non-additive effects. I favour the term 'synergistic' here, because the concepts of 'additivity' and 'non-additivity' can have a variety of meanings in the context of social evolution theory, generating scope for semantic confusion. In particular, the effects of dominance and epistasis are often described as non-additive, yet have nothing to do with social interaction.

In Chapter 4, 'The Scope and Limits of Hamilton's Rule', I argue (building on an argument first made informally by David C. Queller 1992a) that Hamilton's rule cannot be relied upon as a guide to the direction of natural selection in the presence of synergy, at least if the cost and benefit coefficients are interpreted as the average effects of phenotypes. There are two ways round the problem, both of which have been fruitfully pursued by Queller and others in the past two decades: one is to develop extended versions of Hamilton's rule involving additional phenotypic predictors (Frank 1998; Queller 2011); the other is to formulate the rule in purely genetic terms (Queller 1992b; Gardner et al. 2007, 2011). I compare and contrast the two approaches, arguing that the right response to the 'problem of synergy' ultimately depends on the explanatory function Hamilton's rule is intended to serve.

Do neighbour-modulated and inclusive fitness provide equivalent representations of kin selection?

In a process of kin selection, social traits evolve because, for one reason or another, relatives (i.e., saliently similar individuals) interact differentially. But *why* does the differential interaction of relatives (in this sense) make such a difference? In particular, why does it help enable the evolution of altruism? In contemporary kin selection theory, one finds two apparently quite different answers to this question.

On one answer, the differential interaction of relatives matters because it leads to positive correlation between the genes of a social actor and the social effects to which it is exposed. The implication is that agents who carry genes for altruism are more likely than average to be affected by its manifestation in others. The result, in some cases, is that bearers of genes for altruism are fitter, on average, than non-bearers. This thought is formalized in the neighbour-modulated fitness framework, in which we construe 'fitness' in its usual sense within the Price formalism (i.e., the number of direct lineal descendants an individual contributes to the descendant-population) and analyse the ways in which this quantity is affected by the behaviour of one's social partners. The role of relatedness in this

framework is primarily as a measure of the strength of correlation between a social actor's genes and its social milieu (Frank 1997a,b, 1998).

The other answer is that the differential interaction of relatives matters because, when actors and recipients share genes, the recipient provides the actor with an indirect route to genetic representation in the next generation: an indirect channel of transmission, so to speak. The implication of this 'two channels' picture is that, even if an action detracts from one's genetic representation through the direct channel, it may still spread overall if this loss is outweighed by a gain in genetic representation through the indirect channel. This thought is formalized by the inclusive fitness framework, in which the direct and indirect components of an individual's genetic representation in the next generation are aggregated to construct a more 'inclusive' measure of its fitness. Relatedness appears in this framework primarily as a measure of the fidelity with which an actor indirectly transmits its genes to the next generation via the recipient (Frank 1997a, b, 1998).

On the face of it, it would be intuitively surprising if these answers turned out to be equivalent ways of saying the same thing. It would be intuitively surprising, that is, if it turned out that an actor has an indirect route to genetic representation in the next generation if and only if a recipient's genes correlate positively with the fitness effects it receives; *and* if it turned out that an actor's transmission fidelity through the indirect channel is exactly equal to the strength of the correlation between a recipient's genes and its social milieu. Yet it is widely thought that the neighbour-modulated and inclusive fitness approaches to kin selection are indeed formally equivalent perspectives on the same process, despite their significant superficial differences (e.g., Dawkins 1982; Taylor et al. 2007; Rosas 2010; Wenseleers et al. 2010; Gardner et al. 2011; Queller 2011). In Chapter 5, 'Two Conceptions of Social Fitness', I attempt to get to the heart of this subtle and complex issue. I show that, at least on Steven A. Frank's (1997a, b; 1998) influential formalism, the two approaches are formally equivalent (in the sense that they are sure to agree regarding the overall direction of selection, and regarding how it should be

partitioned into 'direct' and 'indirect' components) *some* of the time but not *all* of the time. Moreover, we can specify the precise conditions under which they are equivalent, and we can classify the various kinds of cases in which they are not.

Does kin selection lead to organisms acting as if maximizing their inclusive fitness?

The caveats of Chapter 5 notwithstanding, the neighbour-modulated and inclusive fitness approaches to the analysis of kin selection are equivalent in a substantial range of cases. The former is the simpler of the two, since the latter requires explicit consideration of how social fitness effects are controlled by actors. This inevitably leads to the question of how inclusive fitness earns its keep in contemporary theory, if not through affording greater simplicity. A common suggestion is that inclusive fitness theory is valuable because it identifies a quantity (viz. inclusive fitness) that an individual organism will 'act as if maximizing' (Dawkins 1982; Grafen 1984, 2006a). It therefore underwrites an agential heuristic (or 'individual-as-maximizing-agent' analogy) in which we predict and explain the social behaviours an organism is likely to have evolved by considering the actions a rational agent would choose to perform, if it were maximizing its inclusive fitness.

A heuristic of this sort lies at the heart of informal arguments that appeal to an organism's inclusive fitness interests; it is also central to some versions of optimality modelling and evolutionary game theory (see, e.g., Maynard Smith 1982). But is it justified? Is there any theoretical support for the idea that kin selection produces organisms that act as if maximizing their inclusive fitness? In Chapter 6, I address this question as part of a broader discussion of the relationship between concepts of maximization and concepts of selection. I begin by distinguishing four varieties of 'maximization' in evolutionary theory. I then consider the question of where the most famous 'maximization' principle in 20th Century population genetics—Ronald A. Fisher's (1930) fundamental theorem of natural selection—fits in relation to this four-part distinction. The main moral I want to draw is that Fisher's theorem, by virtue of being a purely population-level principle, tells us nothing about what we should expect individual organisms to act as if maximizing.

Indeed, there is *no* formal result in 20th Century population genetics that definitively settles this latter issue. In recent work, the Oxford geneticist Alan Grafen (1999, 2000, 2002, 2006a, b, 2007a, b, c, 2008, 2009) has sought to close this theoretical lacuna through his 'Formal Darwinism' project. The rest of the chapter explains the structure of Grafen's arguments and subjects them to critical scrutiny. I close on a somewhat sceptical note: much more work needs to be done, I argue, to truly vindicate the notion that kin selection reliably leads to organisms acting as if maximizing their inclusive fitness.

For the most part, this is not a thesis with a thesis: there is no single big idea I will spend the next five chapters defending. The aim is to make a substantial contribution to many debates, rather than a revolutionary contribution to one. Even so, there are some important themes that recur throughout the dissertation, bringing the chapters together as components of a cohesive project. The dissertation ends with a review of these themes.

1.4 Theory and philosophy

As the précis above makes plain, the dissertation engages primarily with the work of evolutionary geneticists. To be specific, the authors cited and discussed most often are Ronald A. Fisher, George R. Price, W. D. Hamilton, David C. Queller, Steven A. Frank and Alan Grafen: authors whose work is undoubtedly rich in philosophical insights concerning the nature of the evolutionary process, but all geneticists or mathematicians by training. In light of this, one might be forgiven for wondering if this dissertation is really a direct contribution to evolutionary genetics rather than a contribution to the philosophy of science.

I reject this suggestion, mainly because I reject the very idea of a sharp distinction between science and its philosophy. An example will help to explain why. At many points in this dissertation, I use formal language to express an argument more precisely than I could do

verbally. The formalism I use is the covariance selection mathematics of Price (1970, 1972a). The Price formalism, like classical logic and the probability calculus, straddles the border between mathematics and philosophy: it has the abstract symbolism characteristic of mathematics, but its role in contemporary theory is quite different to that of other uses of mathematics in population genetics.

Mathematics in population genetics is most commonly employed to construct dynamically detailed but highly idealized models of evolutionary processes. Typically, we begin by writing equations which specify the determinants of fitness, and which specify how fitness relates to frequency change. In both cases, the equations we use tend to belie the true complexity these relationships would have in actual evolving populations. We then analyse or simulate hypothetical evolutionary processes that satisfy our idealized specification. The Price formalism is not like this. Rather than buying dynamical detail at the expense of descriptive accuracy, the Price formalism does the opposite: it buys descriptive accuracy at the expense of dynamical detail. It aims to provide a *literally true* description of evolutionary processes, but the cost of literal truth is that the description proceeds at a high level of abstraction, eschewing any detailed description of the underlying dynamics.[4]

This has led to the accusation that the Price formalism is not serious mathematics at all—that, because it avoids any detailed description of evolutionary dynamics, it is little more than a way of making informal, verbal arguments look more credible than they actually are (van Veelen 2005; van Veelen et al. 2010, 2012; Nowak and Highfield 2011). In my view, this accusation rests on a misunderstanding of the role of the Price formalism in contemporary theory. The Price equation and results derived from it are not in any way supposed to *supplant* concrete mathematical models of evolutionary dynamics. Instead, their role is a unifying one: their job is to bring together otherwise diverse results from

[4] See Godfrey-Smith 2009b for further discussion of abstraction and idealization in evolutionary theory.

modelling work and empirical studies under a common conceptual framework (Grafen 1985a; Gardner et al. 2007; Frank 2012). One could argue, in light of this, that the Price formalism is doing a philosophical job rather than a scientific job: rather than embodying empirical claims about the world, its role is to provide a kind of evolutionary *Aufbau*—a universal formal language in which to unify a wide range of theoretical and empirical results. But I would sooner conclude that the Price formalism blurs the boundaries between science and philosophy, since its development is a project to which philosophers and geneticists are both well placed to contribute.

The same can be said, I think, for many foundational issues in contemporary evolutionary theory. Many evolutionary biologists would regard the five questions I listed in Section 1.3 as too general, too abstract, too conceptual and too distant from immediate empirical concerns to be worthy of serious research time. In my view, the philosophy of science is at its best when addressing such questions: questions that arise directly from contemporary science but that, for whatever reason, are marginalized or ignored by the vast majority of scientists in the area concerned (cf. Chang 2004; Pigliucci forthcoming). Addressing such questions inevitably requires close attention to the details of the relevant science, but it also requires attention to conceptual subtleties to which the scientists themselves only rarely have either the time or the will to attend.

1.5 The wider context

Still, one might ask: why address these questions *now*? What are the prospective payoffs of this discussion for evolutionary geneticists and philosophers of biology? Let me suggest two reasons why this dissertation is of broad and timely importance to evolutionary theory and its philosophy.

A field in disarray?

The first is that—perhaps more than at any other time in its relatively short history—the integrity of social evolution theory is threatened by entrenched and bitter factionalism. Notably, there remain deep divisions between proponents of kin and multi-level selection. There is a longstanding disagreement about whether (and in what in sense) the two frameworks are equivalent (e.g., Nowak and Traulsen 2006; Traulsen 2010; Wenseleers et al. 2010; Marshall 2011a; van Veelen et al. 2012), and there is an equally longstanding (and, in my view, more or less orthogonal) disagreement about which approach is more useful for explanatory purposes (e.g., West et al. 2007a, 2008, 2010; D. Wilson 2008; Eldakar and Wilson 2011). I will bring some novel considerations to bear on this debate in Chapter 3. More generally, however, it is clear that the debate cannot be resolved without a careful and detailed treatment of the conceptual foundations of both theories. Samir Okasha (2006) offers such a treatment of the multi-level framework, but no comparable treatment is currently available for the theory of kin selection. Remedying this deficiency in the philosophical literature is essential if we want to reconcile the kin and group selectionist camps.

There is a less visible but equally worrying chasm opening between kin selection theory and evolutionary game dynamics (Nowak 2006a). As noted above, contemporary kin selection theory is usually formalized using Price's (1970) covariance selection mathematics, a method of analysis that provides a general and accurate description of the evolutionary change between earlier and later time-slices of a population, but one that proceeds without modelling the dynamics responsible for the change. Evolutionary game dynamics, as the name suggests, *does* model dynamics—albeit in a highly idealized way—and is therefore better able to analyse the effects of frequency-dependence on long-run evolutionary outcomes. In recent years, a number of evolutionary game theorists—notably Martin A. Nowak and Matthijs van Veelen—have criticized approaches to social evolution based on the Price equation, and have argued that evolutionary game dynamics provides a more appropriate foundation for social evolution theory (van Veelen 2005, 2009; van

Veelen et al. 2010, 2012; Nowak 2006a, b; Taylor and Nowak 2007; Nowak et al. 2010; Nowak and Highfield 2011). Kin selection theorists—notably Andy Gardner and Steven A. Frank—have replied to these criticisms on behalf of the Price approach (Gardner et al. 2011; Frank 2012).

The differences between the two formalisms are genuine and deep. In particular, it is extremely difficult to formalize the notion of inclusive fitness in any general way within evolutionary game dynamics (Nowak et al. 2010), though it is easy to do so in the Price formalism (Frank 1998; Gardner et al. 2011; Queller 2011; see also Chapter 5). Moreover, while versions of Hamilton's rule formulated via the Price equation tend to hold with a high degree of generality, superficially similar versions couched in the language of evolutionary game dynamics only hold under very restrictive assumptions (Birch forthcoming). There is thus a battle for the heart of social evolution theory: a battle that does not concern any specific empirical issue, but rather concerns the appropriate mathematical formalism in which to address foundational questions about social evolution. The debate rages on, with no sign of an end to the standoff. This dissertation is not neutral in this debate: I employ the Price formalism, with little discussion of the game-theoretic alternative. But just as I hope a detailed examination of kin selection theory will serve as a precursor to a reconciliation between the theories of kin and multi-level selection, so I hope this work will similarly help to foster mutual understanding between kin selectionists and their critics in evolutionary game dynamics.

A social revolution?

The second main reason why kin selection theory seems especially ripe for philosophical examination is that it has, in recent years, come to occupy an increasingly central position in evolutionary theory as a whole. This shift can be traced to the recent upsurge of interest in the major transitions in evolution, triggered in large part by the pioneering work of John Maynard Smith and Eörs Szathmáry (1995). When, instead of taking the biological hierarchy for granted, we view the history of life as a series of episodes in which new,

higher-level individuals evolved from collectives of lower-level entities, we start to see apparently cooperative phenomena where we saw none before: we see cooperation among cells within multicellular organisms, among organelles within cells, among genes within a chromosome. As Andrew F. G. Bourke observes in his recent synthesis, *Principles of Social Evolution*:

> Social evolution has grown outwards from the study of the beehive and the baboon troop to embrace the entire sweep of biological organization. It claims as its subject matter not just the evolution of social systems narrowly defined, but the evolution of all forms of stable biological grouping, from genomes and eukaryotic unicells to multicellular organisms, animal societies, and interspecific mutualisms. (Bourke 2011, 7)

This has lead naturally to the thought that well known approaches to the evolution of cooperation—and kin selection theory in particular—might turn out to explain vastly more than they were originally intended to explain. We can again turn to Bourke for a very clear statement of this view:

> Hamilton's inclusive fitness theory (kin selection theory) provides a general theory of social evolution powerful and versatile enough to serve as the conceptual foundation for understanding the major transitions in evolution. (Bourke 2011, 27)

In a series of recent papers, Joan E. Strassmann and David C. Queller defend a similar line (Strassmann and Queller 2007, 2010; Queller and Strassmann 2009). They, like Bourke, see higher-level individuality as an essentially social phenomenon, and see kin selection as the key to understanding transitions in individuality. Moreover, they suggest that the successes of kin selection theory in explaining the behaviour of social insects provide strong evidence for this claim:

> [Multicellular] organisms are groupings of cells [...] how did they combine and make the transition to the unity of purpose of a single organism? Social insect groups can give us special insight into this question. We will argue that social insect colonies are much like organisms, and we will show how their unity of purpose can arise through kin selection. (Strassmann and Queller 2007, 8620)

In keeping with my emphasis on general and foundational issues, I do not explicitly discuss the application of kin selection theory to transitions in individuality in this dissertation, though I have done so in some detail elsewhere (Birch 2012a, b, 2013a). Even so, the potential for such an application remains in the background at many points in the discussion. The foundational questions highlighted in Section 1.3 are interesting in their own right, but they acquire a new urgency when we realize that they concern not merely the foundations of a theory about social behaviour in insects and mammals, but the foundations of a theory that may well hold the key to understanding how new levels of biological organization come into being.

TWO

Cooperation, Collaboration and Control

Before we can even begin to examine the evolution of cooperation, we first need a firm conceptual grip on the nature of the target phenomenon. In this chapter, I introduce and analyse a number of concepts foundational to behavioural ecology. In Section 2.1, I consider biological cooperation itself, and its relationship to the intuitive notion of cooperation as the joint pursuit of shared goals. I then consider the broader notion of social behaviour; the narrower notions of mutual benefit, altruism, selfishness and spite; and the subtle but pivotal notion of a behavioural strategy. In Section 2.2, I consider the concept of a cooperative task and survey the aspects of task structure that distinguish social complexity from mere sociality. In Section 2.3, I propose an account of genetic control in terms of systematic counterfactual dependence. I stress the need to separate debates about control from debates about manipulation and enforcement: I argue that, while these notions are easily conflated, they in fact refer to three quite different phenomena.

2.1 Biological cooperation: the very idea

2.1.1 Phenomenal cooperation

When we talk of human beings 'cooperating' with one another, we usually have in mind cases in which people work together in pursuit of shared goals. The members of a hunter-gatherer tribe cooperate when they hunt animals larger than any they could ever bring down alone. Drivers impeded by a snowdrift cooperate when they clear the snow from the

road. Prisoners cooperate with the police when they hand over useful information; they cooperate with each other when they withhold that information to protect their co-conspirators.

If human cases are our benchmark—our exemplars of cooperation against which others should be judged—then it is hard to see, on the face of it, why interactions among bees or termites or bacteria should ever warrant description in the same terms. After all, bees, termites and bacteria do not literally have internally-represented goals, interests or intentions; they do not literally take means to promote their ends. We might be willing to ascribe intentional, means-end agency to dolphins, chimpanzees and other cognitively sophisticated species, but insects and microbes seem well outside the sphere of creatures to which such properties might reasonably be attributed. When we start with human examples and work outwards, talk of cooperation in the natural world soon starts to sound like a dubious metaphor—a metaphor we might want to put in a drawer marked 'dangerously anthropomorphic' and keep well away from serious biology.[1]

But that would be much too quick. The following three vignettes are chosen to illustrate the fact that, while the language in which we talk about biological cooperation may often seem metaphorical and anthropomorphic, to take this as any kind of indication that biological cooperation itself is somehow unreal or unimportant—that it is something serious biology should seek to explain away, rather than seek to explain—would be a serious error. For, while 'biological cooperation' is a metaphor, it is not just a metaphor. It is also the best name we can come up with for a real and astonishing natural phenomenon—a feature of the natural world which, like lightning and auroras and rainbows, simply cries out for explanation wherever we find it.

---

[1] Controversies over the merits and dangers of anthropomorphism are, of course, a mainstay of animal cognition research and its philosophy (see, e.g., Bekoff and Jamieson 1995; Mitchell et al. 1997; Bekoff et al. 2002). I will not weigh into these debates here.

Empire of the leafcutters

The eusocial Hymenoptera provide some of the best known and most celebrated instances of cooperation in nature, and perhaps the most remarkable of all are the leafcutter ants of the genera *Atta* and *Acromyrmex* (Hölldobler and Wilson 2009, 2011). As their common name suggests, the leafcutters specialize in cutting and retrieving fragments of leaves—a task they undertake with great efficiency and precision (Figure 1.1B) —but this is only part of the story. For the leaves are not food for the leafcutters, nor are they of any use in constructing the underground megalopolis in which they live. Instead, the ants use the leaves to cultivate subterranean fungus gardens,[2] stocked with a special fungal cultivar passed from one generation to the next. Farming the fungus is a joint endeavour on a colossal scale: the ants plant the fungi in purpose-built chambers, spray it with growth hormones, protect it against parasites and other fungal strains, and supply it with appropriate food. Without the coordinated contributions of vast numbers of workers (leafcutter colonies often number in the millions; see Hölldobler and Wilson 2009), the fungus could never be cultivated in sufficient volumes. The relationship is one of nature's great mutualisms: the ants rely on thriving fungus gardens to provide their larvae with food, while the fungus relies on the steady stream of leaf matter brought by the ants from the world outside.

Microbial towers

While the best known examples of biological cooperation concern multicellular organisms—specifically, vertebrates and social insects— 'best known' does not imply 'most common' or 'most important'. One of the most significant lessons from the last few decades of research in microbiology is that cooperative phenomena are extremely widespread in the microbial world, and that the feats of cooperation performed by microbes need not be any less spectacular than those undertaken by larger and more

[2] A trait the leafcutters share with several other ant genera and the termite genus *Macrotermes*; see Hölldobler and Wilson 2009.

familiar creatures (Crespi 2001; West et al. 2007b). A useful model organism for the growing field of sociomicrobiology is the social amoeba, or slime mould, *Dictyostelium discoideum* (Bonner 1959; Strassmann et al. 2000; Strassmann and Queller 2011). For much of their life cycle, these amoebae conform to our usual expectations of amoebae: they live in the soil, they engulf bacteria, they divide mitotically. When food gets scarce, however, things get interesting: if the amoebae are present in sufficient density (they detect the density of nearby conspecifics through a form of signalling known as quorum sensing; see Waters and Bassler 2005; Williams et al. 2007), the starving amoebae will aggregate, forming a mobile 'slug'. The slug moves as one—and moves further and faster than any individual amoeba ever would—in the direction of heat and light. On reaching a favourable location, the slug stops and begins to transform into a fruiting body (Figure 1.1A). Around a fifth of the amoebae sacrifice their lives in this process, forming a hardy, cellulose stalk of dead cells. The remaining four fifths cluster at the tip of the stalk, where they generate and release spores. The spores are dispersed through the environment, reducing the probability that the amoebae they ultimately produce will encounter the same harsh conditions suffered by their parents. The generation of fruiting bodies through the aggregation of previously separate cells is by no means unique to *D. discoideum*, nor even to amoebae: similar behaviour has also been observed in the social bacterium *Myxococcus xanthus* (Velicer and Vos 2009).

Sperm cells swim together

Some of the most striking examples of apparently cooperative phenomena occur not *between* organisms, but *within* them: almost wherever we look, we find cells interacting in ways which make ascriptions of common purpose difficult to resist (Queller 1997; Queller and Strassmann 2009; Strassmann and Queller 2010, 2011). Sperm cells provide some particularly memorable examples. We tend to imagine sperm as solitary swimmers, competing with one another to fertilize an egg. In the case of human sperm, this is more or less correct, but the picture changes when we consider species in which females mate with multiple males in quick succession. In these cases, the closely related sperm of a particular

male stand a greater chance of winning the race against the unrelated sperm of rival males if they work together; as a result, we often find that selection has favoured cooperation within groups of sperm. For example, in the American opossum, *Monodelphis domesticus*, sperm swim together in pairs, touching at the head: an arrangement which enables faster and straighter swimming (Figure 2.1C; Moore and Moore 1995; Moore and Taggart 2002; Pizzari and Foster 2008). In rodents such as the Norway rat (*Rattus norvegicus*) and the wood mouse (*Apodemus sylvaticus*), we see even more dramatic feats of sperm organization. The sperm use tiny hooks on their heads to latch together into balls, and propel themselves forward with aligned and synchronized beating of their tails (Moore et al. 2002; Immler et al. 2007; Pizzari and Foster 2008).

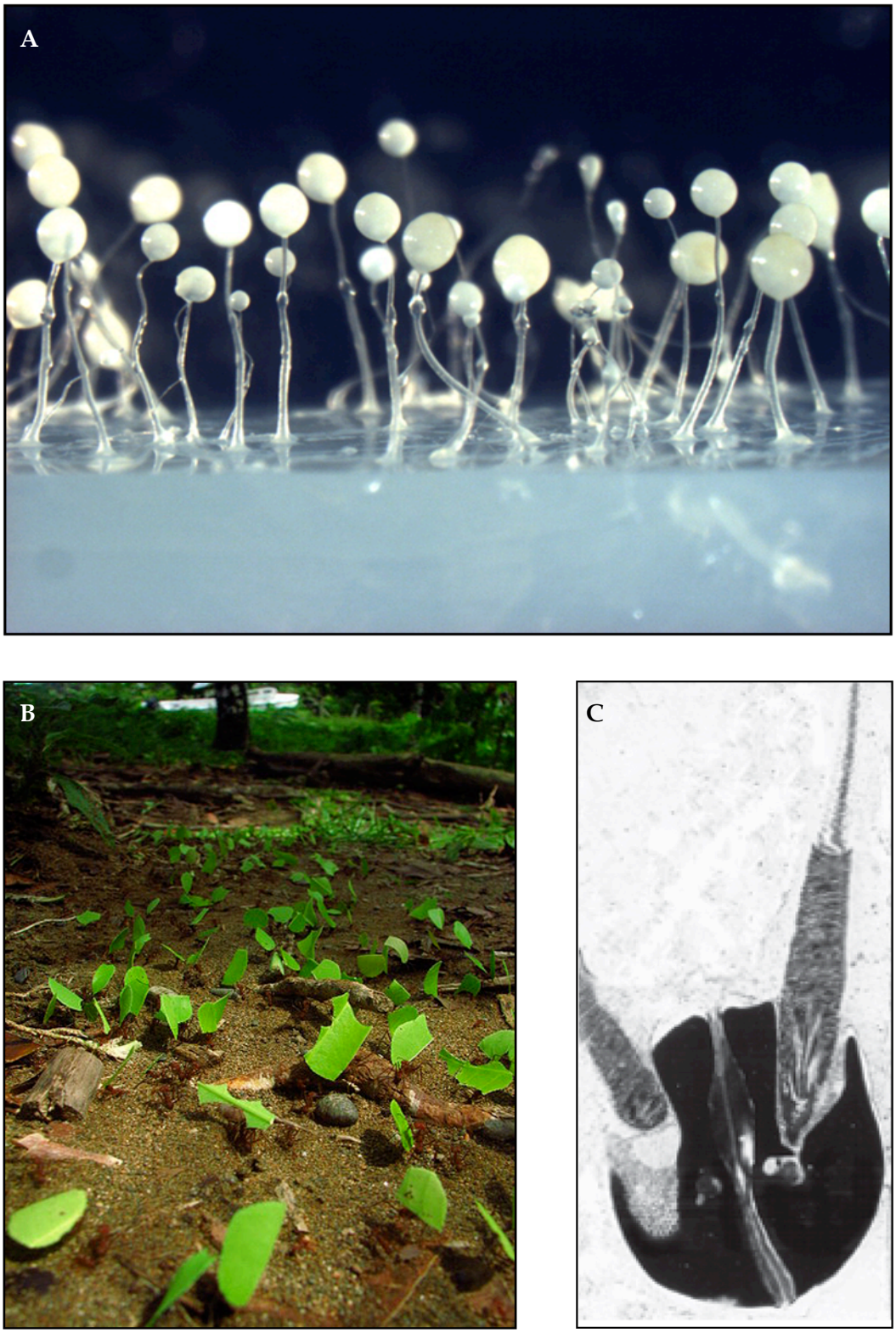


Figure 2.1: Phenomenal cooperation: (A) Fruiting bodies of the social amoeba *Dictyostelium discoideum* (photograph by J. E. Strassmann and D. C. Queller); (B) Workers of the leafcutter ant species *Atta colombica* in action (photograph by bandwagonman at en.wikipedia); (C) The sperm of the American opossum (*Monodelphis domesticus*) align in pairs for more rapid swimming (electron micrograph by Harry Moore).

Three morals ring out from these cases, and from the last few decades of research in behavioural ecology: biological cooperation is *everywhere*, the reality of biological cooperation is *obvious*, and biological cooperation is *amazing*. Hence, while it is hard to deny that biological cooperation is, in some sense, a human projection—a feature we read into the world when we read purpose and agency into living systems—it is equally hard to deny that it is also, in some sense, an objective and unmistakable feature of the biological landscape. While sperm, amoebae and leafcutters may not pursue common goals in a strictly literal sense, it is an objective matter of fact that some of their activities strongly evoke the appearance of common purpose—and thereby invite descriptions in terms of cooperation and related notions—in a way that other natural phenomena do not. Compare, for example, the construction of a slime mould fruiting body with a shark eating a seal: the intuition that the former is a cooperative process while the latter is non-cooperative is so overwhelmingly strong that it would take a seriously impressive error-theory to persuade us that, in reality, the shark and the seal are cooperating while the amoebae are not.

I want to attempt to capture this intuitive, phenomenal sense of biological cooperation by means of the following characterization:

> **Biological cooperation (phenomenal sense):** Any phenomenon in which living entities *appear* to work together or help one another in pursuit of common goals, irrespective of whether or not they are genuinely capable of intentional action.

As a putative definition, this phenomenal characterization of cooperation has obvious defects: it is vague, it is anthropocentric, and it threatens to make cooperation a partly subjective matter—the product of our human tendency to read intention into some patterns of interaction and not others—rather than something objectively measurable. We therefore

have good reason to look for a more technical definition of biological cooperation for the purposes of behavioural ecology and social evolution theory: a definition framed in precise, objective, biologically-respectable terms.

This does not, however, make the preceding discussion of the phenomenal notion redundant. For one reasonable constraint on an adequate technical definition of biological cooperation is that it does justice to the phenomenal notion. Ideally, we would like biological-cooperation-in-the-technical-sense to have more or less the same extension as biological-cooperation-in-the-ordinary-sense: we would like the technical notion to capture at least all the paradigm cases of cooperative phenomena, and to exclude phenomena that we would never ordinarily describe as cooperative. There are two main reasons for imposing such a constraint. One is that, ideally, we would like it to be the case that social evolution theorists are able to communicate their findings to the public: we would therefore like it to be the case that, when experts and laypeople talk about cooperation in the natural world, they are at least talking about approximately the same thing. The other, related reason is that it stops us losing sight of the phenomena social evolution theory is supposed to explain. We care about the evolution of cooperation primarily because cooperative phenomena are amazing; and we build theories and models of the evolution of cooperation with the intention of explaining these phenomena. Yet this explanatory project will never succeed unless biological-cooperation-in-the-technical-sense has at least a reasonable degree of overlap with biological-cooperation-in-the-ordinary-sense. If we want to arrive at the explanations that really mean something, we need to make sure that the explanatory targets of our theories and models actually resemble the phenomena we care about.[3]

[3] I suspect that this is the right way to think about the relationship between intuitive and scientific concepts in many notorious problem cases (e.g., life, organism, species, function, design). On the one hand, a good fit with ordinary intuition cannot be the sole criterion for a useful scientific definition. On the other hand, it seems ill advised to resort to stipulative definitions that run roughshod over the ordinary concept they are

2.1.2 Social behaviour as a fitness transaction

Though cooperative phenomena provide social evolution theory with its most spectacular explanatory targets, it has long been recognized that cooperation is not the only way in which organisms can interact in seemingly social ways. For this reason, it is common to see cooperation introduced as a particularly interesting subset of a broader class of social behaviours. The tendency is for social evolution theorists first to formulate a technical definition of social behaviour, and then to characterize cooperation by augmenting that definition with further conditions (see, e.g., Hamilton 1964; Trivers 1985; Bourke and Franks 1995; West et al. 2007a; Bourke 2011).

How, then, should we conceptualize social behaviour? We can turn to the Oxford sociobiologists Stuart West, Ashleigh Griffin and Andy Gardner (2007a) for a typical recent example:

> From an evolutionary point of view, a behaviour is social if it has fitness consequences for both the individual that performs the behaviour (the actor) and another individual (the recipient) (West et al. 2007a, 418).

At first glance, the West et al. definition may seem to miss out something important: we get a definition of what makes a behaviour social, but no attempt is made to define behaviour itself. In fact, this omission is representative of a general trend in theoretical definitions of social behaviour: the notion of behaviour is usually taken as primitive by behavioural ecologists, and the purpose of a technical definition of social behaviour is taken to be that of delineating, among behaviours, which are social and which are not.

---

intended to define; for if we do this, we risk losing sight of the phenomenon we originally wanted to explain. We need to find a middle way between these extremes (cf. Lewens 2004, Chapter 1).

Taking behaviour as primitive in this way may seem questionable. After all, the notion of behaviour is arguably no less anthropomorphic—and no less in need of a naturalized, technical definition—than the qualifier 'social'. Both normally connote intentional action in the human case, and both extend only metaphorically to non-intentional systems. In defence of West et al., however, there is probably a fairly innocuous explanation for their reluctance to define behaviour: the term tends to be used in such a broad and inclusive sense in behavioural ecology that the only way to capture its extension would be to define it as any interaction between an organism and its environment (including other organisms) that it is helpful to individuate for some theoretical or experimental purpose. The interesting classificatory work is done not in deciding whether or not some interaction counts as a behaviour, but rather in drawing biologically-salient distinctions among types of behaviour. The distinction between social and non-social behaviours is one particularly important distinction.

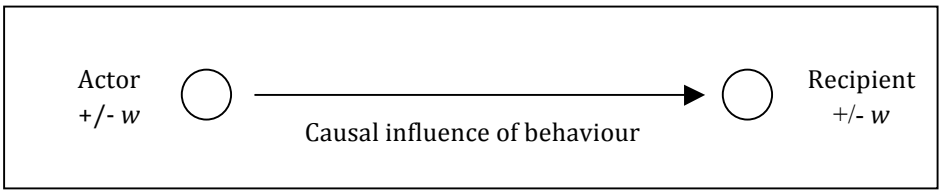


Figure 2.2: Social behaviour as a fitness transaction.

The general picture embodied in the West et al. definition is of social (as opposed to non-social) behaviour as involving a fitness transaction between two individuals, an actor and a recipient. The thought is that, just as a financial transaction involves crediting or debiting one account in order to credit or debit another, a social behaviour involves crediting or debiting the actor's reproductive fitness so as to credit or debit that of the recipient (Figure 2.2). The analogy with financial transactions should not be overstated, however: one important disanalogy is that, in the case of social behaviour, an increase or reduction in the fitness of the recipient need not be counterbalanced by an equal and opposite

reduction or increase in the fitness of the actor. Indeed, the fitness consequences for actor and recipient need not be opposite in sign at all: a social behaviour may be beneficial—or deleterious—to both actor and recipient.

Three further aspects of the fitness transaction conception of social behaviour are worthy of comment. First, note that, on this conception, social behaviour is essentially relational: if we were to remove the recipient, a social behaviour would no longer be social. This may seem obvious, but in some ways it can be counterintuitive: one might suppose that if (for example) a meerkat's alarm call counts as a social behaviour when performed in the wild, it would also count as a social behaviour when performed by a solitary meerkat in the laboratory. But if no other meerkats are present, the behaviour is not social. Second, note that the relation that matters is causation: the actor and recipient must interact causally. If two individuals never interact with one another, then they cannot behave socially with respect to one another, regardless of the other ways in which they may be related (e.g., genealogically or geographically). Third, the relevant causal interactions can involve more than two individuals. The picture in Figure 2.2 (of a fitness transaction involving a single actor and a single recipient) is far too simple to capture many of the most interesting examples of sociality in nature, which often involve multiple actors working in concert, and may implicate multiple recipients too (see Section 2.2).

2.1.3 The four-part schema

Viewing social behaviour as a fitness transaction between an actor and a recipient leads naturally to the thought that we can classify different types of social behaviour in terms of the sign of their fitness effects on the affected parties. The result is a four-part schema, first introduced by W. D. Hamilton (1964), that categorizes social behaviours as mutually beneficial, selfish, altruistic or spiteful (Table 1.1). These categories are easiest to apply when individuals interact in pairs, but note that they can, in principle, apply to interactions involving any number of agents: when more than two individuals partake in a

social interaction, we simply need to average the fitness effects over all actors and all recipients.

Effect on recipient → Effect on actor ↓	+	-
+	**MUTUAL BENEFIT**	**SELFISHNESS**
-	**ALTRUISM**	**SPITE**

Table 2.1: A traditional taxonomy by fitness effects classifies social behaviours as mutually beneficial, selfish, altruistic or spiteful. These are intended as technical terms: the usual psychological connotations of these notions do not apply.

The four-part schema assumes that any social behaviour will effect a positive or negative change in the fitness of both the actor and recipient. In principle, however, we could relax this condition: in principle, organisms could interact in ways that are neutral with respect to actor fitness, recipient fitness or both. This leads to a nine-part schema, with five additional possibilities (see Table 2.2, which also suggests names for the new possibilities).

Effect on recipient → Effect on actor ↓	+	0	-
+	**MUTUAL BENEFIT**	OTHER-NEUTRAL SELFISHNESS	**(OTHER-HARMING) SELFISHNESS**
0	SELF-NEUTRAL BENEFIT	NEUTRAL INTERACTION	SELF-NEUTRAL SPITE
-	**ALTRUISM**	OTHER-NEUTRAL SELF-HARM	**(SELF-HARMING) SPITE**

Table 2.2: A more comprehensive taxonomy by fitness effects allows for partially or wholly fitness-neutral behaviours. The evolutionary significance of fitness-neutral behaviours is an open question.

One might object to this extended taxonomy on the grounds that, while the notion of an actor-neutral social behaviour seems intuitive enough, a behaviour that is fitness-neutral with respect to the alleged recipient does not sound much like a social behaviour at all, because the recipient is not a recipient of anything. This would entitle us to strike out the middle column of the table on conceptual grounds alone, leaving behind a six-part schema. Is this a reasonable restriction? Ultimately, the decision turns on how broad we want the explanatory domain of social evolution theory to be. Suppose, for instance, that a parasite selectively feeds off the by-products of its host in a way that makes no significant difference to the host's fitness. We might well want to analyse the evolution of this host-parasite relationship using social evolution theory (cf. Frank 1998); and, if we take this route, it will be helpful to have a term to describe the general kind of interaction in which the host and parasite are engaged. Naturally, we might equally insist that interactions of this sort fall outside the scope of social evolution theory—that they are not *truly* social—and on these grounds deny the need for any extended taxonomy. This option is always available; the point here is merely that, if we want to extend the explanatory domain of social evolution theory to encompass as many phenomena as possible (cf. Chapter 1), we

should be prepared to embrace a similarly extended conception of the nature of social behaviour. This will be a recurring theme in this chapter.

Of the nine-part schema, the 'neutral interaction' box appears to be by far the least interesting. Even if we grant that social behaviours can be recipient-neutral, is there any reason to grant that social behaviours can be fitness-neutral with respect to both recipients and actors? After all, this stretches the analogy with monetary transactions to breaking point: no one transfers sums of £0.00 from one account to another. Here, I can only reply that I see no good reason to exclude such interactions from the scope of social evolution theory *a priori*. It is true enough that a wholly fitness-neutral interaction will not be directly targeted by natural selection, but fitness-neutral interactions may yet be explained by virtue of their non-causal correlations with behaviours that do affect fitness, or by the fitness effects they may once have had in the population's selection history.

2.1.4 Cooperation and the 'selected for' criterion

The four-part schema classifies social behaviours as selfish, spiteful, altruistic or mutually beneficial according to the sign of their fitness effects; the nine-part schema extends this to incorporate fitness-neutral behaviours. But where does cooperation fit into this picture? Traditionally, cooperation is defined as any social behaviour that confers a positive fitness benefit on a recipient, regardless of the sign of its fitness effects on the actor (cf. Hamilton 1964; Trivers 1985; Bourke and Franks 1995). In the four-part schema, this corresponds to any social behaviour that falls within the altruistic or mutually beneficial boxes; in the nine-part schema, it corresponds to any behaviour that falls within the first column. West, Griffin and Gardner (2007a) depart from this tradition by proposing a slightly more restrictive definition. For West et al., a behaviour that confers a benefit on a recipient does not count as genuinely cooperative unless it has at some point been favoured by selection in virtue of the benefit it confers:

> **Cooperation:** A behaviour which provides a benefit to another individual (recipient), and which is selected for because of its beneficial effect on the recipient (West et al. 2007b, p. 419).

The motivation behind this additional 'selected for' criterion is that, without it, our definition risks including behaviour that merely confers a fortuitous benefit on another organism, and that would therefore not be regarded as cooperation in the phenomenal sense. Suppose, for instance, that an elephant confers a fortuitous benefit on a nearby dung beetle by producing dung in its vicinity. We would not intuitively describe the elephant as cooperating with the dung beetle, nor would we consider them to be engaged in any kind of social interaction: we would sooner say that the dung beetle is merely exploiting a by-product of the elephant's digestive system (West et al. 2007b, 419). The selected for criterion preserves our intuitions here, since, to the best of our knowledge, the elephant's tendency to produce dung was not favoured by selection in virtue of the benefit it confers on nearby dung beetles.

The downside of this additional criterion is that, in making cooperation conceptually dependent on past selection, it rules out *a priori* the possibility of cooperative behaviours that are explained by something other than natural selection. The problem here is not so much that this overstates the importance of natural selection in explaining the evolution of cooperation, since there is no question that selection is extremely important; the problem is that, in bringing natural selection into the very definition of cooperation, we turn the undeniably close empirical connection between selection and cooperation into a definitional stipulation. There are two main reasons why this is best avoided. First, we typically want to be able to classify social behaviours prior to an investigation of the processes through which they have evolved; yet, on the West et al. definition, there is no way for us to know whether or not a social behaviour is genuinely cooperative without first knowing something about its selection history. Second, the discovery that natural selection explains biological cooperation surely ought to rank as an epistemic

achievement—a breakthrough that it took serious scientific work to make. Yet, on the West et al. definition, it is trivial that selection explains cooperation. This seems wrong: it seems as though, if anything should count as a non-trivial breakthrough in evolutionary biology, this should. But it will count as a non-trivial breakthrough only if the explanandum (cooperation) and the explanans (selection) are conceptually distinct.[4] These considerations suggest that we would be well advised to drop the selected for criterion from the definition of cooperation.

With regard to the extension of the term, not a great deal hangs on whether or not the criterion is included, since in many cases it makes no practical difference: sociobiologists tend to focus their attention on social behaviours that confer fitness benefits and that have, in all probability, been selected in virtue of the benefits they confer; and these behaviours are likely to count as cooperative on either definition. On balance, my preference is for the traditional definition, without the selected for criterion; but we should undoubtedly acknowledge that, in counting cases of fortuitous benefit as cooperative, this definition occasionally yields counterintuitive results.

The general moral to draw from this discussion is that accounting for our intuitive, phenomenal conception of biological cooperation in naturalistic terms is harder than one might think. A purely 'forward-looking', effect-based account does not distinguish *bona fide* cooperation from fortuitous benefit, while a partially 'backward-looking' account that appeals to selection history makes it impossible to identify cooperative behaviours prior to an investigation of their evolutionary origins and renders trivial the claim that past selection explains cooperation. Neither fully accords with our pre-theoretical intuitions. This problem is by no means unique to the concept of cooperation. We find ourselves in a similar predicament with respect to many commonplace biological notions: notably,

[4] Lewens (2007b) makes a similar point in the context of adaptation. The argument that historical definitions of adaptation trivialize the claim that selection explains adaptation appears to have first been made by Daniel C. Fisher (1985).

function (Lewens 2004, 2007a; Allen et al. 1998; Buller 1999; Ariew et al. 2002), design (Lewens 2004; Allen et al. 1998; Buller 1999) and adaptation (Rose and Lauder 1996; Lewens 2007b; Gardner 2009). We may feel as though we grasp these concepts well enough, but producing a technical definition that does justice to the intuitive notion often proves surprisingly difficult. In the end, we have to choose among a range of stipulative sharpenings of a murky intuitive concept, all of which have occasionally counterintuitive consequences.

2.1.5 From behaviours to strategies

So far we have been concerned with characterizing and categorizing social behaviours. Yet particular behaviours are only rarely the immediate target of evolutionary explanations. More often than not, the explanatory target is a strategy, where a strategy is thought to in some sense underlie the totality of behaviours that an organism performs over its lifetime. Talk of strategies, though extremely widespread, is sometimes considered controversial. For some authors, such talk imputes to insects, bacteria and other social organisms a dubious capacity for planning and foresight (Kramer 1984); for others, there may be a concern that strategies, if they can be said to exist at all in organisms with very limited cognitive capacities, must somehow be 'programmed' into the genome; and that, while it is not too controversial to suggest that the genome 'codes for' the construction of RNA molecules and proteins, extending the reach of the programme to encompass social behaviour takes a seductive metaphor much too far (for discussion of the limits of the programming metaphor, see Godfrey-Smith 2000, 2007b).[5]

Social evolution theorists have typically sought to evade such concerns by characterizing the notion of a strategy in rather more minimalist terms. For example, John Maynard Smith (1982) offers the following definition:

[5] Similar concerns apply to the notion of control; see Section 1.3.

> A strategy is a behavioural phenotype; i.e., it is a specification of what an individual will do in any situation in which it may find itself. (Maynard Smith 1982, 10)

This looks reasonable enough at first blush, but on closer inspection it becomes apparent that the expressions on either side of the 'i.e.' are subtly in tension with one another. For a 'specification' of an individual's behaviour is most naturally regarded as a set of conditional statements of the form: 'if in context C, the agent performs behaviour B'. Yet if we simply identify a strategy with some such set of statements, strategies will no longer be behavioural phenotypes, because sets of conditional statements are not properties of individual organisms.

One way to escape this conceptual tangle is to distinguish strategies from strategy-descriptions. We can then say that strategies are indeed properties of individual organisms. More specifically, they are dispositional properties:

> **Strategy:** The complete set of an organism's behavioural dispositions (i.e., dispositions to perform some behaviour, B, in some environmental context, C).[6]

A strategy-description, meanwhile, is a set of subjunctive or indicative conditionals that tell us how an organism would or will behave under various hypothetical conditions:

> **Strategy-description:** A set of conditional statements of the form 'if in context C, the agent performs (or would perform)

[6] This is not to deny that an organism's behaviour in some contexts may be chancy, e.g., that it may be disposed in context C to perform behaviour B_1 50% of the time and behaviour B_2 50% of the time. The notion of a strategy is intended to accommodate probabilistic dispositions of this sort.

> behaviour *B*' that specifies the relevant parts of an agent's strategy in the context of formal modelling.

Clearly, a strategy-description can help us characterize or specify an organism's strategy, but it should not simply be identified with the strategy. The strategy is a dispositional property of the organism; the strategy-description is not.

Given this dispositional conception of a strategy, we might informally envision a strategy as a kind of programme that tells the organism what to do in all the different contexts to which it may be exposed over its lifetime. But although strategies are literally programmed rules for behaviour in computer simulations of social evolution, we need not take talk of programming quite so literally in real ecological contexts. We can limit ourselves to a purely dispositional characterization, leaving open the question of whether the relevant dispositions are in any sense programmed into the genome. It is a platitude to say that organisms are disposed to behave in certain ways in certain circumstances, but it is no platitude to say that these dispositions are grounded in a genetic programme. On my account, talk of strategies is committed only to the former claim; the latter is optional.

Note that, strictly speaking, the categories introduced in the previous sections classify social behaviours, not strategies. Therefore, strictly speaking, behaviours can be classed as cooperative or non-cooperative, or as altruistic, selfish, spiteful or mutually beneficial, but strategies cannot be so classified. Yet biologists, notably in the context of evolutionary game theory, routinely talk of altruistic strategies, selfish strategies and so on. Sometimes these terms are labels introduced stipulatively to denote particular strategies in an evolutionary game; but sometimes they are used in a looser sense, with the intention of describing the overall character of a behavioural strategy. I do not think we should outlaw such talk (indeed, I will often use it myself in what follows); however, we should recognize that it is informal, imprecise and potentially misleading. A strategy can encompass a great variety of behaviours under a great variety of conditions, and the same

strategy may lead to altruistic behaviours in some contexts and selfish behaviours in others. If a strategy disposes an organism to be unconditionally altruistic, that is, to perform altruistic behaviours in all circumstances, then we could wholeheartedly call it an altruistic strategy; but altruism in the real world rarely, if ever, works like this. Most of the strategies that we informally describe as altruistic are in fact only conditionally altruistic: they will dispose an organism to behave altruistically under some conditions, but not under all conditions.

There are two, related reasons why the notion of a strategy is valuable in behavioural ecology. One is that it gives us a property which will often be relatively stable over an organism's lifetime—a property that is typically unaffected by the countless environmental contingencies that affect an organism's manifest behaviour. Of course, an organism's behavioural dispositions can change over its lifetime (dogs, even old ones, can learn new tricks), but they will not change rapidly, over very short timescales, in the way that an organism's manifest behaviour often will. The other, related reason is that the strategy is more likely to remain under the organism's control, in the sense that it is less likely to be influenced by other organisms (see Section 1.3 for detailed discussion of the notion of control in behavioural ecology). An organism will often modify its manifest behaviour in response to the behaviour of other, nearby organisms, and in response to the signals and cues they emit. But its strategy is not so easily altered. Consider, for example, the case of the social amoeba *D. discoideum*, introduced in Section 1.1.1. Nearby cells can aggregate to form a mobile slug and eventually a fruiting body, but will do so only if (i) there is a sufficient number of cells in the same area, and (ii) they are starving. The aggregating behaviour is therefore highly sensitive to both ecological cues (the abundance of food) and social signals (which indicate whether or not a quorum can be reached). If a particular amoeba is never short of food, or never senses a quorum of other amoebae in its surroundings, it will never perform the aggregating behaviour. Yet the amoeba's strategy may not be sensitive to any of these variables. For it may be that the amoeba's *disposition* to

aggregate, if food is scarce and a quorum is present, is not affected by changes to its environment nor by the presence or absence of other amoebae.

2.2 Task-based cooperation

2.2.1 The reality of tasks

In Section 2.1.2, we noted that, while the notion of social behaviour is often introduced in the context of a simple interaction between two individuals, there is no reason in principle why a social phenomenon could not implicate multiple actors and multiple recipients. Crucially, this is no mere theoretical possibility: many actual instances of cooperation in nature take the form of large-scale collaborative tasks, in which the successful production of a fitness benefit requires multiple causal contributions from multiple actors (Anderson and McShea 2000; Anderson and Franks 2001; Anderson et al. 2001; Calcott 2006, 2008). Fruiting-body formation in *D. discoideum* is one excellent example of cooperation with this structure; fungal cultivation in the leafcutter ants is another (see Section 2.1.1). Some additional examples will help set the scene.

Heavy-lifting

We noted in Section 2.1.1 that many of the most vivid and remarkable displays of cooperation can be found in the eusocial Hymenoptera. It should come as no surprise, therefore, that they also provide some of the most impressive examples of collaborative tasks. While we could turn again to the leafcutters for numerous illustrations, an equally iconic case is that of the army ants *Eciton burchelli* and *Dorlyus wilverthi*, which form teams of two or more individuals (one at the front, the rest at the back) to retrieve prey items that are too heavy for any single ant to carry alone (Franks 1986, 1987; Anderson and Franks 2001). Indeed, these teams are 'superefficient', in the sense that they are able to carry items

heavier even than the sum of the weights each team member could carry by itself (Franks 1986; Anderson and Franks 2001).

Defending the group

A common form of cooperative task is that of group defence. Again, the eusocial insects provide compelling examples. Consider the joint action of members of the major (large) and minor (small) castes in the dimorphic ant *Pheidole pallidula*: when the colony is threatened by an intruder, the ants form teams comprising one major ant and many minor ants; the minors pin down the intruder so that the major can deploy its strong jaws to decapitate it (Detrain and Pasteels 1992; Anderson and Franks 2001). We see a similar form of defensive teamwork in a very different context: the mammalian immune system, in which dendritic cells (tasked with detecting antigens) signal to nearby memory cells to induce the release of appropriate antibodies. By coordinating in this process of 'cellular teamwork' (Ayres and Vance 2012), the cells are able to repel pathogenic intruders more effectively (Figure 2.3C; Ayres and Vance 2012; Kupz et al. 2012). Of course, on a traditional conception of the explanatory domain of sociobiology, an immune response would not be considered a social phenomenon at all, and it remains an open question how seriously we should take the many suggestive analogies between insect sociality and the organization of multicellular individuals. Here, I simply want to note that, if we do take seriously the suggestion that interactions among the cells within organisms and proto-organisms may usefully be regarded as cooperative, it is clear that much if not all of the cooperation in question will consist of cooperative tasks that implicate numerous cells in different roles.

Pack hunters

Task-based cooperation often yields rich rewards for predators: by working together in structured and organized ways, groups of predators are able to tackle bigger prey, or more prey or to predate more efficiently than they ever could alone. Examples include tribes of humans (*Homo sapiens*), troops of chimpanzees (*Pan troglodytes*), prides of lions (*Panthera*

leo) and packs of wolves (*Canus lupus*) (Anderson and Franks 2001). A particularly spectacular example is provided by pods of humpback whales (*Megaptera novaeangliae*), which occasionally deploy a tactic known as 'bubble net feeding'. A shoal of herring is located and driven upwards from the sea floor by a group of whales; then a separate whale swims around the fleeing shoal, encircling it with a curtain of bubbles. The herring will not swim through the curtain of air; instead, they continue to swim upwards towards the surface, where they are trapped and devoured by the chasing group (Figure 2.3A; Sharpe and Dill 1997; Anderson and Franks 2001). Pack hunting is by no means the sole preserve of vertebrates, however, and may not even be the sole preserve of multicellular organisms. Recent work on the social bacterium *Myxococcus xanthus* has revealed a mysterious behaviour in which the bacteria move collectively in a 'ripple' formation, like waves on the sea (Figure 2.3B; Berleman and Kirby 2009). There is good evidence that rippling is a predatory behaviour, triggered by the proximity of food, but the question remains open as to what predatory advantage, if any, it provides for the bacteria. One hypothesis is that the formation is a kind of battle tactic: by getting underneath prey colonies and performing this synchronized rippling motion, the myxobacteria are able to break down and disrupt the target colony more effectively, enabling its rapid destruction.

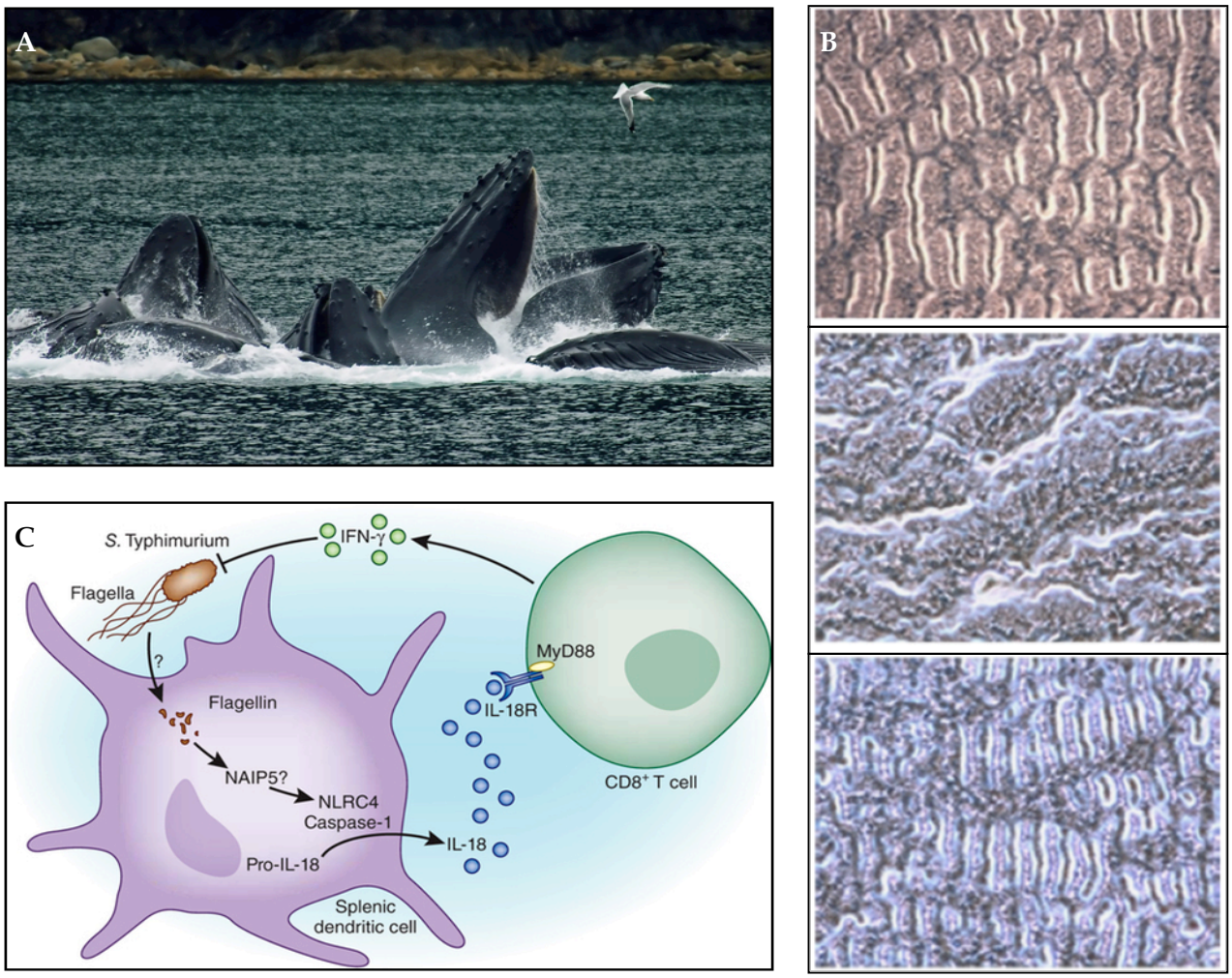


Figure 2.3: Some cooperative tasks: (A) 'Bubble net' feeding by the humpback whale, *Megaptera novaeangliae* (photograph by Evadb at en.wikipedia); (B) Predatory ripple formation in *Myxococcus xanthus* (photograph by J. E. Berleman and J. R. Kirby); (C) 'Cellular teamwork' in the mammalian immune system (drawing by Katie Vicari).

2.2.2 Tasks as mechanisms

We have now seen some particularly vivid examples, but what exactly *is* task-based cooperation? As with cooperation *simpliciter*, we usually know it when we see it, but that does not make it especially easy to characterize in general. In a series of important articles on the subject of collaborative tasks, Carl Anderson, Nigel R. Franks and Daniel W. McShea (Anderson and McShea 2000, Anderson and Franks 2001, Anderson et al. 2001) work with the following definition:

> A task is an item of work that potentially makes a positive contribution, however small, to inclusive fitness. (Anderson and Franks 2011, 534)

While it represents a reasonable first pass, the Anderson et al. definition has two main drawbacks. The first is that the emphasis on inclusive fitness rather than fitness seems questionable. While the only way to generate an inclusive fitness benefit for the actors is to generate a fitness benefit for the recipients, the converse is not true: a task may generate fitness benefits which fall on recipients genetically unrelated to the actors. To accommodate such cases, I suggest that we drop the 'inclusive' and take a positive contribution to the personal fitness of the recipients as the relevant fitness effect. The second is that the notion of an item of work which potentially contributes to fitness seems too broad to be of much use in individuating tasks, or in distinguishing a whole task from its component steps. Suppose a humpback whale produces a ring of bubbles around a shoal of fish. This is an item of work, and there is a sense in which it potentially contributes to fitness, in so far as it enables the humpback and its neighbours to feed when performed as part of the bubble net feeding procedure. Is it therefore a whole task in its own right? Or is it best regarded as part of a larger task, since the overall process of bubble net feeding, of

which it constitutes one step, is also an item of work which potentially contributes to fitness? The Anderson et al. definition does not help us settle such questions.

We can remedy this second weakness by drawing on an idea from Brett Calcott (2006): collaborative tasks may be regarded as a form of mechanism, in the sense of Peter Machamer, Lindley Darden and Carl Craver (2000). Machamer et al. characterize mechanisms in biology as '[composed of] entities and activities, organized such that they are productive of regular changes from start or set-up to finish or termination conditions' (2000, 3). Calcott's proposal is that cooperative tasks constitute a species of mechanism: specifically, they are mechanisms composed of entities and activities organized such that they are productive of regular fitness benefits.

This goes some way towards addressing concerns about task individuation: if tasks are mechanisms, then individuating tasks is a matter of individuating mechanisms; and, while there is no formal method or algorithm for the individuation of mechanisms, it is a feat biologists routinely accomplish in many different areas of biology. In the specific case of the humpback whales, we can see that, while bubble net feeding as a whole is plausibly regarded as a mechanism that produces regular fitness benefits, the component steps of that mechanism are only productive of fitness benefits if the other steps are also completed in the correct order. This provides us with principled grounds for regarding these steps as parts of tasks (or subtasks), rather than as tasks in their own right.

In light of these considerations, I will work with the following, modified characterization of a collaborative task:

> **Collaborative task:** A multi-step mechanism that regularly produces fitness benefits when each of its component activities is completed successfully and in the correct order.

2.2.3 *Aspects of task structure*

While some tasks may be completed by a single individual, many require multiple contributions in order to generate a fitness benefit (including all the examples introduced above). In such cases, the general causal picture is not the intuitive picture of a single actor conferring a benefit on a recipient: it is one of many actors collaborating to confer a benefit through task completion (Figure 2.4). The recipients may be the same individuals as the actors (as in the case of a task performed by a number of individuals for their mutual benefit) but they need not be; indeed, in the kinds of cases that will concern us here—cases in which the recipient is a queen or a germ cell—the recipients rarely participate in any tasks, and the actors rarely gain any personal fitness benefit from their efforts.

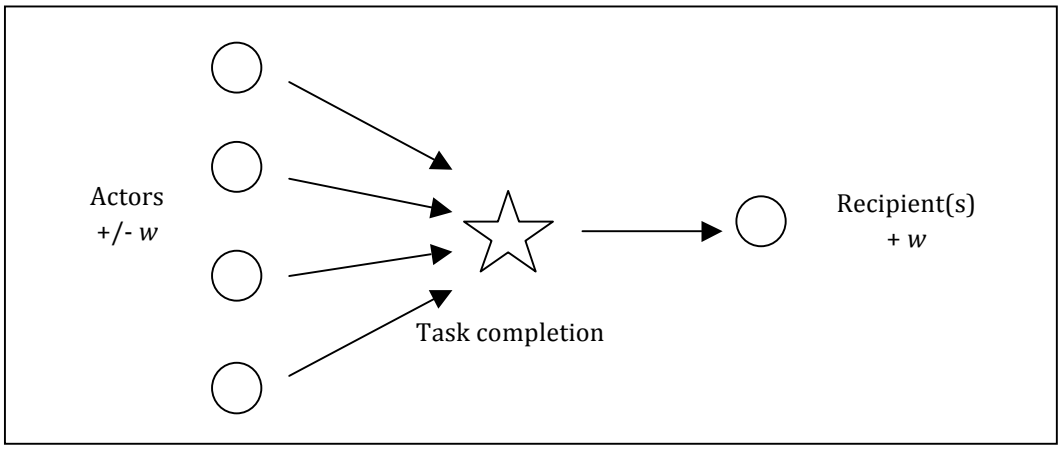


Figure 2.4: An idealized representation of task-based cooperation.

Note that, even on the highly-idealized picture of task-based cooperation shown in Figure 2.4, it is unclear how, if at all, we can resolve the benefit conferred on the recipient into discrete components contributed by each of the actors. Should we say that each actor contributed B/n, where B is the total benefit conferred and n is the number of actors? Or should we say that each actor contributed $B - B^*$, where B^* is the reduced benefit that would have been conferred if that actor had not participated? The former measure takes no account of the fact that some actors may make a greater contribution to the task than

others, while the latter measure allows that the total benefit conferred by the actors may differ from the total benefit received by the recipient.[7] Neither, therefore, is satisfactory. But if the total benefit of task completion cannot be resolved into discrete individual contributions, the pairwise fitness transaction model does not hold: task completion confers a benefit on the recipient that cannot be treated as a sum of the benefits conferred by the individual actors considered separately (Figure 2.5).

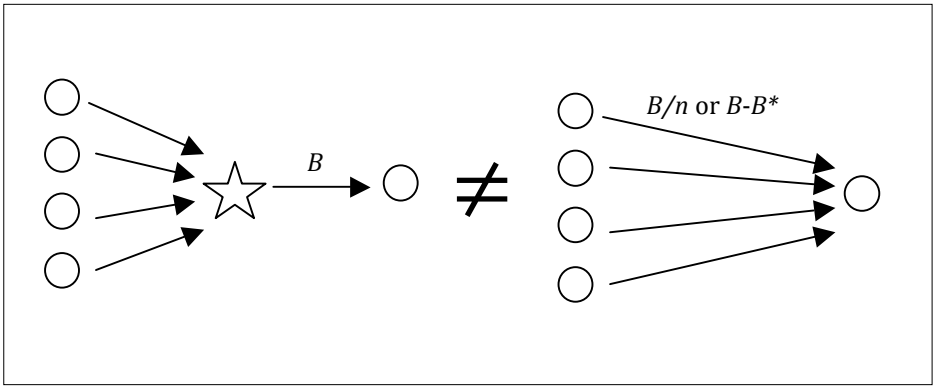


Figure 2.5: The benefit of task completion does not straightforwardly decompose into discrete, additive components contributed by the individual actors.

Of course, the failure of these two simple measures hardly shows that the overall benefit could not be resolved into discrete components attributable to the different actors: some more complicated measure may yet succeed where the simple measures fail (this is an issue I revisit in Chapter 5, Section 5.4.1, and in Appendix D). But it does show that decomposing the benefit of task completion is by no means a straightforward business: even in very simple cases, we can see how acknowledging the task structure of cooperation puts the pairwise fitness transaction model under strain. Moreover, even if we could find an acceptable means of decomposing the benefit into components attributable to individual actors, this procedure would belie the true causal structure of the mechanism by which the benefit was generated. For, causally speaking, the benefit generated by a

[7] Suppose, for instance, that every contribution is needed for the completion of the task, so that $B^* = 0$ and $B - B^* = B$. On this measure, the benefit conferred is nB, but the benefit received is only B.

collective task is not a sum of benefits separately generated by individual tasks: it is the product of a single mechanism in which all the relevant actors are participating entities.

I now want to consider four further features that add to the complexity of task-based cooperation. The list is not intended to be exhaustive (see Anderson and McShea 2001 for a more detailed synthesis). The purpose of this brief survey, in addition to that of introducing key concepts we will revisit in later chapters, is to highlight two general points. The first is that many of the organizational features of the most complex social groups are shared, in some form or another, by multicellular organisms. The second is that the same features are also exhibited, albeit to a much lesser degree, by many simpler animal societies. The overall picture is one in which the same broad types of complex social phenomena recur throughout the biological hierarchy, wherever groups of entities are bound into stable, integrated wholes.

Division of labour

Informal talk of division of labour is widespread in discussions of the major transitions in evolution (see, e.g., Maynard Smith and Szathmáry 1995, Godfrey-Smith 2009, Bourke 2011), but Anderson, Franks and McShea (2001) deploy the notion in a relatively technical sense. For Anderson et al., labour is divided when a task is split into more than one distinct subtask, where a subtask is an item of work that would not by itself confer a inclusive fitness benefit but that fulfils one of the necessary conditions for the completion of a larger task. Subtasks may themselves be divided into further subtasks, and so on. Anderson et al. provide no algorithm for the individuation of subtasks, but suggest that in practice the subtasks are often easily identified. For instance, they describe a grass harvesting task in *Hodotermes mossambicus*, where the workforce is visibly divided into cutters and transporters (Anderson et al. 2001, 645).

By conceptualizing division of labour in this way, we make the notion distinct from that of specialization (see below). Indeed, they are properties of different things: tasks are divided,

while workers are specialized. This conceptual distinction, though rarely drawn explicitly, is a helpful one, because the division of a task into subtasks may occur without the specialization of workers, and vice versa.

Specialization

Specialization is correlation between the properties of workers and the tasks they undertake.[8] It thus requires some form of differentiation among workers. In the eusocial Hymenoptera, two kinds of specialization predominate: specialization based on morphological differences (in which workers undertake different tasks depending on their physical characteristics) and specialization based on age differences (in which workers typically perform different tasks at different life stages). As Anderson and McShea (2001) note, however, some cases of specialization fall in between these categories. These are cases in which workers develop traits that enable them to perform a particular task at a particular life stage, only to lose those traits subsequently (they cite the short-lived production of royal jelly in honey bees, which leads to the temporary specialization of workers in feeding tasks). Polymorphism and age-based polyethism can thus blur into one another: we can do better by seeing these as extremes of a continuum of specialization based on developmental differences, ranging from superficial, short-lived differences in the simplest colonies to spectacular, life-long polymorphism in the most complex.

I want to reserve the term 'extreme specialization' for cases in which specialization is accompanied by a loss of behavioural totipotency—in other words, cases in which workers have lost the ability to undertake some or all tasks other than the task for which they are specialized. While eusocial societies with distinct morphological castes exhibit some degree of extreme specialization, multicellular organisms display this phenomenon to a far greater degree: consider, for example, a human red blood cell, which specializes so

[8] Because specialization may be regarded as a kind of correlation, we can quantify the overall degree of specialization in a social group using information theory (see Gorelick et al. 2004). It is thus perhaps the only aspect of social complexity for which a reasonably straightforward quantitative measure is available.

exclusively in oxygen transport that it lacks even a nucleus, a basic prerequisite for participation in most other tasks.

Germ-soma specialization occurs when some group members specialize in tasks which contribute to the growth and persistence of the collective, while others specialize in tasks which generate new collectives. Germ-soma specialization may be extreme, such that somatic specialists lose the capacity to generate new collectives, but it need not be (in plants, for instance, all cells in the floral meristem can potentially give rise to new individuals; see Clarke 2011). Owing to its consequences for within-group conflict, germ-soma specialization is often assigned special importance in accounts of evolutionary transitions (see, e.g., Buss 1987, Michod 2007, Godfrey-Smith 2009, Bourke 2011).

Coordination

Coordination is a feat of signalling and plasticity, and introduces yet more contingencies on which the success of a task may depend: when a task requires coordination, the subtasks must be performed at the right time and in the right order. As Anderson and Franks (2001) take pains to point out, while coordination presupposes a division of labour, it may not always require specialization: a task must be split into subtasks, but the workers who undertake the subtasks need not belong to different specialized castes.

Among coordinated tasks, Anderson and Franks distinguish partitioned tasks, in which the subtasks take place in a coordinated series, from team tasks, in which the coordinated subtasks occur concurrently. While partitioned tasks are fairly widespread in eusocial societies (particularly tasks which exhibit a 'bucket brigade' style organization; see Ratnieks and Anderson 1999), team tasks appear to be relatively rare. Anderson and Franks cite nest construction in *Oecophylla* weaver ants, prey retrieval in *Eciton burchelli* and *Dorylus wilverthi* (see above), and the decapitation of intruders in *Pheidole pallidula* (see above). They are far from rare in multicellular organisms, however, where teamwork is

rife: consider the cellular teamwork involved in an immune response (see above), or a coordinated muscle contraction, or the coordinated production of enzymes.

Redundancy

A workforce contains redundancy when there are more workers than are strictly needed for task completion. We see two broad kinds of redundancy in insect societies. The first sort (which I will call passive redundancy) occurs when there is a large reserve workforce, idle but ready to step in should any labour shortages arise. This phenomenon is widespread in eusocial societies (see Hölldobler and Wilson 1990, 342-343). The second (which I will call active redundancy) occurs when more workers actively undertake a task than are strictly necessary for its completion. We see this in the foraging strategies of complex ant societies: large numbers of ants search for food in parallel, then work in parallel to retrieve the food that one individual has found (see Oster and Wilson 1978, Herbers 1981). The upshot of redundancy in either form is that 'if one worker doesn't complete the task someone else will' (Oster and Wilson 1978; see also Section 3). We see a clear analogue of this phenomenon in multicellular organisms, where the number of cells that specialize in a given task often dramatically exceeds the minimum required for task completion. To take a particularly extreme example, the human circulatory system can stand to lose one eighth of its total stock of red blood cells during a routine blood donation without any significant adverse effects.

We should, I think, take care to distinguish what I have called redundancy (following Anderson and McShea 2001, and Hölldobler and Wilson 1990) from a very different phenomenon to which the same name has been applied. Andrew F. G. Bourke and Nigel R. Franks (1995, 440) contrast what they term the redundancy of parts (that is, the existence of surplus workers, which I am calling simply redundancy) with what they term the 'redundancy of functions'. By redundancy of functions, they mean an individual worker's latent capacity to undertake tasks that they are never called upon to perform during their lifetime. I will call this phenomenon latent versatility.

2.3 Cooperation and control

2.3.1 Questions of control

I turn now to a third family of conceptual issues. Considerations of control are often central to empirical debates in behavioural ecology. Yet these debates typically proceed in the absence of any strong grip on the meaning of control, or of associated concepts. The result is ample scope for ambiguity and semantic confusion. Two examples will serve to illustrate the point.

Pheromones: manipulation or honest signalling?

There is good evidence that queens in many insect societies produce chemicals—*pheromones*—which influence the reproductive behaviour of workers in such a way as to promote worker sterility (see, e.g, Grozinger et al. 2003; Holman et al. 2010). This has given rise to the thought that pheromones provide a means by which the queen 'manipulates' or 'controls' the behaviour of workers. The general idea has been around for a long time (see, e.g., Wilson 1971; Alexander 1974; Fletcher and Ross 1985; Hölldobler and Wilson 1990) and a version of it has recently been revived in a controversial article by Martin Nowak, Corina Tarnita and E. O. Wilson (2010), who assert (among many other bold and provocative claims) that workers may be regarded as 'the extrasomatic projection of [the queen's] personal genome' (2010, 1061) and go on to claim that '[t]he workers can be seen as "robots" that are built by the queen. They are part of the queen's strategy for reproduction' (2010, 38 (supplementary information)) .

The best-known critique of the notion that queens control their workers at a distance is that of Laurent Keller and Peter Nonacs (1993), who argue that:

> [P]heromonal queen control, defined as manipulation and control through chemical production alone, has never been

> conclusively shown to exist and is unlikely to have evolved [...] True queen control, we argue, is likely to be found only when direct, physical aggression against all subordinates is possible with pheromones serving as honest signals rather than as controlling substances. (Keller and Nonacs 1993, 788)

Note that Keller and Nonacs are not disputing that the queen emits pheromones which influence the behaviour of workers. The debate is not over the reality of pheromones, but rather concerns whether these pheromones can be said to give the queen control over the behaviour of the workers. The alternative favoured by Keller and Nonacs is that pheromones are best regarded as 'honest signals' which indicate the location and status of the queen, and that the workers adaptively adjust their behaviours to these signals without thereby granting her control of those behaviours. While this issue turns in part on empirical considerations, it also turns on what is meant by 'control'. Indeed, without some grip on the meaning of control in this context, we will not even be able to make sense of what the debate is about.

Altruism in insect societies: voluntary or enforced?

In a recent review, entomologists Francis Ratnieks and Tom Wenseleers (2008) explore similar themes. They set out to challenge the idea that, in social insect colonies, altruistic behaviours represent a voluntary sacrifice on the part of the workers, who take a hit to their direct fitness in order to boost their indirect fitness through relatives:

> [I]t is normally assumed that whether an individual is altruistic is under the control of the individual itself, that is altruism is voluntary and not socially enforced. But is this true for social insect altruism? As we argue here, in many cases it is not. (Ratnieks and Wenseleers 2008, 45)

The main contrast with Keller and Nonacs discussion is that, when Ratnieks and Wenseleers talk of social enforcement, they are not talking primarily about manipulation of workers by the queen. Rather, they are referring to apparently coercive or enforcing behaviours that workers inflict on one another. They piece together an impressive body of evidence indicating that such behaviours are widespread in insect societies, 'ranging from the killing of worker laid eggs to preventing larvae from developing into queens via food control' (2008, 45).

As with the question of whether pheromones are a form of signalling or a form of manipulation, the question of whether altruism is voluntary or enforced in the social insects turns on conceptual as well as empirical issues. In addition to reinforcing the need for a firmer grasp on the notion of control (what exactly does it mean, for example, to assert that 'whether an individual is altruistic is under the control of the individual itself'?), this case also highlights the importance of mapping out the conceptual relationships between control and other associated concepts, such as enforcement, volunteering and manipulation. Compare, for instance, the policing of egg-laying with the withholding of food from larvae at an early stage in development. In some sense, both types of behaviour intuitively reduce the control an insect has over its own fate; one may therefore be tempted to group them under a general heading of coercive behaviours and downplay the differences. But the ways in which these behaviours deprive workers of control are very different. In the former case, an adult individual loses its capacity to produce viable offspring, though it may retain control of other aspects of its behavioural phenotype; in the latter case, a developing individual loses control of its own developmental fate, though it may yet retain control of the behaviours it performs as an adult. Moreover, we should contrast both these cases with one in which an individual directly controls the behavioural phenotype of another, as exemplified by the pheromonal control hypothesis considered and rejected by Keller and Nonacs.

2.3.2 Control as systematic counterfactual dependence

The widespread use of 'control' and associated notions in behavioural ecology may give rise to two related worries. One is familiar from the foregoing discussions of other central notions in the study of social behaviour: it is the worry that control talk is problematically anthropomorphic—that, in talking of organisms such as ants, wasps, bees and bacteria controlling their own actions, manipulating one another, volunteering in cooperative endeavours, or enforcing each other's compliance with social norms, we impute to them a level of cognitive sophistication they cannot seriously be considered to possess.

The second worry is that, at bottom, control talk is implicitly committed to the view that social behaviour is in some sense genetically programmed. The thought is that, even if we grant that biologists are not literally imputing sophisticated forms of intentional agency to insects and bacteria, they are assuming a strong analogy between and the role of the mind in determining human action and the role of the genome in determining social behaviour—that to say an organism controls a particular behaviour is to say, in broad terms, that the behaviour is 'in its genes', or 'programmed into the genome' in the same way that a future course of action is represented in the mind. This genetic programme metaphor is controversial (cf. Section 2.1.5), and building a commitment to its validity into the conceptual foundations of control talk would render such talk equally controversial.

In my view, both concerns are misplaced. For I contend that we can explicate the notion of genetic control, at least as the term is used in behavioural ecology, without appealing to the notion of a genetic programme. We can do this by characterizing control in purely counterfactual terms. In short: genetic control of phenotype is a matter of systematic counterfactual dependence of phenotypes on genotypes. In talking of 'systematic

counterfactual dependence', I have in mind David Lewis's (1973) notion, articulated in the following passage:[9]

> Let A_1, A_2, ... be a family of possible propositions, no two of which are compossible; let C_1, C_2, ... be another such family (or equal size). Then if all the counterfactuals $A_1 \,\Box\!\!\rightarrow C_1$, $A_2 \,\Box\!\!\rightarrow C_2$, ...[10] between corresponding propositions in the two families are true, we shall say that the C's *depend counterfactually* on the A's. We can say it like this in ordinary language: whether C_1 or C_2 or ... depends (counterfactually) on whether A_1 or A_2 or ... . Counterfactual dependence between large families of alternatives is characteristic of processes of measurement, perception, or control. (Lewis 1973, 561; his italics and ellipses)

Similar ideas have appeared in recent discussions of 'causal specificity' in biology (Sarkar 2005; Weber 2006; Waters 2007; Woodward 2010; Stegmann forthcoming). A particularly notable example is that of James Woodward (2010), who disambiguates two senses in which the term 'causal specificity' is used in biology and the philosophy of biology. In some contexts, Woodward notes, talk of causal specificity connotes that certain kinds of effect have very characteristic causes, allowing reliable inferences from effect to cause. In other contexts, however, causal specificity has more to do with the 'fine-grained influence' of one variable over another. Woodward suggests that the influence sense of causal specificity can be glossed in terms of a characteristic pattern of counterfactual dependence:

---

[9] Lewis's (1973) paper is mostly remembered for its counterfactual account of singular causation (i.e., the causation of one event by another), but it also contains an account of the systematic counterfactual dependence of one family of propositions on other. This appears earlier in the paper and is often neglected, but it is very relevant to questions of control. Here I employ Lewis's account of systematic counterfactual dependence without endorsing or employing his account of singular causation.

[10] In Lewis's 'box arrow' notation, $A_1 \,\Box\!\!\rightarrow C_1$ denotes the counterfactual conditional 'if A_1 were to obtain, C_1 would obtain'.

> There are a number of different possible states of C (c_1 ... c_n), a number of different possible states of E (e_1 ... e_n) and a mapping F from C to E such that for many states of C each such state has a unique image under F in E (that is, F is a function or close to it, so that the same state of C is not associated with a different state of E, either on the same or different occasions), not too many different states of C are mapped on to the same state of E and most states of E are the image under F or some state of C. This mapping F should describe patterns of counterfactual dependency between states of C and states of E that support interventionist counterfactuals. Variations in the time and place of occurrence of the various states of E should similarly depend on variations in the time and place of occurrence of states of C. (Woodward 2010, 305)

Woodward cites Lewis's (2000) notion of influence as a forerunner; in my view, however, Lewis's (1973) notion of systematic counterfactual dependence represents an earlier and closer precursor.[11] There are two technical differences between Lewis's notion of systematic counterfactual dependence and Woodward's notion of causal specificity: Lewis's talk of propositions as the constituents of the relevant counterfactuals gives way to Woodward's talk of states of variables, while Lewis's simple counterfactuals relating two families of propositions (of the form: 'if proposition A_1 were to obtain, proposition C_1

[11] Roughly speaking, on Lewis's account an event, C, 'influences' another event, E, when a family of alternatives representing slight variants of E depends counterfactually on a corresponding family of alternatives representing slight variants of C: 'we have a pattern of counterfactual dependence of whether, when and how on whether, when and how' (2000, 190). Note, however, that this is an account of singular causation (intended as a replacement for his 1973 account; see footnote 9), not an account of control or causal specificity. Lewis's earlier notion of systematic counterfactual dependence (on which his 2000 notion of influence is parasitic) is more directly relevant to questions of control, though it is neglected by Woodward.

would obtain') are replaced by Woodward's interventionist counterfactuals (of the form: 'if we were to intervene on C to bring about c_1, then E would adopt the value e_1'). While these differences are not merely superficial,[12] they are minor enough that a preference for Lewis over Woodward, or vice versa, will make little difference for current purposes. I will talk of systematic counterfactual dependence rather than causal specificity, but substituting the latter for the former would not imperil the claims I want to make.

As Woodward notes, it is plausible that the structure of an organism's proteins and RNA molecules depends counterfactually, in a reasonably fine-grained way, on the nucleotide base sequence in its DNA—and that this at least partly captures the thought that the DNA controls the synthesis of RNAs and proteins: [13]

> [T]here are many possible states of the DNA sequence and many (although not all) variations in this sequence are systematically associated with different possible corresponding states of the linear sequences of the mRNA molecules and of the proteins synthesized. [...] To the extent that such dependency is present, varying the DNA sequence provides for a kind of fine-grained and specific *control* over which RNA molecules or proteins are synthesized. (Woodward 2010, 306; his italics)

It is rather more controversial, however, to suggest that an organism's behavioural phenotype depends on its DNA sequence in the same way. Behaviours are, in many cases, hugely sensitive to aspects of an organism's environment, including the signals and cues it

[12] See Briggs (2012) for illuminating discussion of the semantic and metaphysical differences between simple and interventionist counterfactuals.

[13] There may well other useful ways of cashing out the idea that DNA controls the synthesis of RNA and proteins (see Stegmann forthcoming). I focus on control-as-systematic-counterfactual-dependence because I take it to be the most relevant sense of control for my purposes.

receives from other organisms. The worry naturally arises that, if control requires systematic counterfactual dependence, then genetic control over behavioural phenotypes is at best an idealization, at worst a dangerously misleading myth.

Three separate considerations help to deflate this worry. The first is that attributions of genetic control in behavioural ecology are normally directed at strategies rather than token behaviours. As we saw above (Section 2.1.5), a strategy is best envisioned as a set of behavioural dispositions that specify how an organism would behave across the range of ecological and social contexts in which it might find itself. Given this conception of a strategy, the suggestion that genotypes often possess a significant degree of control over an organism's strategy is compatible with an acknowledgement that they usually possess rather less control over an organism's manifest behaviour. This is a special case of the more general observation that, while an organism's realized phenotype may vary greatly across environments—and hence cannot be regarded as being specified, determined or controlled to any high degree by the genotype—the norm of reaction relating environment to phenotype may still be under largely genetic control (see Schlichting and Pigliucci 1998; Pigliucci 2001; West-Eberhard 2003).

The second consideration is that systematic counterfactual dependence of strategy on genotype is a matter of degree, not an all or nothing affair. I will not attempt to formulate a precise metric of control here; this remains a challenge for future work in this area. Note, however, that Lewis's account of systematic counterfactual dependence suggests an intuitive measure in terms of the grain at which the propositions in the A and C sets partition the space of possible alternatives. If we can only achieve counterfactual dependence by partitioning the possible alternatives at an extremely coarse grain (e.g., if I were to set the thermostat to 'hot', the temperature would be between 20 and 30°C; if I were to set it to 'cold', the temperature would be between 10 and 20°C), then the A-possibilities have only a low degree of control over the C-possibilities. If, by contrast, we can still obtain systematic counterfactual dependence even when we partition the

possibilities much more finely (if I were to set the thermostat to '20˚C', the temperature would be 20˚C to the nearest degree; if I were to set it to '21˚C', the temperature would be 21˚C to the nearest degree, and so on), then the *A*-possibilities control the *C*-possibilities to a much higher degree. It is very likely that an organism's behavioural dispositions will depend on its genotype to some degree, though the precise degree may vary greatly from case to case.

The third is that control is not exclusive: assigning a degree of control to an organism's genotype is compatible with other causes having an equal or perhaps greater degree of control. Hence, while control is often cited as a special feature of genetic causes—a feature which objectively distinguishes their role in development from those of other causes, such as environmental causes[14]—talk of genetic control is not essentially committed to this view. We can, in principle, assign degrees of control to aspects of an organism's environment.[15]

These considerations go some way towards validating talk of genetic control, conceived in terms of systematic counterfactual dependence, in the context of social behaviour. Bearing all these points in mind, I propose the following account of notion of control, as it applies in behavioural ecology:

> **Genetic control:** The degree to which an organism, *O*, genetically controls a behavioural strategy, S_0, is the degree to which the strategy set, **S**, of which S_0 is a member, exhibits systematic counterfactual dependence on the set of alternative genotypes, **G**, of which *O*'s genotype, G_0, is a member.

[14] See Sarkar 2004; Weber 2006; Waters 2007; Woodward 2010.

[15] For instance, we might allow that the ambient temperature controls, to some extent, the sex of reptiles with mechanisms for temperature-dependent sex-determination; though since the dependence is coarse-grained, the degree of control is low.

Though I have not provided a quantitative measure, we can see that the degree of control will depend on the grain at which we can partition **S** and **G** while retaining counterfactual dependence of alternative strategies on alternative genotypes. If we can only achieve systematic counterfactual dependence by characterizing both strategies and genotypes in a highly coarse-grained, disjunctive fashion (if O were to have G_0 or G_1 or G_2 or ..., it would do S_0 or S_1 or S_2 or ...), we have a low degree of genetic control; if we can move to finer and finer grains of analysis without losing counterfactual dependence, we have higher and higher degrees of genetic control, arriving eventually at the ideal limit of a one-to-one mapping of the members of **G** on to the members of **S** (if O were to have G_0, it would do S_0; if O were to have G_2, it would do S_2, ...).

On the account I have outlined, attributions of degrees of control are relative to a set of possible alternative strategies, **S**, and a set of possible alternative genotypes, **G**. The implication is that reasonable attributions of degrees of control are dependent on a reasonable choice of **S** and **G**: oddly gerrymandered sets of strategies and genotypes may yield strange control attributions.[16] How, then, should membership of **S** and **G** be determined? One possibility is to restrict membership of **S** and **G** to strategies and genotypes actually present in the population (cf. Waters 2007). But while this might produce useful control attributions for some theoretical purposes (we could talk of 'actual control', where an organism actually controls a strategy if and only if it controls it relative to the actual sets of strategies and genotypes in the population), it seems too stringent a restriction to apply across the board. The reason is that actual control requires actual variation with respect to the strategy in question, and we may want to attribute degrees of control with respect to some behavioural strategy even when the strategy does not actually vary in the population under study. I therefore leave open the question of how **S**

[16] This problem is not avoided by adopting Woodward's notion of 'causal specificity' in place of systematic counterfactual dependence. For, on Woodward's account, assessments of causal specificity are relative to the choice of the cause and effect variables. In this context, the variables would be **S** and **G**, and the problem of how to determine membership of these sets would remain.

and **G** should be determined. Like prior probabilities in Bayesian epistemology, they will typically have their origins in somewhat murky intuitive judgements about the plausibility of different alternatives. All we can say is that, just as reasonable priors are needed for reasonable posterior probabilities, reasonable choices of **S** and **G** are required for reasonable attributions of genetic control.

2.3.3 Related notions

How does the notion of control, conceived in terms of systematic counterfactual dependence between a strategy set and a genotype set, relate to the associated concepts we encountered in the cases discussed in Section 2.2.3? I will first consider the relationship between control and manipulation. I will then turn to the relationships between control, enforcement, acquiescence and volunteering. In both cases, I will stress the need to separate questions of control from other, conceptually distinct questions with which they are easily conflated.

Signal-induced behaviours: coordination versus manipulation

The common cuckoo (*Cuculus canorus*) is an iconic manipulator. A notorious brood parasite, the mother discreetly drops her eggs into the another bird's nest (there are many host species, perhaps the best-known being the Eurasian reed warbler, *Acrocephalus scirpaceus*); the host then feeds and raises the cuckoo chick as if it were her own, deceived by the calls of the cuckoo chick, which mimic with uncanny precision the calls of the conspecific chick she might otherwise have raised (see Payne 2005 for further details). In formulating an account of manipulation, it will be helpful to keep this example in mind: we want an account that correctly identifies the cuckoo's behaviour as manipulative, without also counting as manipulative the barely distinguishable calls of a *bona fide* reed warbler chick.

One might intuitively imagine that manipulation is a form of control, and that the notion as it applies in behavioural ecology should be defined in terms of control. But if we are to think of control as I have urged—that is, in terms of systematic counterfactual dependence—this approach will not do. The reason, in a nutshell, is that considerations of counterfactual dependence do not correctly identify the difference between the behaviour of the cuckoo chick and that of the reed warbler chick. The reed warbler's feeding behaviour depends counterfactually on call of the cuckoo chick to some degree; the chick can therefore be said to have some degree of control over the mother's feeding behaviour. Crucially, however, the call of the *bona fide* reed warbler chick has exactly the same effect on its mother's behaviour; so the reed warbler chick can be said to possess exactly the same degree of control.

If the difference between these cases is not one of systematic counterfactual dependence, then what is it? Plausibly, it has something to do with the information the chicks convey to the reed warbler mother when they call. Yet both signals convey the same information. Informally, they both say 'I am your offspring and I need food'. More formally, both signals raise the probability, conditional on the evidence available to the mother,[17] that the mother is interacting with one of its own chicks, and that the chick is hungry. In this sense, they both carry the same informational content (see Skyrms 2010 for a detailed treatment of informational content in terms of conditional probability-raising). Crucially, however, there is a difference in the accuracy of the two signals: the call of the reed warbler chick is accurate, in that it raises the conditional probability, from the mother's point of view, of a state that actually obtains; whereas the call of the cuckoo chick, though aurally

[17] There is no intended connotation here that the mother is capable of consciously making probability judgements. Skyrms' (2011) account of informational content is intended to be far more general than this: for instance, it is intended to apply to signalling among bacteria. The probabilities are taken to be objective conditional probabilities; they can be computed in principle by human observers, but need not be computed by agents in the population under study.

indistinguishable, is deceptive, in that it raises the conditional probability, from the mother's point of view, of a state that does not obtain.

With this in mind, I propose that we restrict the term 'manipulation' to describe only those cases in which behaviours are induced by deceptive signalling:

> **Manipulation:** One individual manipulates another with respect to a particular behaviour iff the first induces the second to perform that behaviour by means of a deceptive signal.

Thus construed, the appropriate contrast with manipulation is coordination, in which a behaviour is induced by an accurate or 'honest' signal:[18]

> **Coordination:** One individual coordinates with another with respect to a particular behaviour iff the first performs the behaviour only in response to an honest signal from the second.

This account of manipulation (and coordination) makes sense of Keller and Nonacs contrast between manipulation and 'honest signalling', a contrast that may seem puzzling at first glance. But it also suggests that Keller and Nonacs are wrong to equate manipulation with 'queen control' of worker strategies. For, on my account, the question of whether a pheromone-induced behaviour represents a case of coordination or manipulation—that is, the question of whether it is the product of honest or deceptive signalling—is conceptually distinct from that of whether pheromonal signals afford the queen any degree of control over the behavioural strategies of the workers. The former question concerns the content of the signal—is it honest or deceptive? —while the latter

[18] Compare Section 2.2.3: to say that a task requires coordination between its participants is to say that it cannot be completed without the participants signalling honestly to each other (typically about the subtask they are performing and its state of completion) and responding appropriately to one another's signals.

concerns the extent to which a worker's behavioural strategy exhibits systematic counterfactual dependence on the content of the signals it receives. It is a conceptual possibility that a queen manipulates her workers into performing particular behaviours by means of deceptive signals without thereby achieving any significant degree of control over their strategies; it also a conceptual possibility that queen achieves control over worker strategies through honest signalling. Hence, even if we settle definitively the question of whether pheromonal signals are honest or deceptive, the question of whether they facilitate queen control of worker strategies remains open.

Note that the open question here is not whether pheromones afford the queen some degree of control over the manifest behaviour of the workers, but whether they afford the queen a degree of control over the workers' strategies, construed as sets of behavioural dispositions. A pheromone-producing queen is very likely to have some degree of control over which behaviours her workers actually manifest, since worker strategies often issue in different behaviours in response to different pheromonal stimuli; indeed, if the dependence of manifest behaviour on pheromonal stimulus is sufficiently fine-grained, the queen may even have a high degree of control over the manifest behaviour of the workers. But this is quite compatible with the claim that the worker alone controls the strategy—that is, the set of behavioural dispositions that determine its responses to stimuli. It may well be the case in many species of social insect that the queen is able to influence the manifest behaviour of workers, yet has no control over how the workers will respond to the signals she emits. Compare: I can easily control whether or not the glass on the table in front of me is actually smashed, but I have no control over whether or not it is fragile.

Of course, it may turn out that queens in some species do enjoy a degree of control over the strategies of workers, as well as over their manifest behaviour. The crucial point is that we should take care not to conflate these hypotheses: positing a degree of queen control over the workers' manifest behaviour is not equivalent to positing a significant degree of queen control over the workers' behavioural dispositions. The former can be achieved

through any form of signalling; the latter is much less easily achieved. And we should take even greater care not to conflate either of these hypotheses with the claim that pheromonal signals are manipulative, in the sense of inducing behaviour through deceptive informational content.

Enforcement, acquiescence and volunteering

Like manipulation, the notions of enforcement, acquiescence and volunteering sound, at least on first hearing, as though they ought to be defined in terms of control. Again, however, I suspect that this intuition, though perhaps reasonable in the human context, is not correct in the context of behavioural ecology. In contemporary theory, enforcement is most commonly used to refer to mechanisms that help maintain cooperation in a social group by differentially imposing fitness costs on defectors, and/or fitness benefits on cooperators, to counterbalance the costs of cooperating relative to defecting. I propose that, to avoid conflating it with other notions, we should restrict the use of the term 'enforcement' to only those cases in which such a mechanism is present:

> **Enforcement:** A strategy is enforced iff fitness penalties are differentially imposed on individuals that fail to adopt it and/or fitness rewards are differentially conferred on individuals that do adopt it.

An enforcement mechanism is active if penalties and rewards are actually imposed on members of the social group. Enforcement mechanisms need not be active, however. If they are effective enough, the population may evolve such that virtually all individuals cooperate. If, at this point, the mechanism were simply to break down and disappear, defection might once again start to gain a foothold. An alternative is for the mechanism to be retained but in a latent state: inactive, but ready to spring into action should a significant number of defectors arise (see Wenseleers et al. 2004; Ratnieks and Wenseleers

2008). I propose that we reserve the term 'acquiescence' for cases in which a strategy is maintained by virtue of a latent enforcement mechanism:

> **Acquiescence:** An individual acquiesces to performing a particular strategy or behaviour iff (i) the strategy or behaviour is not actively enforced, but (ii) there exists a latent enforcement mechanism without which the strategy would not be evolutionarily stable.

Both active enforcement and acquiescence can then be contrasted with volunteering. Volunteering, like any other term in behavioural ecology, should not be construed as imputing psychological motives to evolutionary agents, but should rather be construed in a technical sense in terms of the absence of an enforcement mechanism:

> **Volunteering:** An individual volunteers to adopt a particular strategy iff it is not exposed to any enforcement mechanism, whether active or latent, with respect to this strategy.

It is important to recognize that none of these enforcement-related notions has any close conceptual connection to the notion of control, conceived in terms of systematic counterfactual dependence. An individual may retain a high degree of control over its behavioural strategy and yet only adopt that strategy as a response to active or latent enforcement. Similarly, an individual's strategy may be voluntary, in the sense of being unenforced, while still being heavily influenced by another individual. Suppose, for example, that a particular ant is caused to develop particular morphological characteristics by differential feeding in the larval stage, but is not subject to any kind of enforcement mechanism, whether active or passive, as an adult. The ant's behavioural strategy as an adult will not be under its own full control: it will have been specified to some degree by

the feeding regime imposed upon it in the larval stage. But this does not mean the strategy it adopts is enforced, in the sense of incurring differential penalties or rewards.

By keeping enforcement and related notions separate from the notion of control, we can clearly distinguish a case in which an individual's strategy is self-controlled but subject to penalties/rewards, from a case in which an individual's strategy is not fully under its own control but is not subject to penalties/rewards. This distinction is important, for the two cases exemplify two different mechanisms for the maintenance of cooperative phenotypes. If we use enforcement and control loosely and interchangeably, we will end up glossing over this distinction.

The question of whether altruism in insect societies is voluntary or enforced is also quite different from the question of whether it is a product of coordination, manipulation or neither. The latter concerns the informational content of the signal by which a behaviour is induced, while the former depends on whether there is a regime in place that differentially penalizes those individuals who make the selfish choice, or differentially rewards those who cooperate.

We therefore have at least three different questions we can ask about the mechanisms by which social behaviours and strategies are induced and maintained. We can ask who controls the behaviour or the underlying strategy, and to what degree. We can ask whether a particular behaviour is a product of coordination (honest signalling) or manipulation (deceptive signalling). And we can ask whether the strategy or behaviour is enforced (and, if so, whether the enforcement is active or latent) or if it is voluntary. The distinctions between these questions should not be blurred—and we should not assume that, by finding the answer to one of them, we will find the answer to either of the other two.

THREE

Selection, Transmission and the Price Formalism

The previous chapter addressed a cluster of foundational concepts in behavioural ecology. In this chapter, I change tack, and consider a number of related philosophical issues raised by formal representations of evolution. These projects may initially appear unrelated, but of course they are not: in later chapters, we will see how formal representations of evolution are able to shed light on the origins of social phenomena.

I focus in particular on the 'covariance selection mathematics' of George R. Price (1970, 1972), which has in recent decades become the preferred framework of many evolutionary biologists for the formulation of fundamental theory, and which provides the basic theoretical apparatus for subsequent chapters. There are many good existing introductions to the Price formalism (see, e.g., Grafen 1985a; Frank 1995, 1997a, 1998; Rice 2004; Okasha 2006; McElreath and Boyd 2007; Gardner et al. 2007; Gardner 2008; Wenseleers et al. 2010; Gardner et al. 2011). But although I stay close these standard treatments in some respects, my discussion differs in emphasizing interpretative questions relevant to the study of social evolution that the mathematics alone does not settle.

In Section 3.1, I outline the derivation of the central principle of Price's formalism—the standard version of the Price equation—and explain the meaning of the variables it describes. In Section 3.2, I contrast phenotypic formulations of the Price equation with genetic formulations. I show that there is a substantive difference between them with regard to how they account for the effects of heritability on evolutionary change, and I

argue for the superiority of genetic formulations in many social-evolutionary contexts. In Section 3.3, I discuss the causal interpretation of the Price equation as a separation of the effects of selection and transmission. I argue that we need to distinguish primary, secondary and tertiary effects of natural selection, and I show how we can use a three-term variant of the Price equation to separate all three effects. In Section 3.4, I examine three different ways in which organisms may usefully be sorted into equivalence classes within the Price formalism (namely: trait-groups, genotypic classes and developmental classes). I then bring these considerations to bear on two philosophical debates: one concerning the relationship between the Price formalism and 'evolutionary nominalism' (Godfrey-Smith 2009a; Nanay 2010), and the other concerning the relationship between kin-selectionist and multi-level approaches to the analysis of social evolution.

3.1 Introducing the Price equation

3.1.1 Ingredients

The Price equation is a highly general, highly abstract description of the change in aggregate properties between two sets. It is a piece of mathematics: its biological interpretation and application to organic evolution are entirely optional. To derive the equation, all we need is two sets of countable entities. In biology, the entities will often be organisms, but the derivation of the Price equation does not assume this. We label one population the ancestor-population (A) and the other population the descendant-population (D). In biology, the sets will usually be earlier and later time-slices of the same evolving population, so the labels are usually apt. But again, the derivation does not assume this.

The sets A and D must satisfy two conditions. First, the members of the two sets must be related by some salient mapping relation. In the abstract, we can represent this mapping relation as R. We need to be able to say, for each member of A, to which descendants it is

connected by *R*; and, for each member of *D*, to which ancestors it is connected by *R*. In biology, *R* will often be the relation of direct lineal descent[1]; that is, *R* will connect each member of *A* to all and only those members of *D* of which it is a direct, genealogical ancestor. Again, however, the derivation of the equation does not assume any particular biological interpretation of the *R*-connections.

Second, we must be able to attribute to each member of *A* and *D* a property, *z*; and we need to be able to attribute to each member of *A* two additional properties, *w* and z'. Let us consider each of these properties in turn.

z: The first property, *z*, is the property we are interested in studying—perhaps because its mean changes between the two sets, or perhaps because its mean stays the same. In biology, this will usually be a phenotypic or genotypic property or some kind (see Section 2.2). The only constraint on the nature of *z* is that we must be able to assign to each member of *A* and *D* a number representing its value for that property (if the property is qualitative, we can represent its presence as $z = 1$ and its absence as $z = 0$).

w: The second property, *w*, represents, for any particular ancestor, the number of entities in the *D* to which it is connected by *R*. In biology, this will usually be the number of entities to which it is connected by direct lineal descent. This quantity is often glossed as fitness (or realized fitness, when it is important to distinguish realized from expected fitness; I do not discuss this distinction here). I adopt this

[1] Why 'direct'? When generations overlap, some organisms within the ancestor-set may be the offspring of other organisms in the ancestor-set. For instance, the ancestor-set may contain the parents *and* grandparents of a particular descendant. In order to avoid double-counting, we usually ignore these relationships when assigning descendants to ancestors. So if the ancestor-set were to contain the parents and grandparents of a particular descendant, this descendant would by connected by *R* to its parents, but not to its grandparents (inclusive fitness analysis complicates this picture somewhat; cf. Chapter 5).

terminology here, but an important disclaimer is needed: an entity's fitness, thus construed, may come apart significantly from the intuitive notion of (realized) fitness as a measure of an organism's total number of offspring. First, *w* in the Price formalism can, in principle, be ascribed to any countable entity, whenever we have two sets of these entities connected by an appropriate mapping relation. It might, in principle, be ascribed to molecules, genes, cells, groups, species, ecosystems, cultural variants, and more; and there is no formal requirement that the entities to which it is ascribed are capable of 'reproduction' in any intuitive sense. Second, even when the *w*-bearers *are* organisms, and even when the mapping relation R is direct lineal descent, an organism's number of offspring is rarely the best measure of its value for *w*. Fitness, in the sense of the Price formalism, will reliably align with number of offspring only when generations do not overlap, and when the A and D populations are separated by a single generation. When generations do overlap (so that organisms of different ages coexist in the same population), or when the ancestor- and descendant-sets are separated by multiple generations, the two notions will often come apart: an organism's total reproductive output may not be a good indicator of the number of direct descendants it contributes to the descendant-set.[2] For this reason, some authors prefer to gloss w as reproductive value, and, following R.A. Fisher (1930), label it with the letter 'v' (Grafen 2006b; Wenseleers et al. 2010; Gardner et al. 2011).

z': The third property, z', represents, for any particular member of A, the average value of z in the members of the descendant-set to which it is related by R. To

[2] For example, in an age-structured population, younger organisms are likely to contribute more direct descendants to the descendant-set than older organisms with the same lifetime reproductive output. In a class-structured population (e.g., an eusocial insect society), organisms which produce offspring of more productive classes are likely to contribute more direct descendants to the descendant-set than organisms which have the same total output but which produce offspring of less productive classes (assuming the ancestor- and descendant-sets are separated by more than one generation).

calculate z' for the i^{th} individual (z_i'), we look at the value of z in its descendants, and take the average of these values. Importantly, however, although we may calculate z' by looking at D, it is still a property of a member of A: it is a *relational* property of an ancestor, a piece of information about the way it has transmitted its z-value to its descendants.

3.1.2 The derivation

Given the definitions of z, z_i' and w, the derivation of the Price equation is fairly straightforward. We begin by writing the change in the mean value of z between A and D as its mean value in D, minus its mean value in A:

$$\Delta\bar{z} = \bar{z}_D - \bar{z}_A$$

We then express each of these averages as a sum over properties of members of A. To calculate the average z-value in A, we simply sum over the z-values of each of the n members of that set, and divide by n:

$$\bar{z}_A = \frac{1}{n}\sum_i^n z_i$$

Crucially, however, we do *not* calculate the average z-value in D by summing over the z-values of the members of D. Instead, we sum over the z_i'-values of individuals in A, weighting each ancestor by its relative value for w (i.e., by the relative number of descendants to which it is connected by R):

$$\bar{z}_D = \frac{1}{n}\sum_i^n \frac{w}{\bar{w}} z_i'$$

At first sight, writing the average z-value in D as a fitness-weighted sum of z'-values in A may seem eccentric. Why not simply sum over the z-values of the individuals in D? This

move, however, is absolutely critical to the derivation. By expressing the average z-value in D as a sum over properties of A, we lay the foundations for a result that describes how the properties of the two sets relate to one another. It is also the only point at which the derivation makes a substantive assumption about the populations it describes. To be specific, it is assumed that *all descendants have the same number of ancestors*, since it is only on this assumption that $\bar{z}_D$ is equal to a fitness-weighted average of the ancestors' z'-values. Since there are possible pairs of ancestor- and descendant-populations in which the R-connections violate this assumption, there are possible pairs of populations for which the standard Price equation does not hold (see Kerr and Godfrey-Smith 2009 for an extension of the Price equation to accommodate these cases). Nevertheless, the assumption is widely applicable to evolving populations in nature.

Combining our expressions for $\bar{z}_A$ and $\bar{z}_D$, we obtain the following:

$$\Delta\bar{z} = \frac{1}{n}\sum_{i}^{n}\frac{w_i}{\bar{w}}z_i' - \frac{1}{n}\sum_{i}^{n}z_i \tag{3.1.1}$$

The rest of the derivation is a simple exercise in rearrangement and re-labelling. First, we rewrite (3.1.1) as follows:

$$\Delta\bar{z} = \frac{1}{n}\sum_{i}^{n}\frac{w_i}{\bar{w}}\left(z_i' - z_i\right) + \frac{1}{n}\sum_{i}^{n}\frac{w_i}{\bar{w}}z_i - \frac{1}{n}\sum_{i}^{n}z_i$$

By merging the second and third summations, and by re-labelling $z_i' - z_i$ as Δz_i, we obtain the following:

$$\Delta\bar{z} = \frac{1}{n}\sum_{i}^{n}\frac{w_i}{\bar{w}}\left(\Delta z_i\right) + \frac{1}{n}\sum_{i}^{n}z_i\left(\frac{w_i}{\bar{w}} - 1\right) \tag{3.1.2}$$

This is the Price equation in all but name. By applying the standard definitions of 'expectation' and 'covariance'[3], we can re-label the terms as follows:

$$\Delta\bar{z} = \frac{\mathrm{E}(w\Delta z)}{\bar{w}} + \frac{\mathrm{Cov}(w,z)}{\bar{w}}$$

By reversing the order of the terms, we obtain the equation in its original and most commonly used form:

$$\Delta\bar{z} = \frac{1}{\bar{w}}\left[\mathrm{Cov}(w,z) + \mathrm{E}(w\Delta z)\right] \tag{3.1.3}$$

Although relabelling the terms in equation (3.1.2) as 'covariance' and 'expectation' allows for a compact and convenient expression of the Price equation, it can also lead to confusion. This is because, in statistics, 'covariance' and 'expectation' are normally understood either as properties of the probability distributions of random variables, or as properties of a sample drawn at random from a larger population. Yet the derivation of the Price equation makes no assumption that z is a random variable, or that A and D are samples from a larger population (cf. van Veelen et al. 2012). As a result, the use of statistical notation arguably gives the impression that the equation is less general than it actually is.

I will stay reasonably close to Price's statistical notation here, though I note that various alternatives are available: in addition to the statistical form given in equation (3.1.3) and the 'naked', algebraic form given in equation (3.1.2), it is also possible to rewrite the Price

---

[3] Expectation: For a discrete random variable, $\mathrm{E}(X) = \sum_i p_i x_i$, where x_i is the i^{th} possible state of X, and p_i is its probability of obtaining. In the present context, we are weighting values of $w\Delta z$ by *frequency* rather than probability; and, since we are counting each individual's value for $w\Delta z$ separately, $p_i = 1/n$ for all i.

Covariance: The covariance of two random variables is the expected product of their deviations from the mean, i.e., $\mathrm{Cov}(X,Y) = \mathrm{E}\left[\left(X - \mathrm{E}[X]\right)\left(Y - \mathrm{E}[Y]\right)\right]$ or, equivalently (and often more conveniently), $\mathrm{Cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)$.

equation in vector notation, or even in information-theoretic notation (Frank 2012). Since these variants are all equivalent statements of the same theorem, one's preference in this regard will no doubt depend on one's prior views as to how the evolutionary process ought to be represented. Price's statistical formulation is in keeping with Fisher's (1930) conviction that natural selection, like the behaviour of gases, is a phenomenon properly described in the language of statistics.

3.2 Genetic *versus* phenotypic formulations

We noted above that, in the Price equation, z can be used to represent *any* property for which every individual in the ancestor- and descendant-sets has a value. Although the letter z is conventionally used to denote phenotypes, there is no requirement that the Price equation be used to describe only phenotypic change; indeed, social evolution theorists more commonly deploy the equation to describe change in the *genetic* properties of individuals (see, e.g., Price 1970, 1972a; Grafen 1985a; Wade 1985). In this section I compare several genetic versions of the Price equation, and contrast these with a purely phenotypic formulation. The aim is to draw attention to the substantive differences in how the different formulations partition the overall evolutionary change.

3.2.1 The genetic Price equation(s)

Allelic values

We usually think of an 'allele frequency' as a property of a *population*: namely, it is the total number of copies of that allele in the population, divided by the total number of copies there *would* be, if every individual possessed the maximum number of copies it could possess (i.e., if every individual had the maximum 'allelic dosage'). As Price (1970) notes, however, allele frequencies can also be ascribed to *individuals*. Roughly speaking, we can define an individual's personal frequency for a particular allele as the number of

copies of that allele that individual possesses (i.e., its personal allelic dosage), divided by its personal ploidy. Following Frank (1998), I will refer to these personal allele frequencies as allelic values:

> **Allelic value**: An individual's allelic value (or individual allele frequency), x, with respect to a particular allele at a particular locus, is the number of copies of that allele it possesses, divided by its ploidy.

Note that, given the definition of x, the population mean $\bar{x}$ is equal to the overall frequency of the allele in the population. Exploiting this convenient relationship, Price (1970) originally derived his equation as an expression for the change in overall frequency of a particular allele:

$$\Delta\bar{x} = \frac{1}{\bar{w}}\left[\mathrm{Cov}\left(w,x\right) + \mathrm{E}\left(w\Delta x\right)\right] \tag{3.2.1}$$

In his original paper, Price writes equation (3.2.1) without the expectation term. The reason is that this term can be assumed to be zero in the absence of mutation, gametic selection or intragenomic conflict, and Price follows many population geneticists in neglecting these effects. If we do neglect these effects, we obtain the highly elegant result that the change in the overall frequency of an allele is equal to the covariance between an individual's relative fitness and its personal allele frequency. It seems probable that Price was already well aware that the same result would hold much more generally, indeed for any property for which every individual in a population can be assigned a numerical value.[4] Not unreasonably, however, he took x to be the most salient property for the purposes of studying genetical evolution.

---

[4] Price writes: 'This is a preliminary communication describing the application to genetical selection of a new mathematical treatment of selection in general' (1970, 520).

p-scores

Price's original formulation is extended, in subtly different ways, by Alan Grafen (1985a) and David Queller (1992a,b). Grafen reformulates the Price equation in terms of *p*-scores, where a *p*-score is a quantitative characterization of an individual's genotype, obtained by aggregating its allelic values (see Grafen 1985a, 2006):

> ***p*-score**: A *p*-score, p, is a weighted sum of an individual's allelic values. The weights can take any value, and each possible set of weights defines a different *p*-score.

The notion of a *p*-score is intended to be as broad as possible: we can assign non-zero weights to any linear combination of allelic values we happen to care about, and we can weight them in any way we see fit—we can even weight them completely arbitrarily if we want to (Grafen 1985a). The most appropriate *p*-score to use when analysing a particular problem will depend on the precise details of the problem. In the simplest case, in which we only care about a single allele at a single locus, the relevant *p*-score is simply an individual's allelic value for that allele. If we care about various alleles at one locus, or at multiple loci, the relevant *p*-score will be one that sums over a larger number of allelic values.

The Price equation holds for any possible *p*-score. There are two ways to see this. One is to begin with the fact that the equation holds for a single allelic value; and then to note that, since the operators Cov and E are linear[5], the same equation will hold for any linear combination of such values. The other is to begin with the fact that the Price equation holds for *any* property for which we can assign a value to all members of the ancestor- and

[5] Formally, $\mathrm{E}(aX+bY)=a\mathrm{E}(X)+b\mathrm{E}(Y)$ and $\mathrm{Cov}(aX+bY,Z)=a\mathrm{Cov}(X,Z)+b\mathrm{Cov}(Y,Z)$. Strictly speaking, Cov is *bilinear*: it is linear in both arguments.

descendant-sets; and then to note that every possible *p*-score is one such property, even though a *p*-score is unlikely to have any meaningful biological interpretation when the alleles are weighted arbitrarily. Either way, we obtain the following variant of the Price equation (Grafen 1985a, 2002, 2006a):

$$\Delta\bar{p} = \frac{1}{\bar{w}}\left[\mathrm{Cov}(w,p) + \mathrm{E}(w\Delta p)\right] \qquad (3.2.2)$$

Breeding values

The drawback with formulating the Price equation in terms of allelic values or *p*-scores is that we often want to explain the evolution of *phenotypic* characters. Yet studying the sign and magnitude of $\Delta\bar{x}$ or $\Delta\bar{p}$ can tell us nothing about the sign and magnitude of evolutionary change in any phenotype, unless we can establish that the average value of some character is more likely to increase if the frequency of a particular gene increases, or if the mean value of a particular *p*-score increases. There is a gap between genotypes and phenotypes that neither allelic values nor *p*-scores can bridge satisfactorily, for they concern purely genotypic properties of their bearers.

Queller (1992a,b) suggests an alternative: formulate the Price equation in terms of breeding values. The breeding value, a central notion in quantitative genetics, is an individual's value for a phenotypic character *as predicted by the average effects of relevant alleles* (see Falconer and Mackay 1996):

> **Breeding value**: An individual's breeding value (or additive genetic value), *g*, for a phenotypic character, *z*, is a sum of its allelic values at relevant loci, weighted by their average effects on *z* so as to give the best possible prediction of the phenotype.

In this context, the 'average effect' of an allele is interpreted in statistical terms (following Fisher 1930) as the partial regression of the phenotypic value on the allelic value, computed by the method of least-squares. Regression analysis and the method of least-squares, in addition to their role in defining the notion of a breeding value, are also implicated in many derivations of Hamilton's rule. We will therefore revisit these topics in Chapter 4, and I will postpone detailed discussion of regression until then. For now, it is enough to note that weighting alleles by least-squares regression coefficients yields the best prediction of the phenotype that it is possible to obtain from a linear combination of allelic values: least-squares theory guarantees that no other system of weightings could predict the phenotype with greater accuracy (i.e., with an average error of smaller magnitude).

The breeding value bridges the gap between genetic and phenotypic properties. It is 'genetic' in the sense that it is simply a linear combination of allelic values, and allelic values are unambiguously genetic. But it is an unusual genetic property, in that it aggregates over as many alleles as it takes to generate the best prediction of the relevant phenotype, and weights them with whatever combination of weights maximizes their predictive accuracy. It is thus an essentially *phenotype-relational* property: there can be no question of calculating an individual's breeding value *simpliciter*, without first specifying a phenotypic character with respect to which the breeding value is to be evaluated.

The version of the Price equation we obtain by taking the breeding value as the property of interest has a particularly useful feature. From now on, let us no longer use the letter z to represent any property at all, but instead use it to represent the phenotypic character of interest for which g is the corresponding breeding value. Usefully, the change in the

population mean for z will always equal the change in the population mean for g (Frank 1998, 2012):[6]

$$\Delta\bar{z} = \Delta\bar{g} = \frac{1}{\bar{w}}\left[\mathrm{Cov}(w, g) + \mathrm{E}(w\Delta g)\right] \tag{3.2.3}$$

In this sense, the breeding value provides the link between genotypic and phenotypic change that allelic values or *p*-scores alone do not provide.

Grafen (1985a, 2006a) suggests that we can regard the breeding value as a special type of *p*-score, but it seems to me that there is a subtle difference. Like a *p*-score, a breeding value is a weighted sum of allelic values. But unlike a *p*-score, the weights which attach to each allelic value are not *constants* assigned by the theorist, but *variables* that depend on the average effect of each allele on the phenotype. The average effect of an allele can change between the ancestor- and descendant-sets. If it does, the weights used to calculate breeding values will also change. The upshot is that, in switching from *p*-scores to breeding values, we introduce a new set of influences on the expectation term in the Price equation. Like $\mathrm{E}(w\Delta x)$ and $\mathrm{E}(w\Delta p)$, $\mathrm{E}(w\Delta g)$ is affected by mutation, gametic selection and intragenomic conflict. But it is *also* affected by changes in the average effects of alleles on z. Such changes might arise from changes in the environment; or from changes in allele frequency, if these in turn alter the frequencies of dominance or epistasis effects (see Section 2.3.2; see also Frank 1997a, 1998).

---

[6] One might intuitively imagine that $\Delta\bar{z} = \Delta\bar{g}$ would require substantive assumptions, such as fair meiosis or the absence of gene-environment correlation, but this is not the case: the equality is guaranteed by the definition of breeding value. Breeding value is a sum of allelic values weighted by their average effects, as estimated by the method of least-squares (Falconer and Mackay 1996). In other words, $z = g + \varepsilon_z$, where ε_z is the residual phenotypic value not predicted by the average effects of alleles. Consequently, $\bar{z} = \bar{g} + \bar{\varepsilon}_z$ and $\Delta\bar{z} = \Delta\bar{g} + \Delta\bar{\varepsilon}_z$. Least-squares theory guarantees than the mean of residuals on any regression line will be zero, implying that $\bar{\varepsilon}_z = 0$ in both the ancestor- and descendant-sets. It follows that $\Delta\bar{z} = \Delta\bar{g}$.

3.2.2 Two aspects of heritability

I now want to consider the differences between a genetic formulation of the Price equation in terms of breeding values (i.e., equation (3.2.3)) and a purely phenotypic formulation (i.e., equation (3.1.3), when *z* is interpreted as a phenotypic character). Both express the overall change in $\bar{z}$ as a sum of two components, but they carve up the change in subtly different ways. The difference, in a nutshell, concerns where they count the effects of heritability. As a preliminary, it is important to distinguish, following Frank (1997a, 1998), two aspects of heritability. In the most general sense, the heritability of a character *z* is the overall extent to which differences between the *z*-values of ancestors predict differences between the *z*-values of their descendants. The most inclusive measure of the heritability of a character is $\beta_{z'z}$, the regression of descendant phenotype on ancestor phenotype (Rice 2004; Okasha 2006, 2010).[7] In many cases, however, we can partition this overall measure into two components, each attributable to a separate causal process (Figure 2.1).

The first is the transmission, from ancestors to descendants, of genetic material relevant to the character.[8] All else being equal, greater fidelity in the transmission of genes will issue in greater resemblance between parents and offspring; this is the transmission aspect of heritability. The second is development, which connects transmitted genetic material to realized phenotypes. Since phenotypic differences between ancestors will only recur in their descendants if they are underpinned by transmissible genetic differences (though see footnote 8), the overall phenotypic resemblance between ancestors and descendants will

---

[7] As before, I postpone detailed discussion of regression until Chapter 4.

[8] Other forms of transmission may often be important in evolution, including epigenetic and cultural transmission (Jablonka and Lamb 2005; Helantera 2011). In talking of 'genes', I do not mean to downplay the significance of non-DNA-based transmission. Indeed, the measure of transmission fidelity as the regression of offspring breeding value on parental breeding value can potentially accommodate other forms of transmission. While breeding value is usually defined as the phenotypic value predicted by the best additive system of *allelic* predictors, there is no reason why other relevant transmissible properties could not be included in the predictor set used to determine breeding value.

depend in part on the extent to which phenotypic variation among ancestors is explained by transmissible genetic variation. This is the *variance aspect* of heritability.

The transmission aspect of heritability, with respect to a particular character, is most naturally measured by $\beta_{g'g}$, the regression of descendant breeding value on ancestor breeding value. The variance aspect of heritability can be quantified by either of two well-established measures. The first, known as the 'narrow-sense heritability', considers the ratio of variance in *breeding value* (i.e., the *additive* genetic variance) to the overall phenotypic variance. The second, known as the 'broad-sense heritability', considers the ratio of *all* relevant genetic variance to the overall phenotypic variance (in particular, variation in *products* of allelic values, which predict dominance and epistasis, are also considered). In evolutionary biology, the 'narrow-sense' measure is almost invariably preferred to the 'broad-sense' measure. This is not because evolutionary biologists naively assume that the proportion of the overall genetic variance accounted for by dominance or epistasis is invariably small or negligible. Rather, it is because the role of (the variance aspect of) heritability in evolutionary theory is to relate phenotypic variance to its transmissible basis; and, when reproduction is sexual and meiosis is fair, the additive genetic variance represents the transmissible portion of the overall genetic variance.

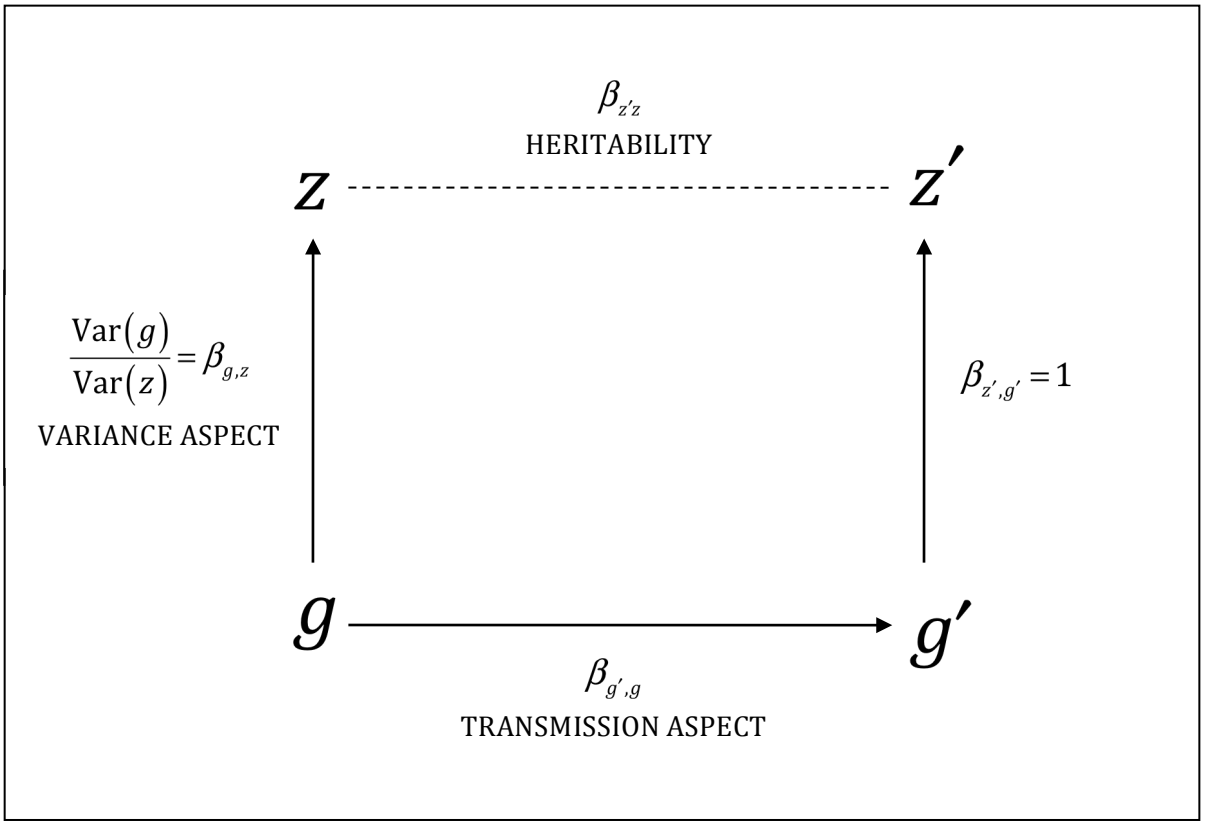


Figure 2.1: A diagram showing the two aspects of heritability and their measures. If all heritability is explained by genetic transmission, then the phenotypic correlation denoted by the dashed line is fully explained by the correlations denoted by the arrows, i.e., $\beta_{z',z} = \beta_{g',g}\beta_{g,z}$. Note that, by definition, the regression of a phenotypic character on its breeding value is 1; but the same does not apply to the reversed regression of breeding value on phenotype, which will normally be less than 1.

3.2.3 Comparing the genetic and phenotypic equations

We can now return to the phenotypic and genetic formulations of the Price equation, and consider: *which terms are affected by which aspects of heritability?* It should be clear enough that, in the phenotypic formulation, neither aspect of heritability affects the covariance term: $\mathrm{Cov}(w,z)$ is sensitive to neither the variance nor transmission aspects, since it depends only on the relationship between phenotypic differences and fitness differences *within* the ancestor-set. *All* the effects of heredity are therefore compressed into the expectation term, $\mathrm{E}(w\Delta z)$.

When we look at the *genetic* formulation, however, we see a different story. The covariance term, $\mathrm{Cov}(w,g)$, is still insensitive to the *transmission* aspect of heredity, since it concerns only the covariance between fitness and breeding value within the ancestor-set. Importantly, however, it does incorporate the effects of the *variance* aspect of heritability, since it is sensitive to the narrow-sense heritability of *z*: for a given value of $\mathrm{Cov}(w,z)$, lower narrow-sense heritability implies a lower value of $\mathrm{Cov}(w,g)$.[9] The phenotypic and genetic versions of the Price equation thus agree regarding the term to which they assign the transmission aspect of heredity, but they differ regarding to the term to which they assign the variance aspect.

The two formulations therefore differ substantively in the way they partition evolutionary change. Is there any reason to prefer one to the other? In some contexts, the phenotypic formulation will be more useful. This is notably true in cases in which a character is transmitted by wholly non-genetic means (e.g., a cultural variant that is uncorrelated with any allele). In such cases, the breeding value for that character will not even be well defined: the breeding value is the phenotypic value as predicted by *relevant (i.e., correlated) alleles*, and this notion has no meaning if there are no such alleles. We can use the Price equation to describe the evolution of such characters (for the cultural case, see Henrich and Boyd 2001; Henrich 2004), but in doing so we must employ the phenotypic formulation, not the genetic formulation.[10]

For the study of social evolution in non-human societies, however, the genetic formulation can be particularly useful. This is because it allows us to take account of the evolutionarily salient correlations that lurk below the surface when genes are differentially expressed. It is a truism that the phenotype of any given individual depends not merely on its genes,

---

[9] In a nutshell, this is because $\mathrm{Cov}(w,g) = \beta_{gz}\,\mathrm{Cov}(w,z)$ in many cases (see Queller 1992a).

[10] We may still be able to define a useful analogue of the breeding value, where allelic predictors are replaced by whatever transmissible predictors are relevant to the phenotype (cf. footnote 8). I do not explore this possibility here.

but also on how those genes are expressed during its life cycle. The result is that two individuals can have the same genetic value but very different phenotypic values for a given character; moreover, the phenotypic value of any given individual can vary hugely over its lifetime. It is sometimes suggested that this truism about development and physiology vitiates a gene-centred approach to evolution, but the opposite is closer to the truth: differential gene expression leads to two connected problems for any purely phenotypic approach to the analysis of social evolution. One is that, in populations where differential gene expression is rife, the overall fidelity of phenotypic transmission will often be very poor. As a result, the expectation term in the phenotypic Price equation will often be large, and may well be more significant than the covariance term (i.e., the term we usually want to analyse) with regard to the overall direction of evolution. The other problem is that phenotypic differences between social partners can belie their underlying genetic similarity, and this genetic similarity can have important evolutionary consequences regardless of whether it is manifested phenotypically.

Abstract as they may seem, both these problems are vividly illustrated by considering a eusocial insect colony. In the most complex eusocial societies, workers are highly differentiated in both morphology and behaviour, and there is a strict division of reproductive labour: all reproduction is undertaken by the queen (see Chapter 5 for further details). Usually, theoretical inquiry into the evolution of eusociality pays close attention to the genetic relatedness between the workers and the queen. This genetic relatedness can help explain how conditionally expressed genes for altruistic behaviour can positively co-vary with fitness, since these genes are present in the queen (i.e., the *beneficiary* of the workers' altruism), as well in the workers who incur the cost. Moreover, the reliable transmission of these genes from the queen to her offspring helps explain how altruistic behaviour reappears in each newly founded colony.

What would happen if we ignored genotypes, and looked only at *phenotypic* correlations between workers, and between the queen and her offspring? The answer is that the

evolutionary stability of eusociality would be much more difficult to explain. The altruistic *phenotypes* of the workers do *not* co-vary positively with fitness: they are genuinely altruistic in that they detract from the lifetime fitness of their bearers, and as such (in contrast to the genes underlying them) co-vary *negatively* with fitness. Without considering genetics, we would expect these behaviours to disappear rapidly from the population. What we would see instead, however, is that they are retained in the population by what would look like a bizarre bias in the transmission of phenotypes: the queen, rather than producing offspring that resemble her phenotypically, continually produces offspring with morphological and behavioural phenotypes that differ dramatically from her own, and that systematically tend to be a great deal more altruistic. This 'bias in phenotypic transmission' would appear to fortuitously counterbalance the selection against altruistic phenotypes in each generation. Until we see the underlying genetic similarity behind the phenotypic heterogeneity within and across generations, the stability of altruism defies any deeper explanation.

3.3 Selection, transmission, and 'spill-over'

3.3.1 Interpreting the Price formalism

The standard Price equation is commonly thought to separate the overall evolutionary change into a component attributable to natural selection and a component attributable to biased transmission (see, e.g., Frank 1995, 1997a, 1998; Gardner et al. 2007; Gardner 2008; Gardner and Foster 2008; Wenseleers et al. 2010; Gardner et al. 2011). The covariance term is taken to quantify the former, while the expectation term is taken to quantify the latter. As Samir Okasha (2006) notes, however, it is doubtful whether this standard interpretation is correct in general. The problem is that, in the standard Price equation, *both* terms functionally depend on differential fitness. For recall that the second term—the term supposedly attributable to 'transmission bias'—is an expectation of $w\Delta g$. Since each individual's value for Δg is weighted by its fitness, the personal transmission biases of

fitter individuals will make a bigger difference to the value of this term than those of less fit individuals.

A rearrangement of the Price equation, first derived by Frank (1997a, 1998), yields a 'modified Price equation' with an expectation term that is independent of fitness differences:

$$\Delta\bar{g} = \frac{1}{\bar{w}}\left[\mathrm{Cov}\left(w, g'\right) + \bar{w}\mathrm{E}\left(\Delta g\right)\right] \tag{3.3.1}$$

Frank's equation differs from the standard equation in two important respects. First, the covariance term replaces g, an individual's personal breeding value, with g', the average breeding value of its descendants. Second, the expectation term is no longer weighted by fitness: we look at the difference between an individual's breeding value and the average breeding value of its descendants *without* taking into account its relative contribution to the descendant-set.

Okasha (2006) suggests that the modified Price equation succeeds where the standard Price equation fails: that is, it *does* provide a clean separation of the effects of selection and biased transmission. In response to Okasha, Peter Godfrey-Smith (2007a) and Ken Waters (2011) have separately argued that matters are not quite so straightforward. For, although replacing $\mathrm{E}\left(w\Delta g\right)$ has the effect of making the second term independent of fitness differences, replacing g with g' has the effect of making $\mathrm{Cov}\left(w, g'\right)$sensitive to variation in transmission biases. Suppose, for example, that $\mathrm{Cov}\left(w, g\right)$ is positive, but that fitter individuals tend to transmit their breeding value less reliably than less fit individuals. In this scenario, $\mathrm{Cov}\left(w, g'\right)$ would be less than $\mathrm{Cov}\left(w, g\right)$.

Here, then, is the overall picture: the standard Price equation collates all the effects of biased transmission in the expectation term, yielding a covariance term that is independent of transmission bias. But it does not *cleanly* separate the effects of selection

and transmission, because the expectation term is sensitive to fitness differences as well as transmission biases. The modified Price equation, by contrast, collates all the effects of differential fitness in the covariance term, yielding an expectation term that is independent of fitness differences. But it too cannot be said to *cleanly* separate the effects of selection and transmission, because the covariance term is sensitive to transmission biases as well as fitness differences. Hence, neither version separates the effects of selection and transmission without some degree of 'spill-over'.

Why does this apparently inescapable spill-over arise? Why is it so difficult to partition the overall change cleanly into two terms, one attributable to selection alone, and the other to transmission alone? As Godfrey-Smith (2007a) and Okasha (2011) note, there is in fact a simple explanation. An individual's personal transmission bias is a character, and, like any other character, it is possible for it to co-vary with fitness. When such covariance occurs, there will be a component of the change in g that depends on *both* differential fitness and biased transmission, and so cannot be attributed to either process acting alone. This component is equal to $\mathrm{Cov}(w,\Delta g)$, and it accounts for the 'spill-over' in both the standard and modified Price equations. This is easier to see if we note the following notational identities:

$$\mathrm{Cov}(w,\,\Delta g)=\mathrm{Cov}(w,\,g')-\mathrm{Cov}(w,\,g)$$

$$\mathrm{Cov}(w,\,\Delta g)=\mathrm{E}(w\Delta g)-\bar{w}\,\mathrm{E}(\Delta g)$$

The standard Price equation accounts for $\mathrm{Cov}(w,\Delta g)$ as part of the expectation term, and, as a result, this term is sensitive to variation in fitness as well as to individual transmission biases. The modified Price equation, by contrast, accounts for $\mathrm{Cov}(w,\Delta g)$ as part of the covariance term, and, as a result, this term is sensitive to variation in transmission biases as well as to variation in fitness.

The only way to avoid spill-over of this kind is to represent $\mathrm{Cov}(w,\Delta g)$ explicitly as a separate term in the Price equation, rather than incorporating it into one of the other terms. The result is a third version of the equation which partitions the overall change into *three* components rather than two (Godfrey-Smith 2007a; Okasha 2011):

$$\Delta\bar{g} = \frac{1}{\bar{w}}\left[\mathrm{Cov}(w,g) + \mathrm{Cov}(w,\Delta g) + \bar{w}\,\mathrm{E}(\Delta g)\right] \tag{3.3.2}$$

The first term, $\mathrm{Cov}(w,g)$ is identical to the covariance term in the standard Price equation and is independent of transmission bias. The third term, $\bar{w}\,\mathrm{E}(\Delta g)$, is identical to the expectation term in the modified Price equation and is independent of differential fitness. The second term, $\mathrm{Cov}(w,\Delta g)$, is sensitive to variation in fitness *and* to variation in transmission bias. The terms are interpretable as quantifying, respectively, the effects selection has independently of biased transmission, the effects biased transmission has independently of selection, and the effect of directional selection on transmission bias (Figure 2.1).

When $\mathrm{Cov}(w,\Delta g) = 0$, the differences between the standard, modified and three-term Price equations collapse: all three provide quantitatively identical partitions of the overall change. But when $\mathrm{Cov}(w,\Delta g) \neq 0$, only the three-term version provides a complete causal decomposition of the effects of selection and transmission, since both the standard and modified Price equations have too few terms to separate the distinct effects of selection, transmission and the interaction of the two processes. The standard Price equation treats the change due to the interaction of selection and transmission as if it were attributable to transmission alone; while the modified Price equation treats this effect as if it were attributable to selection alone. By adding a third term explicitly representing the interaction of selection and transmission, we avoid both types of causal misattribution.

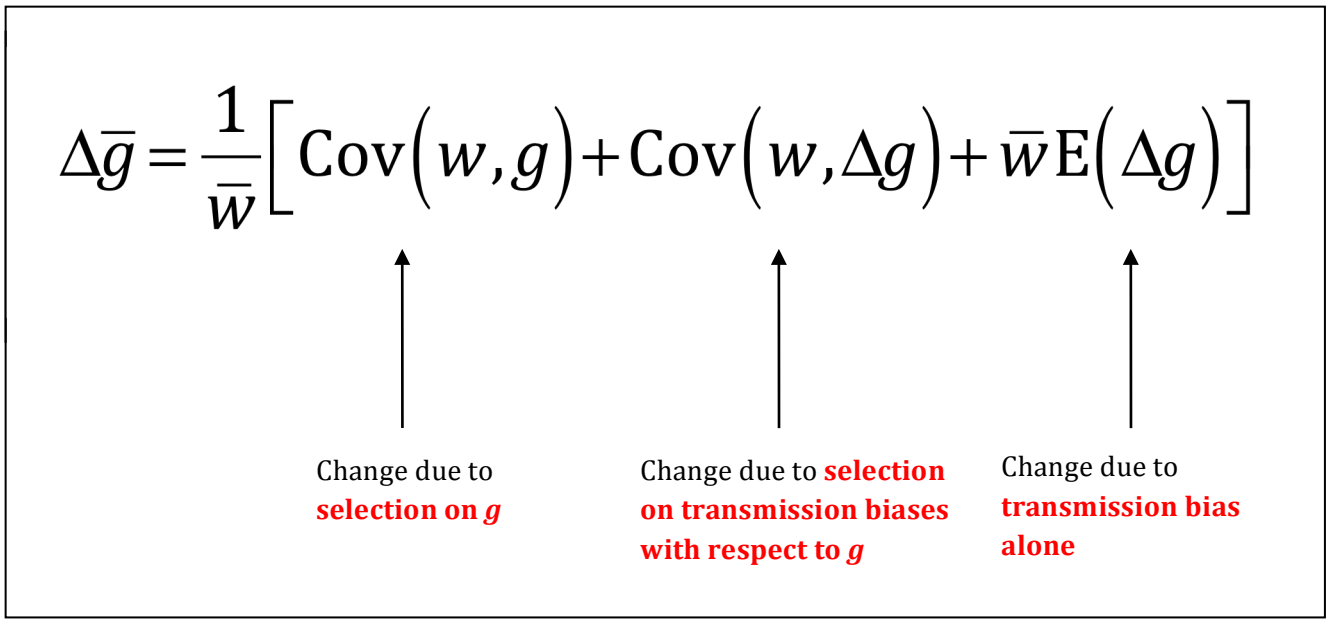


Figure 2.2: The three-term Price equation, with its associated causal interpretation.

Under what ecological conditions should we expect to find a non-zero $\mathrm{Cov}(w,\Delta g)$ term? Such circumstances may be rare (cf. Okasha 2011), but they are certainly not inconceivable. Imagine, for example, a population in which a mutation arises that disposes its bearer to help its parents raise additional offspring. Let g represent the breeding value for this cooperative trait; and suppose that, owing to the appearance of the mutation in their offspring, some parents have positive values of Δg. Now suppose that these parents receive a fitness benefit by virtue of their positive value for Δg, since their offspring help them produce additional offspring. This effect would show up in the Price equation in the form of a positive $\mathrm{Cov}(w,\Delta g)$ term.[11]

3.3.2 A further complication

The three-term Price equation distinguishes two different ways in which natural selection may cause the population mean of a character, z, to increase. One is fairly intuitive: if

[11] Another possible case in which $\mathrm{Cov}(w,\Delta g) \neq 0$, involving horizontal gene transfer, is considered in Chapter 5.

individuals with the genes for z tend to have more offspring than individuals without those genes, then (in the absence of a countervailing transmission bias) those genes will increase in frequency, and this may issue in a positive change in $\bar{z}$. This effect is captured in $\mathrm{Cov}(w,g)$, the covariance between an individual's fitness and its breeding value for z. The other effect is much less intuitive: if individuals whose offspring have a greater breeding value than their own tend to enjoy increased fitness as a result, this too can issue in a positive change in $\bar{z}$. This effect is captured in $\mathrm{Cov}(w,\Delta g)$, the covariance between an individual's fitness and its personal transmission bias with respect to the character of interest. The latter effect seems likely to be much rarer than the former, and smaller when it obtains, but it remains a mathematical and conceptual possibility. I suggest we call the former effect the *primary* effect of natural selection, and call the latter effect the *secondary* effect of natural selection.

That is not quite the end of the story. For there is a *third* way in which natural selection may affect $\bar{z}$. To see what it is, we need to return to the definition of breeding value. A breeding value, recall, is a sum of allelic values weighted by their *average effects* (sensu Fisher) on the character of interest. When introducing the notion of a breeding value, we noted briefly that the average effects of an allele can be altered by changes in allele frequency in the presence of non-additive interactions (i.e., dominance or epistasis) between alleles. In essence, this is because non-additive interactions imply that the difference an allele makes to the phenotype of its bearer is *context-dependent* (cf. Sterelny and Kitcher 1988; Okasha 2006). For instance, if an allele A dominates an allele a at a particular locus, the effect of adding a second copy of A to a diploid individual who already has one copy will not be the same as adding a copy of A to a diploid individual with no copies. Whenever allelic effects are context-dependent, the *average* effect of a substituting one allele for another will depend on the *relative frequency* with which a copy of that allele finds itself in one genetic context rather than another—and this in turn will depend on the overall allele frequencies in the population. Since changes in allele frequency can be brought about by natural selection, the implication is that natural

selection can alter the weightings that determine the breeding values of individuals in the descendant-set (cf. Okasha 2008). This change in average effects may in turn produce a change in $\bar{g}$ (and, by implication, $\bar{z}$). I will call this the *tertiary* effect of natural selection, though this is not intended to imply that it will be any smaller or less important than the secondary effect.

If such an effect occurs, it will not be accounted for in $\mathrm{Cov}(w,g)$. Some of it may be accounted for in $\mathrm{Cov}(w,\Delta g)$, since it is possible that the alleles subject to changes in their average effects will be differentially possessed by fitter (or less fit) individuals. But some of this change is likely to be independent of fitness differences; and the Price equation will account for this portion in the expectation term, $\bar{w}\mathrm{E}(\Delta g)$, which we had originally hoped to interpret as a term attributable to 'transmission bias alone'. Of course, there is a sense in which a change in the average effects of alleles due to natural selection *is* a source of a transmission bias, for it impairs an individual's ability to transmit its breeding value faithfully to its descendants. But there is also a sense in which this label is misleading, since nothing about the process of genetic transmission is responsible for this effect. It comes about not because of any bias in the transmission of alleles, but rather because the breeding value, by definition, requires us to weight alleles by their average effects, and these average effects can change between generations. And it seems particularly misleading to attribute this change to transmission bias *alone*, given that a change in average effects may be due to changes in gene frequency caused by natural selection.

One might see this further complication as a reason to regard breeding values with suspicion. For it implies that, when we formulate the Price equation in terms of breeding values, even the three-term version fails to provide the clean separation of effects we hoped for: it is quite possible for natural selection to influence the expectation term as well as the two covariance terms (Box 2.1 summarizes the overall picture). We could remove this possibility by switching to *p*-scores, which are not phenotype-relational, and do not require us to weight alleles by their changeable average effects. Yet we need not see this

feature of breeding values as a drawback. Indeed, seen in a different light, it actually provides further justification for using breeding values rather than 'raw' allelic values or *p*-scores in the formulation of fundamental theory. For it also suggests that, if we were to look only at the evolutionary change in allele frequencies or average *p*-scores, and assume a simple relationship between alleles and phenotypes, we would be liable to overlook the tertiary effect of natural selection on phenotypic change. This effect is real, and any model of phenotypic change that aims for causal completeness should accommodate it.

Box 2.1: The three effects of natural selection on the evolution of z

- Natural selection may bring about covariance between the genes for z and fitness. This is the *primary effect* of natural selection on the evolution of z. It is quantified by $\mathrm{Cov}(w, g)$, the covariance between an individual's fitness and its breeding value for z.
- When transmission fidelity is imperfect, the difference an individual's personal transmission bias makes to the overall change in z depends on the number of descendants it leaves. Because of this, selection can have a *secondary effect* on the evolution of z. This effect is captured in $\mathrm{Cov}(w, \Delta g)$, the covariance between an individual's fitness and its personal transmission bias.
- Natural selection may also have a *tertiary effect* on the evolution of z when changes in allele frequency alter the average effects of alleles. This effect may contribute towards $\mathrm{Cov}(w, \Delta g)$, but it may also contribute towards $\bar{w}\mathrm{E}(\Delta g)$.
- When we talk informally about the 'effect of natural selection', we should be clear regarding which of these effects we have in mind.

3.3.3 Analysing partial change

At the start of his *Genetical Theory of Natural Selection* (1930), Ronald A. Fisher famously remarks that 'natural selection is not evolution'. Selection *contributes* to evolutionary change, but it is not the whole story: other processes—notably, mutation, migration and genetic drift—contribute too. In this sense, theories of 'social evolution' are inaptly named, because they are usually theories about the conditions under which *natural selection* will favour a social behaviour, rather than theories about the conditions under which social behaviours will in fact evolve. As we see in the next chapter, it is often useful when formulating such a theory to focus on the component of the overall evolutionary change attributable to natural selection (or rather, natural selection at a particular level or levels, since we also usually want to ignore change attributable to within-organism intragenomic conflict or gametic selection). To do this is *not* to assume that other factors are insignificant. It is simply to *abstract away* from them, so as to focus on the *partial* change caused by the process we take to be largely responsible for the evolution of cooperation.

The reason for the Price formalism's rise to prominence in recent decades is that, at least on the face of it, it allows theorists to identify this partial change in a form that makes it a convenient target for further analysis. The moral of Section 2.3.2, however, was that the true picture is somewhat more complicated. The term in the Price equation that is usually taken to represent the partial change attributable to natural selection, $\mathrm{Cov}(w,g)$, in fact represents only the *primary effect* of natural selection. There are two further ways in which natural selection may influence evolution which are not accounted for in this term. Nevertheless, in some contexts we may be chiefly interested in determining the conditions under which the primary effect of natural selection will favour a social behaviour; and, in these contexts, $\mathrm{Cov}(w,g)$ will be the correct target for analysis. I will introduce the symbol $\Delta_{1^\circ}\bar{g}$ to denote this partial change:

$$\Delta_{1^\circ}\bar{g} = \frac{1}{\bar{w}}\left[\mathrm{Cov}(w,g)\right] \tag{3.3.3}$$

In other contexts, we will be able to gain additional insight into the effects of selection by taking the secondary effect into account, and by taking $\mathrm{Cov}(w, g')$ as our target of analysis (see Chapter 4; see also Frank 1997a, 1998). I will introduce the symbol $\Delta_w \bar{g}$ to denote this partial change, since it represents the component of the overall change that directly depends on fitness differences:

$$\Delta_w \bar{g} = \frac{1}{\bar{w}}\left[\mathrm{Cov}(w, g')\right] \tag{3.3.4}$$

These partial changes will be a frequent target of analysis in subsequent chapters. Sometimes one sees these partial changes denoted by the subscript 'S' or 'NS', to indicate that they reflect the partial change attributable to natural selection. I avoid this here because it is misleading: it encourages us to neglect the tertiary effect of natural selection, an effect that is not accounted for by the covariance term in the Price equation but that may still make a significant contribution to the overall evolutionary change (cf. Chapter 6).

3.4 Grouping organisms

When we derived the Price equation in Section 3.1, we framed the entire discussion in terms of the properties of *individuals*. No attempt was made to sort organisms into groups, types or classes of any kind. In some ways, the fact that we *can* formulate the Price equation in purely individualist terms is important (see Grafen 1985a; Godfrey-Smith 2009a). In practice, however, whenever biologists *actually use* the Price equation as the starting point for the analysis of a (real or modelled) population, they more commonly group organisms together in one way or another, so as to focus their attention on the average properties of groups. In this section, I unify the various ways in which one might go about grouping organisms under a common formal framework; I use this framework to

show why certain methods of grouping are particularly useful; and I relate this discussion to philosophical issues.

In Section 3.4.1, I show (following Price 1972a) how, provided we can partition a population into non-overlapping subsets, it is always possible to partition the overall *w*-*g* covariance into between- and within-subset components. In Section 3.4.2, I consider three applications of this general principle: trait-groups of interacting organisms, genotypic classes, and developmental classes. The next two subsections bring the preceding discussion to bear on philosophical questions. Section 3.4.3 considers the subtle relationship between the Price formalism and 'evolutionary nominalism', the view that sorting organisms into classes is never obligatory in evolutionary theory (Godfrey-Smith 2009a). Section 3.4.4 turns to the troubled relationship between kin- and group-selectionist approaches to social evolution. I suggest that the main methodological difference between these approaches lies not in *whether* organisms are sorted into groups for the purpose of analysis, but *how*.

3.4.1 The general case

The overall covariance between *w* and *g* is defined as the expected product of individual deviations from the population mean with respect to each variable:

$$\mathrm{Cov}(w,g)=\mathrm{E}\left[(w-\bar{w})(g-\bar{g})\right]$$

As with any expectation value, this quantity may be computed in a variety of ways. Here is one possible procedure: first, add up the values of $(w-\bar{w})(g-\bar{g})$ for each individual in the ancestor-set; then, divide this number by the total number of individuals. This procedure is reflected in the following expression:

$$\mathrm{Cov}(w,g)=\frac{1}{n}\sum_i (w_i-\bar{w})(g_i-\bar{g}) \tag{3.4.1}$$

It would be equally reasonable, however, to compute the covariance by means of the following, three-step procedure:

1. Sort the members of the ancestor-set into *N* non-overlapping subsets (the subsets need not correspond to natural groupings of any kind; in principle, individuals can be assigned to subsets arbitrarily).
2. For each subset, compute the average value of $(w-\bar{w})(g-\bar{g})$ within that subset.
3. Take an average of the subset averages, weighting each subset by its relative size.

To rewrite equation (3.4.1) in a way that reflects this alternative averaging procedure, we can re-label the members of the ancestor-set. Instead of labelling them with a single index, *i*, we can label them with two indices, *i* and *j*, such that g_{ij} represents the breeding value of the i^{th} member of the j^{th} subset, and w_{ij} represents the fitness of the i^{th} member of the j^{th} subset. We can then replace $\sum_i$ with $\sum_{ij}$, indicating that we are to sum over all *j* for each value of *i* (i.e., over all entities in each subset) and then sum over all *i* (i.e., over all subsets). Finally, we can define a quantity $q_i = m_i/n$, where m_i represents the size of the i^{th} subset; q_i thus represents the *relative* size of the i^{th} subset. Combining these ingredients, we can rewrite equation (3.4.1) as follows:

$$\mathrm{Cov}(w,g)=\sum_{ij}\frac{q_i}{m_i}(w_{ij}-\bar{w})(g_{ij}-\bar{g}) \tag{3.4.2}$$

Equations (3.4.1) and (3.4.2) give us two equivalent expressions for the overall *w*-*g* covariance. The former is the simpler of the two, but the latter still has its uses. In

particular, it is possible to partition the summation in equation (3.4.2) as follows, where G_i represents the *average* breeding value of the *i*th subset, and W_i represents the *average* fitness of the *i*th subset (see Appendix A for details):

$$\mathrm{Cov}(w,g)=\sum_{ij}\frac{q_i}{m_i}\left(w_{ij}-W_i\right)\left(g_{ij}-G_i\right)+\sum_i q_i\left(W_i-\bar{w}\right)\left(G_i-\bar{g}\right)$$

Both terms on the right-hand side of this equation can be given a statistical interpretation. The first term is the *size-weighted expectation* of the *w*-*g* covariance within each subset, while the second term is the *size-weighted covariance*[12] between the subset averages, *W* and *G*. We can therefore rewrite the equation in statistical notation (where the *m* subscripts indicate weighting by size, and Cov^i denotes the within-subset covariance of the *i*th subset):

$$\mathrm{Cov}(w,g)=\mathrm{E}_m\left[\mathrm{Cov}^i(w,g)\right]+\mathrm{Cov}_m(W,G) \tag{3.4.3}$$

This general result, first presented by Price (1972a), bifurcates the overall covariance into two components, the first of which depends only on genetic variation *within* subsets, and the second of which depends only on genetic variation *between* subsets. Nothing is assumed about the nature of these subsets, except that they do not overlap.

---

[12] The notions of 'weighted expectation' and 'weighted covariance' are used frequently in Price's (1970, 1971, 1972a) original papers, but are rarely seen in other contexts. In standard probability theory, *all* computations of expectation and covariance involve weighting outcomes by their probabilities, so this is not explicitly mentioned; and any other weightings are included in the arguments of E and Cov, rather than being incorporated into the functions themselves. Price's stipulative definitions are as follows, where E_k denotes '*k*-weighted expectation' and Cov_k denotes '*k*-weighted covariance':

$$\mathrm{E}_k(X)=\mathrm{E}\left[\left(k/\bar{k}\right)X\right]$$

$$\mathrm{Cov}_k(XY)=\mathrm{E}_k\left[\left(X-\mathrm{E}_k(X)\right)\left(Y-\mathrm{E}_k(Y)\right)\right]$$

3.4.2 Three special cases

Section 3.4.1 was deliberately abstract. We saw that, when we sort the members of our ancestor-set into subsets, we can partition the overall *w-g* covariance into a component that depends on genetic variation within subsets, and a component that depends on genetic variation between subsets; but we said nothing at all about the *biological interpretation* of these 'subsets'. The reason is that biologists sort organisms into groups in a variety of ways for a variety of theoretical purposes, and considering the general case allows us to bring all these cases within a unifying framework. In this section, I want to consider three such cases: groups of interacting organisms, genotypic classes, and developmental classes.

First, however, it will be helpful to introduce a formal framework in which different ways of grouping organisms can be conceptualized and compared. As we have emphasized, equation (3.4.3) holds for *any* partition of a population into non-overlapping subsets; in principle, the subsets can be completely arbitrary. But in practice, little insight is gained by grouping organisms in an arbitrary fashion: we want to group organisms non-arbitrarily. In broad terms, the way to do this is by identifying a biologically meaningful *equivalence relation* among the members of the population under study.[13] In set theory, an equivalence relation, $x \sim y$, can be any binary relation among the elements of a set that is reflexive, symmetric and transitive. If we can find a relation with these properties for a given set, we can use it to partition the set into non-overlapping subsets such that each comprises elements related to each other by $x \sim y$. These subsets are known as *equivalence classes*. For example, suppose we have a set of balls of varying colours. We can identify an equivalence relation for this set:

$$(x \sim y) = (x \textit{ is the same colour as } y)$$

[13] To my knowledge, the first author to apply this notion specifically to the problem of grouping organisms was Godfrey-Smith (2006), and here I am indebted to his illuminating discussion.

Note that the relation is reflexive, symmetric and transitive. We can then use this relation to partition the set of balls into equivalence classes, where each equivalence class comprises all and only those balls of a particular colour.

Case 1: Groups of interacting organisms

In biology, we are rarely interested in grouping organisms by colour. We often are interested, however, in grouping them by *patterns of interaction*. Sometimes it is possible to partition a population into discrete 'trait-groups', where the members of each trait-group engage in fitness-affecting interactions (with respect to the character(s) of interest) only with their fellow group members. We can think of 'trait-groups' as equivalence classes defined by the following equivalence relation (see Godfrey-Smith 2006):[14]

$$(x \sim y) = (x \textit{ has its fitness affected by the character of } y)$$

Grouping by fitness-affecting interactions is a standard approach in the literature on multi-level selection (see, e.g., Price 1972a; Hamilton 1975; Wilson 1975; Wade 1985; Queller 1992a; Sober and Wilson 1998; Pepper 2000; Okasha 2006; Gardner and Grafen 2009). Indeed, this was the application for which Price (1972a) originally derived his partition of the *w-g* covariance into between- and within-subset components. For ease of analysis, it is often assumed that the trait-groups are equal in size. In this special case, we can replace Price's somewhat confusing 'weighted expectation' and 'weighted covariance' functions with their standard, unweighted equivalents:

$$\mathrm{Cov}(w,g) = \mathrm{E}\left[\mathrm{Cov}^{i}(w,g)\right] + \mathrm{Cov}(W,G) \tag{3.4.4}$$

[14] Such a partition will not be possible for all populations; see Section 2.5.4.

Why group organisms in this way? One reason is that the equivalence classes one obtains will undoubtedly seem more 'natural' than purely arbitrary groupings. But another reason is practical. If we grouped organisms arbitrarily, any subsequent analysis of the covariance would have to take account of interactions between the members of different subsets. After all, we could not rule out the possibility of such interactions, nor could we dismiss them as irrelevant to the overall response to selection. By contrast, grouping organisms by patterns of relevant interaction allows us to discount any interactions that cut across group boundaries. For it ensures that, at least with respect to the character of interest, the fitness of a given organism is affected only by its fellow group members; and that the average fitness of the group depends only on how its members behave. In other words, it guarantees that all interactions relevant to the response to selection will take place within groups—not across them.

Case 2: Genotypic classes

While it is often useful to sort organisms by patterns of interaction, this is not the only way in which we may wish to sort organisms for the purposes of analysis. Here is another possibility: we could sort organisms into subsets by their *breeding value* for a character/s of interest, such that there is exactly one subset for each value of g instantiated in the population, and each subset contains all and only those members of the population which instantiate that value. We can call these subsets *genotypic classes*.[15] Genotypic classes, which are again relative to the character under study, are defined, for a given character, by the following equivalence relation:

$$(x \sim y) = (x \textit{ has the same breeding value as } y)$$

[15] It may be that all members of a genotypic class have the same alleles at all loci relevant to the trait of interest, but this need not be the case: a given breeding value might be 'multiply realizable' by various allele combinations.

Since every individual within a subset has the same breeding value, an individual's breeding value always equals the average breeding value of its subset, i.e., $g_{ij} = G_i$ for all i, j. This equality eliminates the first term in equation (2.1.5), yielding:

$$\mathrm{Cov}(w,g) = \mathrm{Cov}_m(W,G) \tag{3.4.5}$$

In effect, sorting organisms into genotypic classes entitles us to assume that *the within-subset covariance will be zero in all subsets*, so that all that remains is the size-weighted covariance between the class averages for fitness and breeding value.

Sorting organisms into genotypic classes—in order to track variation in the average properties of genotypes rather than variation in the properties of individuals—is extremely common in the kin selection literature. This is in part because kin selection theory gives paramount importance to considerations of *genetic relatedness* at relevant loci (see Chapters 4 and 5). Usefully, we know that the relatedness among any two members of a genotypic class will always equal 1; moreover, we know that the relatedness between a member of one genotypic class and a member of another will be the same for all possible pairings of one member from each class. Hence, by sorting organisms into genotypic classes, we save ourselves the trouble of having to track degrees of relatedness between individual actor and recipient pairs: if we know the *average* relatedness between the actor's genotypic class and the recipient's genotypic class, this tells us everything we need to know (see Frank 1997b, 1998; see also Chapter 5).

Sorting organisms by genotype is a standard practice in population genetics more generally. This too is understandable, since the functional relationship between an individual's genes and its fitness will often be enormously complex. Individuals with the same breeding value for the trait of interest will often end up with very different fitness

values, depending on the other traits they inherit, the environment in which they develop, and the chance events that befall them during their life. It can be very helpful to abstract away from all this micro-level variation in fitness among individuals with the same genotype. Equation (3.4.5) shows that we are formally entitled to do this, for it shows that fitness differences *within* genotypic classes, no matter how dramatic, have no effect at all on the overall *w-g* covariance. All that matters is the variation in average fitness *between* genotypic classes.

Case 3: Developmental classes

The third main type of equivalence class arises when organisms are grouped by some important, non-genotypic property that is functionally distinct from the character we are studying, but that nonetheless has profound consequences for an organism's fitness and/or behaviour. For example, when generations overlap (as they do in insect societies), it is helpful to sort organisms into age classes; when sexes differ in their ploidy (as they do in insect societies), it is helpful to sort organisms by sex; and when organisms exhibit significant morphological differentiation that impacts on their reproductive and behavioural capacities (as they do in insect societies), it is helpful to sort them by morphological caste. I will refer all these forms of equivalence class as *developmental classes*; since, in one way or another, they all group organisms by the developmental features they instantiate, be it the temporal stage of development they are passing through or the morphology they have come to exhibit. Like genotypic classes, developmental classes are defined by a similarity-based equivalence relation, though in this case it is some relevant *developmental* similarity that matters:

$$(x \sim y) = (x \textit{ is in the same age-group/sex/caste as } y)$$

Developmental classes, like genotypic classes, are commonly encountered in kin selection models. This is because the kin selection approach requires that we track the fitness effects

of particular social behaviours, and there are many cases in which a particular social behaviour will yield different payoffs depending on the developmental class of the actor and recipient. For instance, in an age-structured population, an altruistic act that benefits a very old recipient will tend to confer a smaller fecundity benefit than it would confer if it were to fall on a younger recipient; conversely, an altruistic act performed by a very old recipient will tend to impose a smaller cost than it would have done if it were performed by a younger actor with more to lose.

For ease of analysis, kin selection theorists typically assume that there is no genetic variance between developmental classes at the relevant loci (see Frank 1997b, 1998). This is often a reasonable assumption. When classes are differentiated by age, there is usually no particular reason why the overall genotypic composition of one age class would be different to that of any other; the same applies when classes are differentiated morphologically, provided the relevant differences arise from differential gene expression (i.e., phenotypic plasticity) rather than from genetic differences; and the same also applies when classes are differentiated by sex, provided the genes related to social phenotypes are not correlated with the genes for sex determination. Of course, one can construct hypothetical scenarios in which this assumption would fail, but it will hold in many cases. If we succeed in individuating classes such that there is no genetic variance between classes, we can eliminate the *second* term in equation (3.4.3), yielding, in Price's statistical notation:

$$\mathrm{Cov}\left(w,g\right)=\mathrm{E}_{m}\left[\mathrm{Cov}^{i}\left(w,g\right)\right] \tag{3.4.6}$$

Verbally, the overall covariance is equal to a size-weighted (i.e., frequency-weighted) average of the within-class covariance. A version of this expression, which was (to my knowledge) first derived by Peter Taylor (1990), is frequently deployed in the kin selection literature when a problem requires explicit accommodation of class-structure (e.g, Taylor

1990; Taylor and Frank 1996; Frank 1997b, 1998; Wild and Taylor 2006; Taylor et al. 2007; Wenseleers et al. 2004; Wenseleers et al. 2010).

The three types of equivalence class we have considered provide cross-cutting ways of grouping organisms: they are very unlikely to ever align with one another.[16] Nevertheless, for any partition of a population into equivalence classes, we are free to treat each equivalence class as a new population in its own right, and partition it into yet smaller equivalence classes. As a result, there are various ways in which the three types of class could be nested within each other. For instance, we could partition a population into trait-groups, then separately partition each trait-group into developmental classes, then separately partition each developmental class of each trait-group into genotypic classes. In practice, the usual procedure in the kin selection literature is to index organisms purely by genotype and developmental class; while the usual procedure in the multi-level selection literature is to index organisms purely by patterns of interaction. To my knowledge, the possibility of combining all three types of equivalence class in a single model has not yet been explored.

3.4.3 The Price formalism and evolutionary nominalism

I now want to bring the foregoing discussion of equivalence classes to bear on two (related) philosophical matters. Godfrey-Smith (2009a) suggests that the Price formalism is a natural ally of a view he terms *evolutionary nominalism*:[17]

[16] This would require that each developmental class is genotypically unique, that the members of each class all have the same genotype, and that organisms only interact with other members of their class.

[17] A similar view is defended by Nanay (2010). Nanay, however, frames his view in terms of property-types and property-tokens; specifically, he argues that 'biological property-types do no play any explanatory role in evolutionary explanations' (2010, 93), a view he terms 'trope nominalism'. This may or may not be more radical than Godfrey-Smith's formulation, depending on whether or not the trope theorist can sanction the various equivalence relations discussed in Section 2.5.2 (cf. Footnote 10).

> [T]he grouping of individuals into types is in no way essential to Darwinian explanation. Such groupings are convenient tools. But one always has the choice of using finer or coarser groupings, ignoring fewer or more differences between individuals. As categories become finer, they may be occupied by only one individual each. (Godfrey-Smith 2009a, 35)

One moral from the foregoing discussion is that, while Godfrey-Smith is broadly correct to highlight an affinity between the Price formalism and evolutionary nominalism, the true relationship between the two ideas is not straightforward.

We saw earlier in the chapter that the Price equation may be formulated purely in terms of individuals and their properties, without acknowledging equivalence classes of any kind (Section 3.1). The result is an extremely general and strictly individualist description of the evolutionary change in some character. This naturally leads to the suggestion that the Price equation vindicates evolutionary nominalism. The equation, however, is not always formulated in individualist terms. Notably, Frank (1995, 1997a, 1998, 2012) derives the Price equation purely in terms of the properties of *types* (typically genotypes), where each type is weighted by its frequency in the computations of covariance and expectation; we then regard w as a measure of the total number of descendants each type contributes to the descendant-set. This formulation is no less correct than the more familiar version expressed in terms of properties of individuals, provided we can individuate types such that there is no variation within each type (and hence no covariance with w) with respect to the property of interest. In effect, rather than starting (as I have done) from a purely individualist formulation and extending it to accommodate various forms of equivalence class, Frank takes for granted the identity given in equation (3.4.5) and works with types from the outset.

In a sense, therefore, the Price formalism is promiscuous: it permits both individual-centred and type-centred formulations. Moreover, we saw in the preceding section that, in addition to explaining why it is permissible and useful to sort organisms by genotype, the Price formalism also explains why it is permissible and useful to sort them by causal or developmental equivalence relations. In all three cases, sorting organisms into equivalence classes allows us to highlight the variation that matters to the response to selection, while abstracting away from variation that does not: sorting by patterns of fitness-affecting interaction allows us to discount interactions that cross subset boundaries; sorting by breeding value allows us to discount variance in fitness within each class; and sorting by developmental properties, if they are chosen well, allows us to discount the variance in fitness between classes.

The upshot is that, if we construe evolutionary nominalism in the strongest possible sense—as the thesis that evolutionary theory can *and should* proceed without sorting organisms into equivalence classes—then the Price formalism is not the ally it may at first appear. For while it shows that an individualist description of evolutionary change is always available, it *also* shows why sorting individuals into equivalence classes often facilitates a more convenient description. It is doubtful, however, whether anyone would seriously defend so a strong a version of the thesis[18]; and Godfrey-Smith certainly does not. We can more cautiously formulate Godfrey-Smith's claim as the two-part thesis that (i) evolutionary explanations never *require* that we sort organisms into equivalence classes; but that (ii) there is often a *pragmatic justification* for doing so. If we construe evolutionary nominalism like this, the Price formalism provides support for both parts.

[18] Nanay 2010 (see footnote 17) may be an exception, though this is not clear. On the face of it, nothing prevents trope-nominalists from admitting equivalence relations (including causal relations and similarity relations) into their ontology. They would, however, deny that a similarity relation between two objects is reducible to their co-exemplification of a common property-type.

3.4.4 The Price formalism and the 'kin selection' versus *'group selection' debate*

For the past five decades, the question of the relationship between kin and group selection has been one of the most divisive issues in evolutionary theory (see Borrello 2010 for a historical overview). John Maynard Smith (1964) originally coined the term 'kin selection' to describe a process he took to be distinct from group selection; and this view of kin- and group-selectionist approaches as rivals—a view promoted by, among others, George Williams (1966) and Richard Dawkins (1976, 1979, 1982)—persists in many quarters to this day. But theorists have recognized for some time that the true relationship between the two approaches must be rather more subtle, since kin and group selection models appear to yield identical predictions under a wide range of conditions, and also seem to have similar limitations (Hamilton 1975; Grafen 1984; Queller 1992a; Wilson and Dugatkin 1997).

In recent years, a consensus-of-sorts has emerged that the two frameworks are 'formally equivalent', in the sense that they will never disagree regarding the direction of the response to selection (see Lehmann et al. 2007; Wenseleers et al. 2010; Gardner et al. 2011; Marshall 2011a,b; for dissent, see van Veelen 2009, 2011; Nowak et al. 2010; van Veelen et al. 2012). This is certainly correct if one is comparing the most general partition of the Price equation deployed in multi-level selection theory (i.e., equation (3.4.3)) with the most general version of Hamilton's rule deployed in kin selection theory (see 'HRG' in Chapter 4). Yet, for those of us who believe there ought to be room for both approaches in mainstream social evolution theory, this 'equivalence' result is a double-edged sword. On the plus side, if kin and group selection are, at a very general level, formally equivalent, then it is futile to argue over which approach is objectively superior in general: we can let both flowers bloom, at least in principle. But the downside is that theorists who have hitherto completely ignored one or other of these approaches will consider themselves entitled to carry on doing so. This is often the case in the current kin selection literature, where multi-level selection is typically introduced as a 'formally equivalent' alternative

only so that it can be set aside for serious explanatory purposes (see West et al. 2007a, 2008; Bourke 2011; Birch 2012b). The theoretical pluralist faces a new challenge: given that, at a very general level, the partitions of the Price equation employed by kin and group selection theory are formally equivalent, how exactly do the theories *differ* in a way that justifies us in retaining and developing both? Is there still any interesting methodological difference between the two approaches, or should we regard them as nothing more than redundant notational variants of the same theory and jettison one for the sake of the other?

One often encounters the suggestion that the theories of kin and group selection, though formally equivalent, offer usefully different 'perspectives' on social evolution. The apparent implication is that, while they may not constitute substantively different theories as to how social evolution proceeds, the differences between them are not simply notational. But how exactly should we cash out this idea? One might assume that the difference lies in the fact that kin selection, in contrast to group selection, provides a fundamentally *individualist* perspective on social evolution—a perspective that puts the individual at the centre of the analysis, and avoids any sorting of organisms into groups. But I think this is a mistake, and the preceding sections illustrate why. Though it is possible to formulate the basic idea of kin selection in strictly individualist terms (see Gardner et al. 2011; see also Chapter 4), detailed modelling of kin selection almost never holds to strict individualist scruples. When applying kin selection theory to particular problems, theorists almost always sort organisms into genotypic classes, developmental classes, or both.

With this in mind, I suggest that one fundamental difference between kin- and group-selectionist approaches to social evolution consists not in *whether* one sorts organisms into subsets for the purposes of analysis, but *how*. A kin selection analysis prioritizes considerations of genetic and developmental similarity in assigning organisms to subsets. Organisms are indexed to genotypic and developmental classes, so that variation within genotypic classes and between developmental classes can be ignored. The kin selectionist

then studies the ways in which organisms of different classes interact with one another, with a view to ascertaining which genotypes are likely to be favoured by selection within each developmental class. By contrast, a multi-level analysis prioritizes considerations of causal interaction from the beginning. Organisms are indexed to a particular subset on the grounds that they interact only with the other members of that subset. Considerations of genetic and developmental similarity are thus trumped by considerations of causation: if two organisms have the same genotype or belong to the same developmental class but never interact with one another, the group selectionist will not assign them to the same subset. The group selectionist then analyses the effects of selection in two parts. She studies how within-group differences in character cause differences in individual fitness, and studies separately how differences in the average character of groups cause differences in their average fitness.

This contrast shows how the frameworks of kin and multi-level selection can be something less than rivals, but something more than redundant systems of notation. The differences between the approaches are not particularly substantive, because we will often be able to switch from a kin selectionist analysis to a group selectionist analysis simply by re-labelling the individuals in the population, in the same way that we can switch perspective as we view a duck-rabbit, or a Necker cube (cf. Dawkins 1982; Godfrey-Smith and Kerr 2002, forthcoming). But the differences between the two frameworks are not merely notational, because this form of re-indexing will not *always* be possible. We can identify, at least in the abstract, cases in which only one of the two methods of indexing will work.[19]

The multi-level method of indexing will not always be available, because *x has its fitness affected by the character of y* is not always reflexive, transitive and symmetric (Godfrey-

[19] In Chapter 5, I make a similar point regarding the 'neighbour-modulated fitness' and 'inclusive fitness' frameworks within kin selection theory. Here, I am considering 'kin selection theory' in the broadest possible sense, glossing over the interesting differences between alternative kin selectionist approaches.

Smith 2006, 2008). In other words, it is not always possible to sort organisms into non-overlapping subsets such that each subset contains all and only those organisms which engage in fitness-affecting interactions with other members of the subset. So-called 'neighbour-structured' populations, in which each organism interacts with its nearest neighbours, often have this property; since it is not true in general that my neighbours' neighbours are my neighbours (Maynard Smith 1964, 1976, 1987, 2002; Godfrey-Smith 2006, 2008). For a simple illustration (adapted from Godfrey-Smith 2006), imagine a population with a spatial structure that can be represented by a square lattice. Each organism occupies one node on the lattice, and no node is unoccupied; and each organism interacts with all and only those organisms on the four adjacent nodes. In this scenario, *x has its fitness affected by the character of y* is not transitive: it is not true in general that, if some organism A has its fitness affected by an organism B, and B has its fitness affected by a third organism C, then A has its fitness affected by C (in fact, this is true only if $A = C$).

On the face of it, the equivalence relations we need in order to individuate genotypic and developmental classes (e.g., *x has the same breeding value as y*, *x is the same age as y*) seem less likely to fail the conditions of reflexivity, symmetry and transitivity. But the kin selection method of indexing *might* still lead to difficulties, particularly when developmental classes are employed. For recall that, for equation (3.4.6) to apply, we need to be able to identify developmental classes such that there is *no genetic variation between classes with respect to the character of interest*, and this may not always be possible. For instance, we can imagine cases in which a population is age-structured, and in which bearers of a particular allele at a relevant locus tend to die off more rapidly than non-bearers—so that the genotypic composition of the older classes differs from the genotypic composition of the younger classes. In such cases, we could still assign organisms to classes, but we would not be able to eliminate the between-class portion of the overall w-g covariance. As a result, we would not be able to employ any methods of analysis that start from equation (3.4.6), including the usual methods of neighbour-modulated and inclusive

fitness analysis for class-structured populations (see Taylor 1990; Taylor and Frank 1996; Frank 1998; Taylor et al. 2007; see also Chapter 5).

Can we say anything about which method of grouping is in general preferable, when both options are available? Partisans on both sides are likely to insist that their preferred equivalence relations are more 'natural' or 'meaningful' than the alternative: the ardent kin selectionist will claim that classes defined by genetic and developmental similarity are more natural than groups defined by patterns of causal interaction; the ardent group selectionist will claim the reverse. But it is hard to see how either party could substantiate its claim to superior 'naturalness'. Both approaches involve abstracting, from the great milieu of objective relations in which organisms stand to one another, some subset of relations that is particularly salient to the problem at hand. Which approach is superior in any particular instance is likely to depend on the precise nature of the problem. This conclusion, of course, is in keeping with the modest brand of evolutionary nominalism advanced in Section 6.4.3.

FOUR

The Scope and Limits of Hamilton's Rule

Chapter 2 examined the ecological aspects of social behaviour, and the concepts we use to sort behaviours into types. Chapter 3 examined an abstract formalism for the representation of natural selection—the Price formalism—that provides the foundation for much of contemporary social evolution theory, but which by itself says nothing in particular about the evolution of social behaviour. The most important bridge from the abstract world of population genetics to the real world of behavioural ecology is Hamilton's rule, a deceptively simple statement of the conditions under which we can expect a social behaviour to be favoured by natural selection. The rule states, broadly speaking, that a social behaviour will be favoured by natural selection if and only if $rb - c > 0$, where b represents the benefit the behaviour confers on the recipient, c represents the cost it imposes on the actor, and r represents the relatedness between actors and recipients. To describe Hamilton's (1964) presentation of the rule as 'enormously influential' would be an understatement: 48 years and 9056 citations later,[2] Hamilton's rule remains a result of paramount importance both to theorists, for whom it is the foundational principle of kin selection theory, and to field biologists, for whom it is a versatile rule of thumb with which to rationalize social behaviours observed in the wild.

Yet despite (or perhaps, in part, because of) its great influence, Hamilton's rule has proved a powerful magnet for controversy and debate.[3] The reason, in a nutshell, is that Hamilton

[2] According to Google Scholar, as of 03/09/12.

[3] For a recent example, see Nowak, Tarnita and Wilson's (2010) incendiary claim that Hamilton's rule 'almost never holds'; a claim fiercely rebutted by 157 social evolution theorists (Abbot et al. 2011). I discuss

first derived the rule in a one-locus population-genetic model that made a number of substantive modelling assumptions, including weak selection, fair meiosis, random mating, the absence of mutation and the additivity of genic effects on fitness. In the following decades, many theorists (including Hamilton himself) explored the extent to which these assumptions could be relaxed. The upshot was a variety of routes to '$rb - c > 0$'-type results, often with apparently incompatible implications about the conditions under which the result obtains.[4]

The Price formalism provides a route to a particularly general formulation of Hamilton's rule (Hamilton 1970; Queller 1985, 1992a, b, 2011; Grafen 1985a, b; Frank 1998; Gardner et al. 2007; McElreath and Boyd 2007; Wenseleers et al. 2010; Gardner et al. 2011; Birch forthcoming). Yet even among proponents of the Price formalism, disagreement persists regarding the rule's scope and limits. In particular, a dispute about the consequences of synergy for Hamilton's rule has divided two of the most influential living theorists of social evolution—David C. Queller of Washington University in St Louis and Alan Grafen of the University of Oxford—for almost thirty years. In the 1980s, Queller argued that the familiar '$rb-c>0$' form of Hamilton's rule must be extended in cases of synergistic interaction between social partners (that is, cases in which the effect of two individuals performing a social behaviour differs from the sum of the two effects each behaviour would have had in the absence of the other) (Queller 1984, 1985). Grafen replied that this was simply incorrect: the standard version of the rule still holds, he argued, regardless of whether synergy is present or absent (Grafen 1985a, b). One might expect this to be a mathematical issue that could be resolved quickly and definitively, one way or the other. For reasons discussed below, however, matters are not so straightforward and, three

---

this particular controversy in detail in a separate article (Birch forthcoming). That article can be seen as a companion piece to this chapter, since they address closely-related issues.

[4] See, e.g, Hamilton 1970, 1972, 1975; Levitt 1975; Orlove 1975; Charnov 1977; Charlesworth 1980; Uyenoyama and Feldman 1980, 1981, 1982; Uyenoyama et al. 1981; Michod 1982; Toro et al. 1982; Cheverud 1984; Grafen 1984, 1985a.

decades on, the two theorists remain unchanged in their views (Queller, personal communication; Grafen, personal communication).

The dispute has taken on broader significance in the intervening years, since the question of whether Hamilton's rule can or cannot accommodate synergistic interactions now lies at the heart of a heated and ongoing debate concerning its usefulness and generality (see Fletcher and Zwick 2006; Gardner et al. 2007; van Veelen 2009; Nowak et al. 2010; Gardner et al. 2011; Marshall 2011b; van Veelen et al. 2012; Birch forthcoming). In recent literature, Grafen's insistence that the standard version of Hamilton's rule can accommodate synergy has been picked up and defended at length by a number of his Oxford colleagues (Gardner et al. 2007; Gardner et al. 2011), while Queller's note of caution on this score has (to Queller's evident frustration[5]) been picked up by theorists who would like to see the rule expunged altogether from serious theorizing (van Veelen 2009, 2011; van Veelen et al. 2012).

In this chapter, I dissect the long-running debate regarding the validity of Hamilton's rule in cases of synergy, and I propose an irenic resolution that identifies what is right about both positions in the original Queller/Grafen dispute. In Section 4.1, I provide the necessary theoretical background, presenting a derivation of Hamilton's rule from the Price equation that closely follows that of Queller (1985). In Section 4.2, I explain the problem synergy poses for the formulation of the rule derived in the preceding section. In Section 4.3, I critically examine Queller's proposed extension to the rule and more recent developments in a similar vein; in Section 4.4, I critically examine ways of preserving the original '$rb - c > 0$' form of the rule in the tradition of Grafen 1985b. Finally, in Section 4.5, I propose a resolution to the debate. I argue that, in choosing whether to employ generalized (two-term) or extended (three-or-more-term) versions of Hamilton's rule in social

[5] See his recent article on Joan Strassmann's 'Sociobiology' blog (URL: http://sociobiology.wordpress.com/2012/04/06/agreement-and-disagreement-in-social-evolution-insight-from-david-queller/ [Accessed 03/09/12]).

evolution theory, we inevitably face a trade-off between conceptual unification and causal explanation: generalized formulations buy unification at the expense of causal content, while extended formulations buy causal content at the expense of unifying power. The upshot is that the generalized and extended formulations of Hamilton's rule are apt to perform different theoretical functions, and may peacefully coexist. Problems arise only when we choose the wrong formulation for the task at hand.

4.1 The regression route to Hamilton's rule

4.1.1 Regression analysis of partial change

Partitioning the Price equation

The Price equation is a useful starting point for the analysis of social evolution (and, indeed, evolution more generally) chiefly because the covariance operator is bilinear, or linear in both arguments. This implies that, if we can write either one of the arguments as a function of other variables, we can partition the overall covariance into a sum of components, one for each of the terms in that function (for instance, $\mathrm{Cov}(\alpha X+\beta Y,Z)=\alpha\mathrm{Cov}(X,Z)+\beta\mathrm{Cov}(Y,Z)$, where α and β are constants). To put the point in less abstract terms: if we can express an individual's fitness as a function of relevant phenotypic predictors weighted by appropriate coefficients, then, by substituting this function into the Price equation, we can partition the overall *w*-*g* covariance into a sum of components, one for each of those predictors (see Lande and Arnold 1983; Queller 1992a, b).

Here is a simple illustration. We start with equation (3.3.3), which expresses the partial change due to the primary effect of natural selection as the covariance between an individual's fitness, *w*, and its breeding value, *g*:

$$\Delta_{1^\circ}\bar{g}=\frac{1}{\bar{w}}\left[\mathrm{Cov}(w,g)\right]$$

Next, we express *w* as a function of an individual's value for some phenotypic character, or set of characters. Let us suppose that *w* depends on two phenotypic characters, z_1 and z_2, and that it depends on them in a perfectly linear way:

$$w = \alpha z_1 + \beta z_2$$

Substituting this expression into the Price equation, and exploiting the bilinearity of covariance, we obtain:

$$\Delta_{1^\circ}\bar{g} = \frac{1}{\bar{w}}\left[\alpha \mathrm{Cov}\left(z_1, g\right) + \beta \mathrm{Cov}\left(z_2, g\right)\right]$$

This resolves the overall primary effect of natural selection into a term that depends on differences in z_1 and a term that depends on differences in z_2. Of course, the true fitness function will rarely (if ever) be this simple, and we have not yet introduced social effects; the purpose of the above example is merely to illustrate the bilinearity of covariance and the partitions of the Price equation it facilitates.

Residuals

It would be naïve to suppose that, in any real biological context, we could read off the exact fitness of each individual in a population simply by plugging its phenotypic characters into a linear fitness function. Usefully, however, the models of fitness we use in the analysis of social evolution do not have to be idealized in this way. We can relax this idealization by adding to our fitness function a residual term (ε_w) that, for any individual, represents the portion of its fitness that is not predicted by the other terms in the equation. For example:

$$w = \alpha z_1 + \beta z_2 + \varepsilon_w$$

Or, more generally, for *n* phenotypic characters, each weighted by a coefficient, β_i:

$$w = \sum_i^n \beta_i z_i + \varepsilon_w$$

Of course, by adding a residual term, we make any fitness function true by definition: there is no individual in any population, whether real or modelled, for which the equation is false. The fitness function will fit some populations better than others, however, and this will be reflected in the size of the residual terms. If the residuals are typically negligible, we have probably captured all the significant influences on fitness; if the residuals are typically enormous, we probably have not.

Partial regression coefficients

In the above fitness function, each phenotype, z_i, is weighted by a coefficient, β_i. While these coefficients could in principle be assigned arbitrarily or by guesswork, they are usually assigned in a more principled way. For any given set of phenotypic characters we might use as predictors of fitness, $\{z_1, z_2, ..., z_n\}$, there will be some corresponding set of weighting coefficients, $\{\beta_1, \beta_2, ..., \beta_n\}$, that minimizes the sum-of-squares of the residuals for the population. These special coefficients are known as the partial regression coefficients for this predictor set. Intuitively, weighting the phenotypic predictors by partial regression coefficients provides the overall 'best fit' between the variable we want to predict (i.e., fitness) and the variables doing the predicting (i.e., the phenotypic characters).

The numerical values of the partial regression coefficients are not, in the general case, easy to compute, since to do so we need to solve a set of *n* simultaneous equations, one for each predictor, and this task rapidly becomes very demanding as *n* increases (Ewens 2010). This does not, however, prevent partial regression coefficients from playing a fundamental role in the context of general, abstract theorizing about the nature of the evolutionary process, a context in which theoretical rigour and conceptual clarity often take precedence over tractable number-crunching (Fisher 1930; Ewens 2010; Gardner et al. 2011). We have

already encountered one important theoretical role for partial regression coefficients: they feature in the definition of a breeding value. By weighting allelic values by partial regression coefficients taken with respect to the phenotypic character of interest, we (by definition) obtain the best possible prediction of an individual's phenotype from a linear combination of its alleles, and this quantity, though difficult to compute exactly, is theoretically useful (see Chapter 2). For similar reasons, partial regression coefficients are also commonly used to weight phenotypic predictors in theoretical analyses of social evolution, where we want to define the best possible prediction of an individual's fitness from its behavioural phenotype.

If we want to predict fitness from two predictors or fewer, the relevant partial regression coefficients are much more easily computed. For a single predictor, 'partial regression' gives way to simple regression, and the simple regression of Y on X can be expressed as a ratio of the covariance between Y and X to the variance in X:[6]

$$\beta_{Y,X} = \frac{\mathrm{Cov}(Y,X)}{\mathrm{Var}(X)}$$

For two predictor variables (X_1 and X_2), the partial regression coefficients can be computed from the corresponding simple regressions by means of the following formulae (Lande and Arnold 1983; Queller 1992a; Gardner et al. 2011):[7]

[6] Here I introduce a common notation for simple regression coefficients, on which $\beta_{Y,X}$ represents the simple regression of Y on X.

[7] Here I introduce a common notation for partial regression coefficients, on which $\beta_{Y,X_1|X_2}$ represents the partial regression of Y on X_1, correcting for correlations with X_2. ρ_{X_1,X_2} is the Pearson correlation coefficient between X_1 and X_2 ($\rho_{X_1X_2} = \mathrm{Cov}(X_1,X_2) \Big/ \sqrt{\mathrm{Var}(X_1)\mathrm{Var}(X_2)}$).

$$\beta_{Y,X_1|X_2} = \frac{\beta_{Y,X_1} - \beta_{Y,X_2}\beta_{X_2,X_1}}{1-\rho^2_{X_1,X_2}}$$

$$\beta_{Y,X_2|X_1} = \frac{\beta_{Y,X_2} - \beta_{Y,X_1}\beta_{X_1,X_2}}{1-\rho^2_{X_1,X_2}}$$

These formulae, though specific to the two-predictor case, shed broader light on what partial regression coefficients are, and how they work. Importantly, although the partial regression of a dependent variable, *Y*, on a predictor, X_1, is sometimes glossed as a measure of the extent to which differences in X_1 predict differences in *Y* when we 'control for' correlated variables, talk of 'control' in this context is usually misleading. There is often no literal sense in which correlated variables are being 'controlled' or 'held fixed' by the analyst when partial regression coefficients are computed. Rather, the 'controlling' occurs purely in the statistics: we take the simple regression of *Y* on X_1 and adjust it, subtracting the portion of the overall association between *Y* and X_1 that is accounted for by correlation between $X_1$and the other predictors. The partial regression of *Y* on X_1 is therefore more accurately described as a measure of the extent to which differences in X_1 predict differences in *Y* when we 'correct for' or 'adjust for' correlations among predictors.

One final preliminary is needed: it will be convenient to introduce a symbol, $\mathbf{z}$, representing the complete set of phenotypic predictors we want to use in a particular regression analysis of fitness:

$$\mathbf{z} = \{z_1, z_2, z_3, ..., z_n\}$$

Naturally, the contents of $\mathbf{z}$ will vary greatly depending on the population and process in question. We may often want to regress fitness on only one or two phenotypic predictors, but in principle there is no limit to the number we could use.

4.1.2 The phenotypic formulation of Hamilton's rule (HRP)

When individuals do not interact socially, a causal analysis of the effects of natural selection can proceed solely by analysing how an individual's fitness depends on its own phenotypic characters. Correlations between characters will often matter to the overall direction of selection—and these correlations need to be taken into account in the analysis—but the characters in question will typically be intrinsic properties of the individual whose fitness they predict (Lande and Arnold 1983). By contrast, when individuals do interact socially, extrinsic properties matter too: to capture the causal influences on an individual's fitness with any degree of accuracy, we need to take into account not merely its own intrinsic character, but also its social milieu. We therefore need to include at least two phenotypic predictors in our predictor set. In addition to z, the focal individual's phenotypic value for the social trait under investigation,[9] we at the very least need to consider $\hat{z}$, the average phenotypic value of its social partners:

$$\mathbf{z} = \{z, \hat{z}\}$$

Partitioning the Price equation using this predictor set takes us directly to a useful version of Hamilton's rule (Queller 1992a). First, we express fitness as a sum of these predictors, weighted by the relevant partial regression coefficients (and including a residual term, ε_w):

$$w = \beta_{w,z|\hat{z}} z + \beta_{w\hat{z}|z} \hat{z} + \varepsilon_w \tag{3.1.1}$$

Recall that, owing to the residual term, this equation cannot be false of any individual in any population, though it may fit some populations much better than others. We then substitute this equation into equation (2.3.3), yielding the following partition:

[9] By 'focal individual', I mean the arbitrary individual whose fitness we are attempting to predict, as opposed to its social partners (Rousset 2004).

$$\Delta_{1^\circ}\bar{g} = \frac{1}{\bar{w}}\left[\beta_{w,z|\hat{z}}\,\mathrm{Cov}(z,g) + \beta_{w,\hat{z}|z}\,\mathrm{Cov}(\hat{z},g) + \mathrm{Cov}(\varepsilon_w,g)\right]$$

To get from this three-term partition of the Price equation to Hamilton's rule, we need to eliminate the third term—that is, we need to assume that $\mathrm{Cov}(\varepsilon_w,g)=0$. Queller (1992a) terms this assumption the 'separation condition' on the grounds that, if it obtains, the two-term partition of the Price equation that is left behind fully separates quantities that relate genotype to phenotype from quantities that relate phenotype to fitness.

There is some confusion in the literature as to whether or not the separation condition amounts to a substantive assumption. Queller (1992a) claims that it does, and I agree. It is true enough that least-squares theory guarantees that the residuals in a regression equation cannot co-vary with any of the predictors (a point I revisit in Section 6.4). At first glance, therefore, one might take it to be trivial that $\mathrm{Cov}(\varepsilon_w,g)=0$. Crucially, however, g is not among the predictors of fitness in our analysis: our predictors are phenotypes, whereas g is a breeding value. There can be no formal guarantee that a variable outside our predictor set will not co-vary with the residuals, so there can be no formal guarantee that g will not co-vary with ε_w when the predictors of fitness are phenotypes. The separation condition thus amounts to a substantive assumption. The circumstances under which it obtains are discussed at length by Queller (1992a), and will also be considered in detail in Section 4.2.

If the separation condition is indeed satisfied, we obtain the following simplified partition:

$$\Delta_{1^\circ}\bar{g} = \frac{1}{\bar{w}}\left[\beta_{w,z|\hat{z}}\,\mathrm{Cov}(z,g) + \beta_{w,\hat{z}|z}\,\mathrm{Cov}(\hat{z},g)\right]$$

This result entails the following condition for a social trait to be favoured by the primary effect of natural selection[10]:

[10] Here I use 'iff' in the standard philosophical sense, as an abbreviation for 'if and only if'.

$$\Delta_{1^\circ}\overline{g} > 0 \text{ iff } \beta_{w,z|\hat{z}}\,\mathrm{Cov}\left(z,g\right)+\beta_{w,\hat{z}|z}\,\mathrm{Cov}\left(\hat{z},g\right)>0$$

Owing to the definition of breeding value, $\mathrm{Cov}\left(z,g\right)=\mathrm{Var}\left(g\right)$.[11] Since variance cannot be negative, we can divide the right-hand side of this inequality through by $\mathrm{Var}\left(g\right)$ without any risk of reversing its sign. The result is the following rule:

$$\Delta_{1^\circ}\overline{g} > 0 \text{ iff } \beta_{w,z|\hat{z}}+\beta_{w,\hat{z}|z}\frac{\mathrm{Cov}\left(\hat{z},g\right)}{\mathrm{Var}\left(g\right)}>0 \qquad (3.1.2)$$

This result has the form of Hamilton's rule. The first term ($\beta_{w,z|\hat{z}}$) measures the association between a social behaviour and the fitness of the actor who performs it, and can be regarded as a generalized measure of the cost of that behaviour. The first part of the second term ($\beta_{w,\hat{z}|z}$) measures the association between an individual's fitness and the behaviour of its social partners, and can be regarded as a generalized measure of the benefit an individual receives from its neighbours. This quantity is weighted by a third coefficient ($\mathrm{Cov}\left(\hat{z},g\right)/\mathrm{Var}\left(g\right)$ or, equivalently, $\beta_{\hat{z},g}$) that measures the overall association between the breeding value of an individual and the character of its social partners. This can be regarded as a generalized measure of the 'relatedness' between actors and recipients (Orlove and Wood 1978; Michod and Hamilton 1980; Seger 1981; Michod 1982; Queller 1985; Grafen 1985a). Hence, to get from inequality (3.1.2) to the more familiar $rb-c>0$ form of Hamilton's rule, we need only relabel the coefficients as follows:

$$\beta_{w,z|\hat{z}}=-c \qquad \beta_{w,\hat{z}|z}=b \qquad \frac{\mathrm{Cov}\left(\hat{z},g\right)}{\mathrm{Var}\left(g\right)}=r$$

Yielding:

[11] By definition, the simple regression of a phenotypic character on its breeding value is 1 (if it were anything other than 1, the breeding value would not be the best possible prediction of the phenotype from a linear combination of allelic values, since we could improve the prediction by multiplying through by $\beta_{z,g}$); and $\beta_{z,g}=\mathrm{Cov}\left(z,g\right)/\mathrm{Var}\left(g\right)=1$ implies that $\mathrm{Cov}\left(z,g\right)=\mathrm{Var}\left(g\right)$.

$$\Delta_{1^\circ}\bar{g} > 0 \text{ iff } rb - c > 0 \qquad \text{(HRP)}$$

For reasons that will become clearer later on, I will refer to this as the phenotypic formulation of Hamilton's rule (HRP). Note that, although talk of costs and benefits intuitively connotes that costs will detract from an agent's fitness while benefits increase it, this need not be the case: the rule is intended to apply regardless of the sign of *b* or *c*. Hence, while the rule is most often associated with the evolution of cooperation (for which *b* is positive) and the evolution of altruism (for which *b* and *c* are both positive), selfish, spiteful and mutualistic behaviours are also intended to fall within the scope of Hamilton's rule (see Hamilton 1964; Trivers 1985; Bourke and Franks 1995; West et al. 2007; Bourke 2011).

4.1.3 The regression definition of relatedness

In the above formulation of Hamilton's rule, relatedness is formally defined as the regression of an individual's social partners' average phenotype on its own breeding value for the character under study. This regression definition of relatedness is highly abstract and highly general. It also has several important but rather counterintuitive implications, and it is worth briefly noting these for future reference. Relatedness coefficients and their various forms will be discussed in greater detail in Chapter 5.

Relatedness is not genealogical kinship

As we noted in Chapter 1, a high value of *r* does not require kinship in the traditional, genealogical sense of the word. What matters is correlation between the breeding value of the recipient and the phenotype of the actor. Kinship is one way of generating such correlations, but it is by no means the only way. As Hamilton (1975) himself notes, the necessary correlation could be ensured by, for instance, genetically-correlated habitat preferences, or 'greenbeard'-style recognition mechanisms that allow the bearers of a particular social gene to seek out and detect other bearers, whether or not they are

genetically similar in other respects (cf. Okasha 2002; Godfrey-Smith 2009; West and Gardner 2010). If genealogical kinship is the only source of relatedness then, on the assumption of weak selection,[13] relatedness coefficients between classes can be usefully approximated by traditional pedigree measures (e.g., for diploid individuals: ½ for offspring and full siblings; ¼ for grandoffspring and half siblings, etc.). But nothing in the definition of relatedness requires that we measure it this way (Frank 1998; Gardner et al. 2011).[14]

Relatedness is character-specific

Since breeding values and phenotypic values are strictly character-specific, relatedness too can be evaluated only relative to a particular character. This raises the conceptual possibility that two individuals might be closely related with respect to one character, yet only weakly related with respect to others; greenbeard effects provide one possible mechanism by which this kind of scenario could arise. It has often been suggested that such a pattern of relatedness would be unstable over evolutionary time, owing to the intragenomic conflict to which it could potentially give rise (Ridley and Grafen 1981; Grafen 1985a; Okasha 2002; Biernaskie et al. 2011). Even so, we should not discount the possibility of such variation, particularly in microbial populations, where horizontal gene transfer is known to issue in ephemeral and highly character-specific correlations between individuals (see Rankin et al. 2011a, b; Birch 2013b; see also Chapter 5).

---

[13] Strong selection distorts family trees, leading to correlations between relatives that differ from traditional pedigrees (cf. Frank 1998; Gardner et al. 2011). For example, if my siblings would probably have died years ago had they not inherited a particular gene, then the probability that my living siblings share that gene will be greater than ½.

[14] The possibility of relatedness without genealogical kinship has led to some uncertainty regarding how the term 'kin selection' should be applied. Some authors use the term broadly, to describe any selection process in which relatedness matters; others reserve the term for only those cases in which relatedness arises through genealogical kinship. As explained in Chapter 1, I will employ the term in the broader sense.

Relatedness is not 'shared genes'

Strictly speaking, a high value of *r* does not even require that the actor and recipient must share genes for the character under investigation. Recall that, since the breeding value is an estimate of a phenotypic character on the basis of any relevant alleles, it is possible (at least in theory) for two organisms to possess the same breeding value while possessing very different underlying allele combinations—combinations which just happen to have similar average effects on the phenotypic character in question. Because relatedness is formally defined in terms of breeding values and phenotypes, and not allelic values, individuals may, in principle, be closely related according to the regression definition and yet differ considerably at the more fine-grained level of particular alleles (cf. Fletcher and Doebeli 2009).

Relatedness can be negative

Any regression coefficient can be negative as well as positive, and relatedness, on the regression definition, is no exception. But in what sort of biological scenario could a negative value of relatedness arise? It would be one in which social partner phenotypes are negatively correlated, so that an individual with the genes for the social behaviour under investigation is less likely than average to interact with a social partner that performs it, while an individual without the relevant genes is more likely than average to interact with a social partner that performs it. Hamilton (1970) suggested that such a scenario would be conducive to the evolution of spite: by inflicting harm on their social partners, even at a cost to themselves, individuals with the genes for spite could increase the relative representation of these genes in the next generation. The regression definition of relatedness allows for such a phenomenon, since, when *r* is sufficiently negative, Hamilton's rule may predict that a social behaviour can be favoured by selection even if its effects on actor and recipient are also negative (see Gardner and West 2004a, b; West and Gardner 2010).

Relatedness is population-relative

On the regression definition, relatedness is a population-level statistic: it is a measure of the overall extent to which, in a given population, differences in breeding value predict differences in social milieu. Any evaluation of relatedness is thus relative to a reference population. One implication of this is that, strictly speaking, it makes no sense to talk of the relatedness between one particular individual and its social partner, because regression coefficients cannot be defined for a single data point. A second implication is that, if we want to use relatedness as a rough indicator of whether or not altruism is likely to evolve, the choice of an appropriate reference population is crucial. This point is especially salient in the case of 'viscous populations', in which organisms are confined to a particular locality and must compete for resources with other nearby organisms. When relatedness is evaluated relative to the population as a whole, viscosity tends to increase relatedness, because genealogical kin are confined to the same region. One might intuitively expect, therefore, that viscosity would be conducive to the evolution of altruism. Yet the fact that r is high when evaluated relative to the global population does not imply that r will also be high when evaluated separately for each local subpopulation. Organisms may be surrounded by kin, but that does not mean they will interact differentially with their closest kin—and it is differential interaction with genetically similar individuals, relative to the reference population mean, that matters to the value of r. The upshot of this is that, if competition within subpopulations is much stronger than competition between them, altruism might not be favoured after all (Taylor 1992; Queller 1992c). The general moral is that, if we want to use r-values as a guide to whether altruism will be favoured, we need to make sure that the reference population with respect to which they are computed is commensurate with the scale of competition (Queller 1994; West et al. 2002).

4.2 The problem of synergy

4.2.1 Why synergy matters

Synergy, as I use the term here, refers to any fitness effect that arises from a combination of social behaviours (performed by two or more individuals) and which is quantitatively different from (i.e., greater or less than) the sum of the fitness effects that those behaviours would have had if performed in causal isolation from each other. It is, more informally, a fitness effect that is either more or less than the sum of its parts, where the parts are the fitness effects the behaviours in question would have conferred by themselves.

Detecting synergy is not an easy business, partly because detecting quantitative fitness effects is never an easy business, and partly because we often have no observed instances of synergy-producing behaviours occurring 'in causal isolation from each other' to use as a contrast class. Nevertheless, there is a widespread consensus that synergy matters in social evolution—that many important social behaviours generate synergistic benefits, and that this often helps explain why they evolved in the first place (see especially Queller and Strassmann 1998; Strassmann et al. 2000; Fletcher and Doebeli 2006; Fletcher and Zwick 2006; Strassmann and Queller 2007, 2011; smith [*sic*] et al. 2010; Cornforth et al. 2012; Damore and Gore 2012). Chapter 1 introduced a number of examples of social interaction that may plausibly be regarded as synergistic, including slime mould aggregation; collective predation in myxobacteria; bubblenet feeding in humpback whales; and collective defence, construction and foraging tasks in eusocial insect colonies. Indeed, any instance of task-based cooperation will count as synergistic if the probability of task completion is a non-linear function of the number of contributions. In short, therefore, synergy is rife in nature—not only in the contexts of microbial evolution and of transitions in individuality, but also in sociobiology's more traditional entomological heartland. Any theory of social evolution with aspirations to serious explanatory power should be able to

accommodate synergy and its consequences for social evolution. If HRP cannot, that is a problem.

4.2.2 Why a one-predictor rule is unreliable when relatives interact

To see why synergy is often thought to present a problem for HRP, it is best to start with a different question: why do we need HRP at all? Why can we not predict the direction of selection with a simpler rule that uses a single phenotypic predictor, namely *z*, the focal individual's own value for the character under study? Consider the following regression equation:

$$w = \beta_{w,z} z + \varepsilon_w$$

By substituting this equation into the Price equation and (for now) assuming that $\mathrm{Cov}(\varepsilon_w, g) = 0$ (as in Section 3.2.2), we can derive the following principle:

$$\Delta_{1^\circ} \bar{g} > 0 \text{ iff } \beta_{w,z} > 0 \qquad (3.2.1)$$

The principle states that a phenotypic character will be favoured by selection if and only if there is a positive statistical association between the phenotype and fitness. One might suppose that the main problem with (3.2.1) is that it is just too simple to tell us anything interesting: in compressing all the causal influences on fitness into a single regression coefficient, it conflates direct and indirect causal pathways that HRP usefully separates. That is certainly true, but there is also a more serious problem: when relatives interact socially, (3.2.1) is liable to be downright false. This is because the separation condition we took for granted in its derivation (i.e., the assumption that $\mathrm{Cov}(\varepsilon_w, g) = 0$) is unlikely to be satisfied in real cases of social interaction.

In general terms, the condition will be satisfied only if the overall association between *g* and *w* is fully accounted for by, on the one hand, the association between *g* and *z*; and, on the other hand, the association between *z* and *w*. Accordingly, the separation condition may be usefully rewritten as $\beta_{w,g} = \beta_{w,z}\beta_{z,g}$. To judge informally whether this condition is likely to be satisfied in any given case, we must consider the following question: if we already knew an individual's value for *z*, would additionally learning its value for *g* enable us to predict its fitness with any greater accuracy? If it would, this can only be because the true value of $\beta_{w,g}$ is not fully accounted for by $\beta_{w,z}\beta_{z,g}$.

Queller (1992a) argues informally but cogently that, if genetic relatives interact socially, then knowing an individual's breeding value *does* provide predictively relevant information about its fitness over and above that provided by its phenotype. The key consideration here is that, in general, an individual's breeding value is a better predictor than its phenotype of the phenotype of its genetic relatives, because deviations from the breeding value are not genetically transmissible (and, hence, to the extent that an individual's phenotypic value deviates from its breeding value, the deviation is not likely to reappear in its relatives). When genetic relatives interact socially, this information becomes predictively relevant to *w*, because the phenotype of one's genetic relatives has consequences for one's own fitness. The upshot is that, in such cases, knowing the conjunction of an individual's breeding value and its phenotypic value tells us more about its fitness than knowing its phenotypic value alone. This implies that the separation condition is violated.

An example may help to bring out the logic of this argument. Consider a particular pair of organisms, *A* and *B*. They do not interact with each other, but both interact socially with their own genetic relatives. Both have exactly the same phenotypic value for some cooperative behaviour, *z*, but they differ in their fitness, *w*, and in their breeding value for that behaviour, *g*. Now consider the following question: if we know that *A* has the greater breeding value, does this tell us anything about which has the greater fitness? It does. The

information that *A* has the greater breeding value tells us that the breeding value of *A*'s relatives is likely to be greater than the breeding value of *B*'s relatives. This tells us that the phenotypic value of *A*'s relatives is likely to be greater than the phenotypic value of *B*'s relatives. And this tells us that *A* is more likely than *B* to receive a benefit from its social partners. Hence, *A* is likely to have the greater fitness.

4.2.3 Why HRP is unreliable when relatives interact synergistically

The failings of the one-predictor rule can be remedied simply by adding an extra predictor, $\hat{z}$, which explicitly represents the phenotype of the focal individual's social partners. This new predictor accounts for the component of the *w*-*g* covariance that *z* alone does not account for; and the result, of course, is HRP, which succeeds where the one-predictor rule fails. HRP faces a problem of its own, however, when genetic relatives interact synergistically (Queller 1985, 1992a, 2011).

The problem is broadly similar to the problem we encounter when we try to apply a one-predictor rule to additive social interactions. That problem arose because one's breeding value is often a more accurate predictor than one's phenotype of the (non-synergistic) social effects one can expect to receive, leading to a situation in which breeding value predicts residual fitness. A two-predictor analysis solves this problem, but runs straight into another: one's breeding value is often a more accurate predictor than one's phenotype of the synergistic social effects one can expect to receive, again leading to a situation in which breeding value predicts residual fitness. The main difference is that, this time round, the problem is not nearly as easy to see, because it arises from the precise way in which partial regression coefficients compensate for interactions among predictors.

To understand the nature of the problem, it will be helpful to consider simple one-shot, two-player, game-theoretic models of synergistic interaction (henceforth referred to as 'synergy games'), characterized by the following payoff matrix (Queller 1984, 1985):

	COOPERATE ($\hat{z}=1$)	**DEFECT** ($\hat{z}=0$)
COOPERATE ($z=1$)	$B-C+D$	$-C$
DEFECT ($z=0$)	B	0

We can translate the two strategies into the language of the Price formalism by defining a dummy variable, *z*, such that $z=1$ if the row-player cooperates and $z=0$ if it defects, and by defining a dummy variable $\hat{z}$ such that $\hat{z}=1$ if the column-player cooperates and $\hat{z}=0$ if it defects. The B, C and D parameters represent fecundity payoffs; nothing is assumed about their sign or magnitude. The D-payoff is what makes the model synergistic, since it implies that the payoff each player receives when both players cooperate differs from the sum of the payoffs that would be conferred by each player cooperating in isolation.

The payoff matrix does not fully specify a model, since models with the same payoff matrix can differ with respect to other parameters. For example, the strategies of interacting individuals may be correlated or uncorrelated, and their strategies may or may not be determined by their genotype. In Appendix C, I analyse two specific synergy games. In both games, some individuals possess a particular allele ($x=1$) while others do not ($x=0$). Social partner genotypes are correlated: a fraction, a, of individuals are assigned a social partner with an allelic value that is guaranteed to be identical to their own, while a fraction, $\left(1-a\right)$, are assigned a social partner that is drawn at random from the (infinite) population. The difference is that, in the first game, allelic value determines strategy: $x=0\rightarrow z=0$ and $x=1\rightarrow z=1$. In the second game, allelic value does not fully determine strategy: only a fraction, *k*, of $x=1$ individuals go on to express the cooperative

phenotype. This difference turns out to be critical, for it turns out that HRP is a reliable guide to the direction of selection in the first game but is not reliable in the second.

What is the source of the trouble? At first glance, one might think it obvious that HRP is going to break down in synergy games, on the grounds that it takes no account of the D-payoff (van Veelen 2009; van Veelen et al. 2012). This, however, misunderstands the meaning of the terms in HRP. The cost and benefit terms in HRP are partial regression coefficients, not fecundity payoffs, and as such they measure the overall statistical association between a predictor and fitness (correcting for other predictors), not merely the payoffs caused directly by that predictor. The implication is that, when computed correctly, they do take the D-payoff into account (Grafen 1985a, b; Gardner et al. 2007; Birch forthcoming). What they do, in effect, is account for the expected effects of synergy through a correction factor that we add to the B and C fecundity payoffs. In other words, they treat the synergistic payoff not as a third phenotypic pathway distinct from the costs and benefits of the behaviour in question, but rather as an effect that modulates these costs and benefits. If D is positive, the cost of the behaviour in question is lessened and its benefit is boosted; if D is negative, the converse is true (see Appendix C, Part I for details).

In fact, although synergistic interaction does pose a genuine problem for HRP, the source of the problem is far from obvious. The real issue is that, in scenarios in which genotype is an imperfect predictor of behaviour, synergistic interaction between genetic relatives leads to a violation of the separation condition (i.e., $\mathrm{Cov}\left(\varepsilon_w, g\right) \neq 0$), and this in turn implies that HRP is not a reliable guide to the sign and magnitude of the overall w-g covariance. Queller (1992a) argues for this conclusion informally, but provides no formal argument. In Appendix C (Part II), I show that Queller's intuition is indeed correct: when genotypes determine phenotypes, the separation condition is satisfied; but when genotypes are imperfect predictors of phenotypes, the separation condition is violated. More specifically, I show that HRP systematically overcompensates for the effects of synergy on the direction

of selection when phenotypes are not genetically determined. If synergistic effects are large, this systematic error threatens to make HRP seriously unreliable.[15]

While the formal argument in Appendix C is made in the context of a particular synergy game, it is important to add that the problem for HRP does not arise from any idiosyncratic features of this model. On the contrary, the problem is likely to recur in *any* scenario in which (i) genetic relatives interact synergistically and (ii) genotype is an imperfect predictor of phenotype. This is because the problem ultimately arises from a very general feature of partial regression coefficients. In correcting for the expected synergistic effect arising from interactions among predictors, a regression analysis only ever takes into account correlations among the predictors; any other biologically relevant correlations are ignored. This means that, if our predictors of fitness are phenotypic (and in real ecological contexts they usually need to be if we want to measure them; cf. Sections 4.4 and 4.5), only phenotypic correlations are taken into account when compensating for synergy, and any underlying genetic correlations are neglected. The upshot is that, whenever the underlying genotypic correlations between social partners are stronger than the manifest phenotypic correlations, knowing an individual's genotype will give us predictively relevant information about the synergistic effects it is likely to receive that a purely phenotypic regression analysis of fitness will fail to take into account. Hence, we have reason to believe that synergistic interaction among relatives will, in general, lead to violations of the separation condition. The synergy game analysed in Appendix C is merely an illustrative case.

[15] van Veelen et al. (2012) argue for a similar conclusion using similar models, but their understanding of Queller's separation condition strikes me as deeply confused. My aim in Appendix C is to show when synergy leads to a violation of the separation condition as Queller understands it.

4.3 Solution 1: Expand the predictor set

The problem of accommodating synergy within a kin selection framework is serious, but it is by no means insoluble. Indeed, one finds two different solutions in the kin selection literature, both of which have proved influential. The first—Queller's original (1984, 1985) solution—involves expanding our phenotypic predictor set to explicitly represent synergistic effects. The second—also developed by Queller (1992b), but more recently championed by Andy Gardner and colleagues (2011)—involves recasting Hamilton's rule as a principle concerned only with the average effects of genotypes, and ignoring phenotypic pathways altogether. Each solution has its merits and demerits. In the next two sections, I present and scrutinize each solution in turn. In Section 4.5, I address the question of which is on balance preferable.

4.3.1 Queller's extension of Hamilton's rule (HRQ)

Above, we saw how the deficiencies of a one-predictor regression analysis in cases of social interaction among genetic relatives can be remedied simply by adding an extra predictor to our predictor set, so that the phenotypes of an individual's social partners are explicitly represented. An analogous response is available in the present context: the deficiencies of HRP in cases of synergistic interaction among genetic relatives can be remedied by adding yet another predictor to our predictor set, explicitly representing the synergistic effect. Since the *D*-payoff in the synergy game obtains if and only if both players cooperate (i.e., $z\hat{z}=1$), the predictor we need to add is $z\hat{z}$, the product of the players' phenotypic values:

$$\mathbf{z}=\left\{z,\hat{z},z\hat{z}\right\}$$

From this predictor set we obtain the following regression equation (Queller 1985, 1992a, 2011):

$$w = \beta_{w,z|\hat{z},z\hat{z}} z + \beta_{w,\hat{z}|z,z\hat{z}} \hat{z} + \beta_{w,z\hat{z}|z,\hat{z}} z\hat{z} + \varepsilon_w$$

As always, the partial regression coefficients are defined as the weightings that minimize the sum-of-squares of the residuals. In the synergy game, the residuals will be minimized—indeed, eliminated altogether—when the three coefficients are equal to the $-C$, B and D fecundity payoffs respectively. Substituting the new regression equation into equation (2.3.3), we obtain the following partition:

$$\Delta_{1^\circ} \overline{g} = \beta_{w,z|\hat{z},z\hat{z}} \mathrm{Cov}\left(z,g\right) + \beta_{w,\hat{z}|z,z\hat{z}} \mathrm{Cov}\left(\hat{z},g\right) + \beta_{w,z\hat{z}|z,\hat{z}} \mathrm{Cov}\left(z\hat{z},g\right) + \mathrm{Cov}\left(g,\varepsilon_w\right)$$

Because the three-predictor regression fully accounts for the fitness of every individual in the synergy game—leaving no residuals at all—it follows that $\mathrm{Cov}\left(\varepsilon_w,g\right)=0$, so the separation condition is satisfied. This allows us to derive the following condition under which selection will favour cooperation in the synergy game (Queller 1985, 1992a, 2011):

$$\Delta_{1^\circ} \overline{g} > 0 \quad \text{iff} \quad \beta_{w,z|\hat{z},z\hat{z}} + \beta_{w,\hat{z}|z,z\hat{z}} \frac{\mathrm{Cov}\left(\hat{z},g\right)}{\mathrm{Cov}\left(z,g\right)} + \beta_{w,z\hat{z}|z,\hat{z}} \frac{\mathrm{Cov}\left(z\hat{z},g\right)}{\mathrm{Cov}\left(z,g\right)} > 0$$

Though broadly similar to Hamilton's rule, this condition includes an extra term corresponding to the effect of synergy. By using d to represent $\beta_{w,z\hat{z}|z,\hat{z}}$ and s to represent $\mathrm{Cov}\left(z\hat{z},g\right)/\mathrm{Cov}\left(z,g\right)$, we can produce the following, more memorable formulation:

$$\Delta_{1^\circ} \overline{g} > 0 \quad \text{iff} \quad rb - c + sd > 0 \qquad \text{(HRQ)}$$

This principle is sometimes known as Queller's rule (Marshall 2011), though Queller himself describes it as a natural extension of Hamilton's rule to the case of synergy. I will refer to it as Queller's extension of Hamilton's rule (HRQ).

4.3.2 The general method

Because it explicitly represents the effects of synergy, HRQ is more reliable than HRP when synergy is present: it accurately predicts the direction and magnitude of selection in cases in which HRP falls foul of the separation condition (and thus fails to account fully for the heritable variation in fitness). Yet it would be naive to suppose that HRQ constitutes a fully general extension of Hamilton's rule. We introduced a third predictor, $z\hat{z}$, in order to accommodate the additional complexity of the synergy game in contrast to a game with perfectly additive payoffs. But the synergy game is still very simple, in the great scheme of things, and many instances of social behaviour in the real world are likely to have far more complicated payoff structures. If we want to account for all the *w-g* covariance in these more complex contexts, then we are probably going to need more than three predictors.

The general moral to draw from the two cases discussed in Section 4.2, I suggest, is that our predictor set will fail to account fully for the overall *w-g* covariance whenever (i) we have too few predictor variables in our regression analysis to account for the full causal structure of the social interactions in the population under study, and (ii) an individual's genotype is a stronger predictor than its phenotype of the variable(s) we have omitted. Such a situation arises when we try to apply a one-predictor regression to social interaction among genetic relatives, and it arises again when we try to apply a two-predictor regression to synergistic interaction among genetic relatives. Queller's three-predictor regression copes with the synergy game only because its predictor set fully captures the causal influences on fitness in that game. If we were to apply HRQ to social interactions that have a greater degree of complexity than a three-predictor regression is able to represent, then there is every chance that the same problem would recur again: some of the *w-g* covariance would be unaccounted for, and the separation condition would be violated.

Queller (2011) is well aware of this problem, and does not claim that HRQ provides a fully general characterization of the conditions under which selection will favour a social behaviour. Instead, he suggests that we should see his derivation of HRQ merely as one instance of a general method for deriving Hamilton's rule-type principles to apply to particular cases. The general method can be characterized in four steps:[16]

> **STEP 1:** Construct a regression analysis of fitness including all phenotypic predictors causally relevant to the direction of selection on the character of interest. In cases of social behaviour, this will include extrinsic ('neighbourhood') predictors as well as intrinsic predictors.
>
> **STEP 2:** Substitute this regression equation into the (standard genetic) Price equation to yield a partition of the overall *w-g* covariance.
>
> **STEP 3:** Assume, if it is reasonable to do so, that the residuals in the regression analysis are uncorrelated with breeding value. This leaves behind a partition that cleanly separates quantities that relate genotype to phenotype from quantities that relate phenotype to fitness.
>
> **STEP 4:** Rearrange to derive a rule describing the conditions under which (the primary effect of) natural selection will increase the average value of the character of interest.

[16] Queller (2011) himself suggests that there are eight steps, because he individuates the steps more finely than I do. This is not a substantive difference.

Queller (2011) shows how to apply the general method to various particular cases. Variants of this method have also been fruitfully employed by Steven A. Frank (1997a, b; 1998; 2006) and by jeff smith [*sic*] and colleagues (2010).

4.3.3 Neighbour-modulated and inclusive fitness

The general method is non-specific enough to encompass a variety of quite different analytical approaches. One important ambiguity arises in Step 1: we are told to construct a 'regression analysis of fitness', but the meaning of fitness in the context of social evolution is deliberately left unspecified. The reason is that there are two influential bodies of theory falling within the purview of Queller's general method, but which conceive of social fitness in very different ways. These are the neighbour-modulated fitness (or direct fitness) and inclusive fitness approaches to the analysis of kin selection.

The intuitive difference between these approaches is easy enough to grasp. The neighbour-modulated fitness approach conceives of an individual's fitness in terms of its own reproductive output, and analyses the ways in which that output depends on the behaviour of its social partners (the above derivations of HRP and HRQ are simple examples of this approach). The inclusive fitness approach, by contrast, conceives of an individual's fitness as (roughly speaking) a sum of the fitness components for which its behaviour is causally responsible, where each component is weighted by the individual's relatedness to the organism that is doing the reproducing. It proceeds to analyse how an organism's fitness (thus construed) depends on its own behaviour.

Recent years have seen considerable debate as to whether or not the two frameworks constitute formally equivalent perspectives on social evolution (see, e.g., Frank 1998, 2006; Taylor et al. 2007; Fletcher and Doebeli 2009; Rosas 2010; Martens 2011). I will weigh into this debate in Chapter 5. For now, I merely want to note that both approaches involve

Steps 1-4 of the general method; the difference between them lies in the conception of fitness they take as the target of analysis. Hence, the general method is inclusive enough to accommodate not only a great plurality of predictor sets, but also a plurality of conceptions of social fitness.

4.3.4 Contextual analysis as a special case

It is also worth briefly noting the extremely close relationship between Queller's general method and the contextual analysis approach to multi-level selection (Heisler and Damuth 1987; Damuth and Heisler 1988; Goodnight et al. 1992; Okasha 2006). In essence, contextual analysis involves applying Queller's method with a predictor set that includes group-level properties. In the simplest case, we might only include Z, the average z-value of the focal individual's group:

$$\mathbf{z} = \{z, Z\}$$

From this predictor set, we can (applying Steps 1-3) derive the following partition of the w-g covariance:

$$\Delta_{1^\circ}\overline{g} = \beta_{w,z|Z}\,\mathrm{Var}(g) + \beta_{w,Z|z}\,\mathrm{Cov}(Z,g)$$

And, from this partition, we can (applying Step 4) derive a variant of Hamilton's rule in which the 'relatedness' is equal to $\mathrm{Cov}(g,Z)/\mathrm{Var}(g)$, a measure of the association between an individual's breeding value and the average z-value of the group in which it finds itself[17]:

[17] This quantity is sometimes known as the 'whole-group relatedness' and has received considerable attention in recent kin selection literature, particularly in the contexts of microbial and human cooperation (see, e.g., Pepper 2000; Frank 2006; El Mouden and Gardner 2008; Ross-Gillespie et al. 2009; El Mouden et al. 2010; Rankin et al. 2011; Cornforth et al. 2012).

$$\Delta_{1^\circ}\overline{g} > 0 \text{ iff } \beta_{w,z|Z} + \beta_{w,Z|z}\frac{\mathrm{Cov}\left(Z,g\right)}{\mathrm{Var}\left(g\right)} > 0$$

Of course, the same general method could be applied to more complicated predictor sets, including sets containing 'emergent' group characters (such as specialization or division of labour). Hence, although contextual analysis is usually considered to fall under the umbrella of multi-level selection theory, it might equally be regarded as a natural extension of Queller's general method for the analysis of kin selection to predictor sets that include group characters. Its ambiguous status shows just how close to one another kin- and group-selectionist approaches to social evolution have become.[18]

4.4 Solution 2: Bypass phenotypes

Queller first presented his extension to Hamilton's rule in two papers in the mid-1980s (Queller 1984, 1985). Shortly afterwards, in a paper entitled 'Hamilton's rule OK', Grafen replied that the proposed extension was unnecessary, since the standard version of the rule already accommodates synergistic effects (see also Grafen 1985a, 81-2):

> The third, synergistic term in Queller's form can be made to disappear by agreeing to define benefit and cost as the average

[18] In Section 3.4, I argued that the main difference between kin- and group-selectionist methods of analysis lies not in whether organisms are assigned to groups, but how. A multi-level analysis typically presupposes that organisms can be sorted into equivalence classes on the basis of their social interactions, while kin selection analyses typically presuppose that organisms can be sorted into genotypic classes, developmental classes, or both. Although I think this is broadly correct, contextual analysis muddies the waters somewhat, because it does not require that the groups to which it assigns *Z*-values are non-overlapping equivalence classes. For instance, suppose we have a population of *N* organisms subdivided into *N* groups. Each group is 'centred' on a particular individual, comprising all and only those individuals with whom it interacts (including itself). We can apply contextual analysis to this population by defining an individual's *Z*-value as average character of the group centred on it (cf. Godfrey-Smith 2006).

> effects on individuals' fitnesses, rather than as arbitrary terms in a model of fitness. (1985b, 311)

For Grafen, the illusion that Hamilton's rule breaks down in cases of synergy highlights the fact that 'care must be taken in applying Hamilton's rule' (1985a, 82). He adds, rather stingingly, that '[a] result has no interest as an exception to Hamilton's rule if it is based on the wrong interpretation of r, b and c' (1985a, 82).

There is a sense in which Grafen is right, and a sense in which he is wrong. As we noted in Section 3.2, one might naïvely assume that Hamilton's rule could not possibly accommodate synergy, because simple synergy games involve a D-payoff and Hamilton's rule makes no mention of a D-payoff (cf. van Veelen 2009, 2011; van Veelen et al. 2012). This, however, mistakes the partial regression coefficients in Hamilton's rule for fecundity payoffs in a payoff matrix. When computed correctly, the b and c coefficients do indeed take the D-payoff into account, as Grafen correctly points out. Nevertheless, owing to the way in which the partial regression coefficients are calculated, there is a systematic tendency for HRP to overcompensate for the effects of synergy on the response to selection. If the error is large, $rb-c$ may even be qualitatively inaccurate regarding the direction of partial change. As Queller (1992a) subsequently made clear (and as I argue more formally in Appendix C), this is the real problem for HRP in cases of synergy.

Even so, there is still a way in which the synergistic term in HRQ can legitimately be made to disappear in the spirit of Grafen's original proposal. We can do this by replacing phenotypic predictors with purely genetic predictors. By ignoring altogether the phenotypic pathways linking genotype and fitness, we can avoid failures of the separation condition and produce a version of Hamilton's rule that almost never fails. Queller himself was the first to point out this alternative response to the problem of synergy (Queller 1992b). In recent years, however, it has been defended chiefly by Grafen's Oxford colleagues (Gardner et al. 2007; Gardner et al. 2011).

4.4.1 Hamilton's rule with genetic predictors (HRG)

The derivation of the genetic version of Hamilton's rule is exactly parallel to the derivation of HRP, and can be regarded as yet another special case of Queller's general method. The only difference is that our phenotypic predictor set is replaced with a genetic predictor set (which we can denote with the letter $\mathbf{g}$) comprising the breeding value of the focal individual, *g*, and the average breeding value of its social partners, $\hat{g}$:

$$\mathbf{g} = \{g, \hat{g}\}$$

Applying Steps 2-4 of Queller's method, we can derive the following rule:

$$\Delta_{1^\circ}\overline{g} > 0 \text{ iff } \beta_{w,g|\hat{g}} + \beta_{w,\hat{g}|g}\frac{\mathrm{Cov}(\hat{g},g)}{\mathrm{Var}(g)} > 0$$

By using b_g to represent $\beta_{w,g|\hat{g}}$, $-c_g$ to represent $\beta_{w,\hat{g}|g}$ and r_g to represent $\mathrm{Cov}(\hat{g},g)/\mathrm{Var}(g)$, we can recast this into more familiar notation:

$$\Delta_{1^\circ}\overline{g} > 0 \text{ iff } r_g b_g - c_g > 0 \qquad \text{(HRG)}$$

I will refer to this as Hamilton's rule with genetic predictors (HRG). It differs from HRP in two respects. First, the b_g and c_g terms represent the average effects of genotypes, whereas the *b* and *c* terms in HRP represent the average effects of behavioural phenotypes. Second, the r_g term represents the correlation between social partner genotypes, whereas the *r* term in HRP represents the correlation between actor phenotypes and recipient genotypes.

4.4.2 On the generality of HRG

As with any other application of Queller's method, the derivation of HRG requires an assumption of uncorrelated residuals, that is, $\mathrm{Cov}(\varepsilon_w, g) = 0$. As we have seen, this assumption is often substantive and contentious. In the case of HRG, however, it is guaranteed to hold. This is because g is now among the predictors in our regression analysis, and the method of least-squares (i.e., minimizing the sum-of-squares of the residuals) ensures that the residuals in a regression equation do not co-vary with any of the predictors (Queller 1992b). In light of this, we could say that HRG cannot possibly fail the separation condition. Yet it would be rather more accurate to say that the separation condition does not even apply to HRG, because HRG does not even attempt to separate quantities which relate genotype to phenotype from quantities which relate phenotype to fitness. In effect, it bypasses phenotypes altogether: it considers only the overarching relationships between fitness and breeding value, without any care for how they are mediated phenotypically.

The reward for bypassing phenotypes is a principle of extraordinary generality. After all, $\mathrm{Cov}(\varepsilon_w, g) = 0$ is the only substantive assumption in the derivation of HRP, and in the derivation of HRG this assumption is trivially satisfied. The upshot is that HRG, as a statement of the statistical conditions under which selection will favour a social behaviour, is true of any population to which the Price equation applies, and in which the relevant partial regression coefficients are well defined.

Accordingly, there are only two kinds of case in which HRG can fail. First, there are cases in which HRG is false because the standard Price equation is also false. Such cases are likely to be extremely rare, but they are not inconceivable, because the derivation of the Price equation involves a substantive assumption to the effect that all descendants have the same number of ancestors (see Kerr and Godfrey-Smith 2009; see also Section 2.1). Second, there are cases in which HRG is false because, although social behaviour can still evolve,

the coefficients in HRG are undefined. Such cases may be rather less rare, since the terms in HRG are undefined whenever g and $\hat{g}$ are perfectly collinear (to see why, note that the formula for partial regression coefficients given in Section 3.1.1 yields an undefined value if $1-\rho^2=0$, and perfect collinearity implies that $\rho^2=1$). It is not hard to envisage scenarios in which such collinearity might arise, particularly in populations of asexually reproducing microbes. Yet, on the face of it, these scenarios seem conducive, not hostile, to the evolution of social behaviour. The fact that HRG does not apply to such cases shows not that social behaviour can never evolve in them, but rather that HRG is not a fully general condition for the evolution of social behaviour.[19]

4.5 The solutions compared

We now have two solutions to the problem of synergy on the table. Solution 1 is to expand the predictor set to more accurately represent the causal structure of the phenotypic pathways linking genotype to fitness. Solution 2 is to bypass phenotypic pathways altogether so as to recast Hamilton's rule as a purely genetic principle. Which should we prefer? In this section, I want to argue that we do not have to choose—or rather, we do not have to choose one solution to apply across the board. The general moral of the problem of synergy, I submit, is that there are two explanatory functions traditionally assigned to Hamilton's rule. Each solution to the problem prioritizes one of these functions but neglects the other, and there is no way for a single principle to satisfy both at once. This gives us grounds for an irenic resolution to the debate: neither solution to the problem gives us everything we might want, but both solutions give us something worth having.

[19] HRP faces a similar problem, but less severely. The coefficients in HRP are still well defined when social partners have identical genotypes, provided there is still some phenotypic variation between social partners with respect to the characters under study.

4.5.1 The dual role of Hamilton's rule

One might intuitively expect that the primary role of Hamilton's rule would be predictive—that biologists would estimate its coefficients in order to predict the social behaviours that natural selection will have built. In fact, the power of Hamilton's rule (in any form derived from the Price equation) to predict the long-run trajectory of social evolution is extremely limited. There are two main reasons for this. The first is that the rule concerns only the direction of (the primary effect of) social selection—it avoids analysing any other relevant influences on evolution, including mutation, genetic drift and intragenomic conflict. We can therefore use Hamilton's rule to predict the overall direction of social evolution only on the assumption that these effects are insignificant. The second reason is that the r, b and c coefficients represent aggregate statistical properties of a population, and these properties are liable to change as gene frequencies change. The implication is that, even if Hamilton's rule is satisfied for a particular social behaviour at a particular time, it will not necessarily be favoured by social selection at a later time (cf. Birch forthcoming).

If the pathways linking breeding value to fitness are frequency-independent (i.e., genes have frequency-independent effects on phenotype, and phenotypes have frequency-independent effects on fitness), then there is no particular reason why the direction of social selection would change as gene frequencies change. In this special case, Hamilton's rule can serve as a rough guide to whether a trait will go to fixation in the long run. But frequency-independence is likely to be the exception rather than the rule in real-world social evolution. Notably, frequency-dependent effects can be generated by synergy (as in the synergy game) or by dominance or epistasis among genes, and both kinds of phenomenon are biologically commonplace. When effects are frequency-dependent, Hamilton's rule will only ever provide a static 'snapshot' of a dynamic process. The direction of kin selection may well change over evolutionary time and, if we want to

understand *how* it changes, we have no choice but to build a more concrete model of the evolutionary dynamics (cf. Grafen 1985a,b; Frank 1995, 1998, 2012; Traulsen 2010).

These limitations do not, however, render Hamilton's rule explanatorily useless. Explanation in evolutionary biology is not all about long-run predictive success, and there are important explanatory roles to which Hamilton's rule, in spite of its predictive limitations, remains well suited. Two such roles stand out as especially important: that of providing causal explanations of observed evolutionary outcomes, and that of unifying diverse dynamical models under a common conceptual framework.

Causal explanation in the field

When we encounter social behaviours in nature, we can usually be fairly confident that they evolved at least in part because they were favoured by natural selection, but this by no means tells us everything we want to know. We also want to know why they were so favoured. In particular, we want to know whether they were favoured by virtue of their effects on the actor performing them, or whether they were favoured by virtue of their effects on other individuals. The standard way to answer this question is to estimate the coefficients in Hamilton's rule (Grafen 1985a). By estimating the value of r, b and c for a given population, a behavioural ecologist can assess firstly whether the rule is satisfied, and secondly how the terms compare. They can thereby make an inference as to why the trait was originally favoured by selection.

Since Hamilton's original (1964) derivation of the rule, numerous empirical studies have put this method into practice (e.g., Grafen 1984; Gadagkar 2001; Oli 2003; smith et al. 2010; Waibel et al. 2011; for reviews, see Foster 2009; Westneat and Fox 2010; Bourke 2011). Importantly, however, the method is usually only practicable if the cost and benefit coefficients in Hamilton's rule are understood as average effects of phenotypes rather than of breeding values. This is because, while we can usually gather data on the behaviours of particular organisms and their fitness consequences, it is rarely possible to gather data on

the genotypes of particular organisms outside of the laboratory. Indeed, as Grafen (1985a) notes, this is the main reason why Hamilton's rule—in contrast to the concepts and methods of classical population genetics—has come to be so influential among behavioural ecologists:

> In applications to data, Hamilton's rule comes into its own. The great differences from models are that usually with data on social traits, the genotypes of individuals are unknown and the genetic system controlling the trait is unknown. This makes worries about dominance, number of loci and mode of gene action purely academic. In modelling, the fundamental population genetics method of finding the number of offspring of each genotype is the main rival to Hamilton's rule. This alternative simply cannot be applied to data if the genotypes of individuals are unknown. Hamilton's rule can be applied, provided enough information is available to measure the effects of social action. (Grafen 1985a, 76)

Conceptual unification of models

Hamilton's rule has a very different explanatory function in theoretical population genetics, where it is seen not as a source of causal explanations of particular phenomena, but rather as a means of unifying diverse dynamical models under a single conceptual umbrella. This conception of the explanatory role of Hamilton's rule is forcefully advocated by Gardner et al. 2007:

> The most powerful and simple approach to evolutionary problems is to start with a method such as population genetics (including the multilocus approach), game theory or direct-fitness maximization techniques. The results of these analyses

> can then be interpreted within the frameworks that Price's theorem and Hamilton's rule provide. The correct use of these powerful theorems is to translate the results of such disparate analyses, conducted with a variety of methodologies and looking at very different problems, into the common language of social evolution theory. (Gardner et al. 2007, 224)

Gardner and colleagues' emphasis on translating results into 'the common language of social evolution theory' shows that the explanatory role being performed by Hamilton's rule is unificatory rather than causal-explanatory: the aim is not add any additional causal detail to that already included in the underlying dynamical model, but rather to show how the results of many particular models, from various different modelling traditions, can all be seen as particular instances of an overarching general principle.

4.5.2 The right rule for the right job

Recognizing the dual explanatory role that Hamilton's rule is often expected to play in contemporary sociobiology allows us to see the value in both solutions to the problem of synergy. For Solution 1 shows us how to extend Hamilton's rule so as to better enable it to perform its causal-explanatory function, and Solution 2 shows us how to reformulate it so as to better enable it to perform its unificatory function.

Let us consider the latter function first. If Hamilton's rule is to provide a common language in which to express the results of diverse modelling approaches, generality is paramount: it is important, in particular, that the rule still holds in models of synergistic interaction. The genetic formulation of Hamilton's rule, HRG, provides maximal generality. What it tells us, in effect, is that all cases of the genetical evolution of social behaviour by natural selection have something in common: in all such cases, an individual's fitness depends not only on its own internal genes, but also on the external genetic milieu in which it finds

itself. It also tells us that, as a consequence, the direction of selection depends on three factors: the association between an individual's fitness and its own genes; the association between an individual's fitness and any relevant external genes; and the correlation between its own genes and its external genetic milieu. This may not tell us very much, but it does tell us something—and it may well be all there is to say in *general* about the nature of broad-sense kin selection.

For all its unifying power, however, HRG is dreadfully ill suited to providing causal explanations of particular social phenomena. There are two main reasons for this. The first is that, since the cost and benefit terms in HRG represent the average effects of genotypes, not phenotypes, its terms are extremely difficult to measure accurately in real ecological contexts. The second is that, even if we could estimate the cost and benefit terms in real contexts, it is not clear that we would gain much in the way of causal explanation by doing so, since HRG takes no account of how the overall association between fitness and breeding value is causally mediated by phenotypic pathways. One consequence of this is that estimating the terms in HRG would not settle the question of whether a particular behaviour is selfish, spiteful, altruistic or mutually beneficial, in the standard technical sense of these terms (see Chapter 2). The information that b_g is positive and $-c_g$ is negative for a particular social behaviour, for example, would not imply that the behaviour is altruistic. It would imply only that the genes for that behaviour correlate negatively with actor fitness and positively with recipient fitness—and this pattern of correlation could in principle by explained by pleiotropic effects of the relevant genes on a quite different phenotype. The point is not that such pleiotropy is especially likely (though see Foster et al. 2004), but merely that HRG does not rule it out: because it says nothing at all about the causal pathways linking g to w, it radically underdetermines the true causal explanation of a trait's evolutionary success.

If we want to derive principles of social evolution that can do serious causal-explanatory work in real ecological contexts, then we must include phenotypes; and this is where Solution 1 earns it keep. Queller's method provides a general recipe for causal analyses of phenotypic pathways, of which HRQ is merely a simple example. The downside, of course, is that in applying this method we lose the generality HRG afforded: the traditional two-term form of Hamilton's rule (HRP) does not fully account for the *w*-*g* covariance even in simple synergy games, and predictor sets that capture all the causally relevant phenotypes (and thereby do account for all the *w*-*g* covariance) will often need to be large and complicated, and to vary significantly from one case to the next, if they are to do justice to the complexity of real, evolving populations (cf. Frank 1998; smith et al. 2010).

Let us review the argument of this section. Hamilton's rule has two roles in contemporary biology: it is a causal-explanatory principle that field biologists use to make sense of real evolutionary outcomes; but it is also employed by theorists as a unifying principle that captures a general feature of processes of social evolution. It would be convenient—but also somewhat miraculous—if a single principle could fulfil both explanatory functions. The moral of the problem of synergy is that this is not the case. To fulfil its causal-explanatory function, Hamilton's rule must be formulated using a causally adequate set of phenotypic predictors, and this leads to extended versions of the rule with as many predictors as it takes to reflect the causal structure of the problem of interest. To fulfil its unificatory function, Hamilton's rule must be formulated in terms of genetic predictors, at considerable cost to its causal-explanatory power. No single principle can do both the causal-explanatory work and the unificatory work: HRG is apt to perform the unificatory work, while the myriad Hamilton's rule-type principles derived via Queller's general method are apt to do the causal-explanatory work. Accordingly, there is no one formulation uniquely entitled to the name 'Hamilton's rule'. The bottom line is that there is no pressing need to choose between our two solutions to the problem of synergy. It is not that one solution is correct and the other incorrect; rather, each provides a means of

salvaging a version of Hamilton's rule that is able to perform one of the rule's traditional explanatory functions at the expense of the other.

FIVE

Two Conceptions of Social Fitness

> [T]here exist two forms of Hamilton's rule, each with its own distinct coefficient of relatedness. ... The similarity in the form of these coefficients often leads to the mistaken conclusion that direct and inclusive fitness models are the same process described in two different ways.
>
> (Frank 1997a, 1719)

In Chapter 4, we saw that Hamilton's rule in its traditional $rb-c>0$ form (where 'b' and 'c' represent average effects of phenotypes) struggles to accommodate the complexities of social phenomena in real ecological contexts, where synergistic effects are rife. We also saw that, although the rule still holds if reformulated in terms of the average effects of genotypes, it loses much of its causal-explanatory power in the process. If we want to use kin selection theory to give causal explanations of particular social phenomena, we are better off applying Queller's general method: construct a regression model with enough predictors to fully capture the causal pathways linking genotype to fitness, and substitute this model into the Price equation to derive a condition for positive social selection on the character under study. At the time (Section 4.3.3), we noted in passing that this general method has been developed in two quite different ways by contemporary theorists: modern kin selection theory incorporates both the neighbour-modulated (or direct) fitness approach and the inclusive fitness approach. Both are instances of Queller's general method, and both are potentially more general than (the phenotypic version of)

Hamilton's rule. The overarching aim of this chapter is to examine the relationship between these approaches.

The terminology of 'neighbour-modulated' and 'inclusive' fitness is originally due to Hamilton (1964, 5-6), who noted in passing the possibility of two alternative accounting schemes for fitness in the context of social behaviour. Hamilton himself chose to focus on developing the second, 'inclusive fitness' method, and the alternative has never received quite the same degree of popular recognition. Nevertheless, a long tradition of theory has seen the notion of 'neighbour-modulated' fitness—often under the name of 'direct' or 'personal' fitness—gradually and inconspicuously grow into a full-fledged framework for the analysis of kin selection (see especially Orlove 1975, 1979; Cavalli-Sforza and Feldman 1978; Grafen 1979; Queller 1985; Taylor and Frank 1996; Frank 1998).

Unsurprisingly, this has led to considerable discussion of when—and in what sense—the two frameworks constitute 'equivalent perspectives' on social evolution, rather than genuine rivals. Many theorists have suggested that the two frameworks are no more than alternative methods of bookkeeping which, if applied correctly, cannot disagree on any substantive questions (see, e.g., Dawkins 1982; Rousset 2004; West et al. 2007; Gardner et al. 2007; Gardner and Foster 2008; Wenseleers et al. 2010; Gardner et al. 2011; Queller 2011). Yet there have always been dissenters from the consensus view, notably including John Maynard Smith (1980, 1983, 1987), who contrasts 'the exact "neighbour-modulated fitness" approach' with 'the more intuitive "inclusive fitness" method' (1983, 315). Although Maynard Smith ultimately advocates inclusive fitness over the alternative (on the grounds that it is easier to apply), he suggests that it is unlikely to apply as widely. A similar sentiment is shared by Jeffrey A. Fletcher and Michael Doebeli (2006, 2009, 2010; see also Fletcher and Zwick 2006; Fletcher et al. 2006), who controversially argue that only a neighbour-modulated/direct fitness approach has the resources to cover all cases of the evolution of altruism. Other authors defend positions that are hard to place squarely in either camp. Peter D. Taylor and colleagues (2007), for example, argue that if certain

conditions obtain (namely, fair meiosis, weak selection and interactions among conspecifics only), then the two frameworks will yield equivalent predictions about the direction of evolutionary change. They allow, however, that when their assumptions are relaxed, the frameworks might well come apart. Steven A. Frank (1997a,b, 1998) is a second example: though regularly cited as a defender of equivalence, he writes (in the passage quoted at the start of this section) of the 'mistaken conclusion that direct and inclusive fitness models are the same process described in different ways' (1997a, 1719).

What is needed is a precise and general statement of the conditions under which the frameworks are equivalent, and of the conditions under which they are not. My goal in this chapter is to advance the debate by providing such a statement. The overview of the chapter is as follows. In Section 5.1, I further motivate the project by introducing two quite different ways of thinking informally about the role of relatedness in social evolution. I suggest that, in asking whether neighbour-modulated and inclusive fitness are 'formally equivalent', we are essentially asking whether these 'ways of thinking' describe different mechanisms for the evolution of altruism, or whether they are merely alternative perspectives on the same mechanism. In Section 5.2, I set out more precisely the conceptual contrast between neighbour-modulated and inclusive fitness, and highlight a number of subtle features of inclusive fitness that are often overlooked. I then introduce Steven A. Frank's (1997a,b, 1998) influential formalism for neighbour-modulated and inclusive fitness, which provides the most appropriate framework within which to address questions of their formal equivalence. In Sections 5.3 and 5.4, I discuss in detail when the frameworks are equivalent, and when they are not. I argue, in short, that, while the frameworks can be shown to be equivalent across a wide range of cases, there are some important classes of case in which they are non-equivalent.

5.1 Why does relatedness matter? Two kinds of answer

It is a platitude of social evolution theory that relatedness between social partners can lead to the evolution of social behaviour that would not be stable in its absence. The paradigm cases are cases of altruism, in which an individual negatively impacts its own fitness while conferring a fitness benefit on another individual. We often say that relatedness between social partners makes altruism possible. But *why* does relatedness matter so much? In contemporary social evolution theory, one finds not one but two kinds of answer to this question; and it is certainly not obvious that they are equivalent.

5.1.1 The 'indirect reproduction' answer

There is a longstanding affinity between kin selection theory and a 'gene-centred' or 'gene's eye' view of evolution (see, e.g., Dawkins 1976, 1979, 1982; Bourke 2011a,b). At first glance, one might wonder whether this is simply an historical accident: a result of the fact that W. D. Hamilton, in addition to being the first theorist to formulate the core idea of kin selection, was also a proponent of the gene's eye view (and therefore formulated his theory in genetic terms). After all, kin selection theory is fundamentally concerned with how interactions and patterns of resemblance between *organisms* influence the outcome of selection, and it is possible to formulate versions of the theory that do not assume particulate inheritance (Gardner 2011). I suspect, however, that the widespread association of these ideas, theoretically separable though they may be, is no accident, for taking a 'gene's eye' perspective on kin selection makes the basic evolutionary logic behind the theory seem incredibly intuitive.

The familiar 'gene-centred' rationale for kin selection goes like this: organisms are designed by natural selection to do whatever they can to maximize their genetic representation in the next generation. Broadly speaking, there are two ways for an organism to do this. The *direct* way for an organism to increase its genetic representation is

for it to have more offspring of its own, since it will be able to transmit copies of its genes to the next generation via these offspring. But an organism can also increase its genetic representation *indirectly*, by helping *other* organisms have more offspring. This strategy will only work, however, if the organism differentially confers benefits on recipients who are *more likely than average* to transmit copies of its genes to their offspring. We can think of these recipients as the organism's 'relatives', although there is no requirement that they are its genealogical kin: what matters is that, for whatever reason, they are disposed to transmit copies of its genes. We can intuitively see how altruistic behaviours might evolve by this route: an organism may sacrifice some of its direct reproduction to help relatives, but it will be a sacrifice worth making if the increased genetic representation it gains through the indirect pathway outweighs the representation it loses through the direct pathway. Here, then, is our first answer to the question of why relatedness matters:

> **Answer 1:** Positive relatedness promotes altruism because it provides altruists with an indirect means of transmitting their genes to the next generation.

When an organism gains genetic representation in the next generation through helping relatives, the process is often glossed as indirect reproduction. Consider, for example, the following quotations:[1]

> Social insects are characterized by indirect reproduction, in which most individuals achieve genetic success by helping to rear the offspring of colony mates. (Strassmann et al. 1989, 268)

[1] See also, e.g., Queller 1989, 1996; Gadagkar and Bonner 1994; Cronk 1991, 2007; Choe and Crespi 1997; Voland 1998; Queller et al. 2000; Oli 2003; Frank 1998, 2006; Ratnieks et al. 2006; Ratnieks and Wenseleers 2008.

> [N]onreproductive workers ... can compensate for their loss of direct reproduction by the indirect reproduction achieved through helping relatives, provided relatedness is high enough. (Hastings et al. 1998, 573)

For this reason, I will call this line of thought the 'indirect reproduction' explanation for the importance of relatedness. The idea, in a nutshell, is that high relatedness matters to the evolution of altruism because it allows social actors to achieve a form of 'indirect reproduction' through altruistic acts.

5.1.2 The 'positive assortment' answer

Talk of 'indirect reproduction' remains widespread in elementary expositions of kin selection, particularly those directed at students and field biologists. It is, however, somewhat out of fashion in theoretical circles. The concern is that, in emphasizing *genetic* resemblance between social actors and the offspring of their recipients, the 'indirect reproduction' story unnecessarily limits the domain of phenomena to which kin selection theory can apply. For many of the current generation of social evolution theorists, kin selection theory can and should be extended to accommodate and explain any process in which salient resemblance between individuals leads to the evolution of social behaviour, be it via 'narrow-sense' kin selection based on genealogical kinship, direct or indirect reciprocity, group selection, greenbeard effects, or any other selection process. Since many of these processes do not involve social actors securing genetic representation in the next generation by an indirect pathway (so the line of thought goes), we need to replace the traditional explanation for the importance of relatedness in social evolution with one that applies more widely.

This theoretically in-vogue story takes the defining feature of a process of kin selection to be assortment between recipient genotypes and actor phenotypes (see, e.g., Kerr and

Godfrey-Smith 2002; Fletcher and Zwick 2006; Fletcher and Doebeli 2006, 2009, 2010; Godfrey-Smith 2009a; Rosas 2010). In the context of the evolution of altruism, what matters is that bearers of altruistic genotypes differentially receive the benefits of the altruism. As Fletcher and Doebeli (2009) put it:

> [W]hat is necessary for the evolution of altruism is assortment between focal genotype and phenotypic help, rather than the assortment among genetic types often emphasized in kin selection theory. (Fletcher and Doebeli 2009, 17)

Informally, the 'new' story goes like this: when some altruistic behaviour evolves 'by kin selection', it evolves because individuals with altruistic genotypes are fitter on average than non-altruists; and they are fitter on average because there is a statistical tendency for the benefits of altruism to fall differentially on bearers of altruistic genotypes. We can still talk of 'relatedness' in this framework, but the relatedness that matters is correlation between one's own genotype and the phenotypes of the social actors with whom one interacts; *genetic* correlation between actors and recipients is strictly optional.

Here, then, is our second answer to the question of why relatedness matters:

> **Answer 2:** Positive relatedness promotes altruism because it implies that the benefits of altruistic acts fall differentially on bearers of the genes for altruism.

I will call this the 'positive assortment' answer to the question of why relatedness matters. The attraction over the 'indirect reproduction' answer, at least in the eyes of its proponents, lies in its ability to extend to cases in which there is no strong genetic similarity between actors and recipients.

5.1.3 The equivalence question

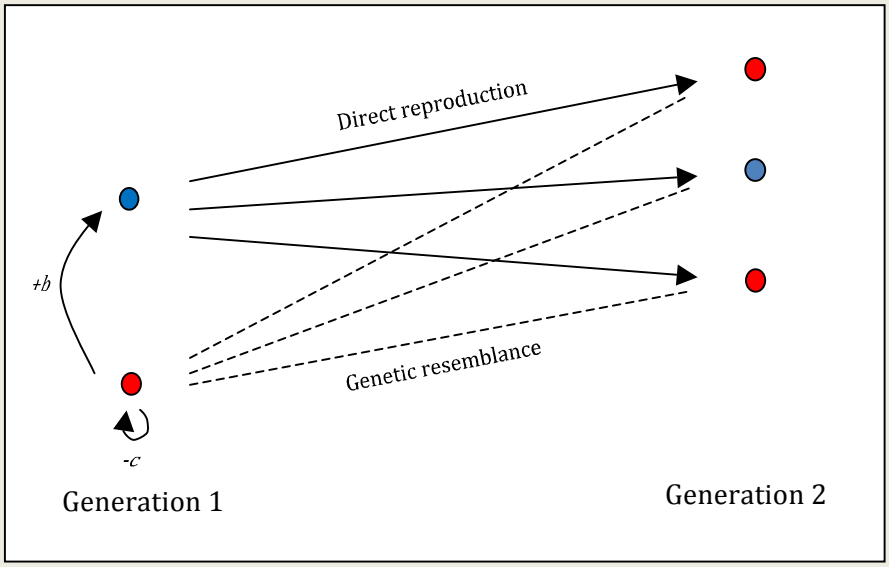
These two answers provide us with pictures of how altruism evolves that are, on the face of it, quite different from each other (Box 4.1). This naturally leads to the question of when, if at all, these two pictures constitute equivalent representations of the same evolutionary process, rather than representations of qualitatively different processes.

Our intuitions on this question could, I think, go either way. On the one hand, we might intuitively take the two pictures to represent qualitatively different mechanisms for the evolution of altruism, since there is a qualitative difference in the sort of 'relatedness' they take to be important. For note that, on the 'indirect reproduction' answer, what really matters is correlation between *the genotype of the actor and the genotype of the recipient's offspring*; whereas, on the 'positive assortment' picture, what matters is correlation between *the phenotype of the actor and the genotype of the recipient*. Though the difference is subtle, neither kind of correlation strictly implies the other (a point I revisit in Section 6.4). On the other hand, we might intuitively suspect that, although the two pictures embody subtly different conceptions of the sort of 'relatedness' that matters for the evolution of altruism, in reality both kinds of relatedness tend to arise from the same family of causal mechanisms: limited dispersal, kin recognition, greenbeard effects, and so on. If the causal mechanisms responsible for both kinds of correlation are the same in many cases, then it would seem more reasonable to regard our two informal pictures (at least in these cases) not as characterizations of different processes, but as two alternative ways of visualizing the same process.

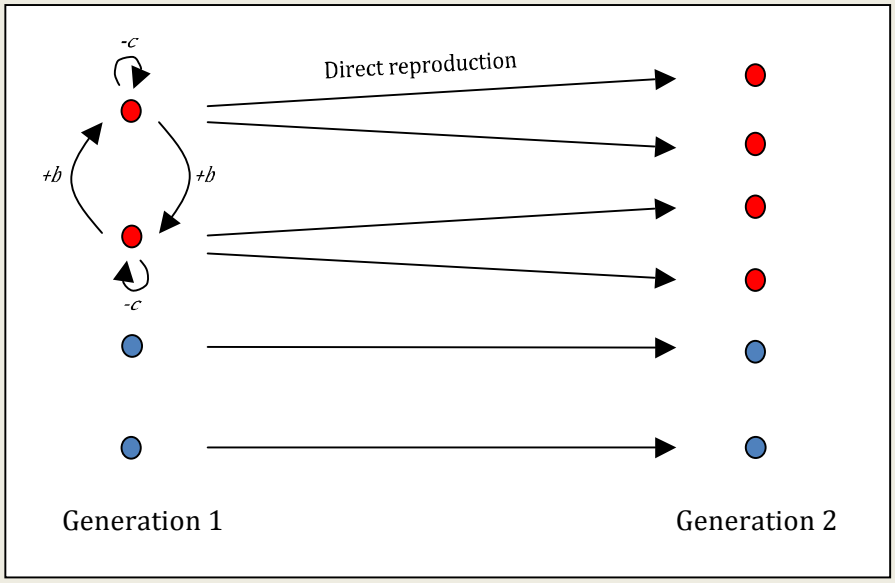
I therefore doubt that we can settle the equivalence question by intuition alone: to arrive at a more satisfactory answer, we must approach the question formally. This is the strategy I intend to pursue in the coming sections. In the next section, I introduce the neighbour-modulated and inclusive fitness approaches to the analysis of kin selection, and I argue that the distinction between these approaches reflects the distinction between our two

informal pictures. As a consequence, the question of when, if at all, 'indirect reproduction' and 'positive assortment' constitute equivalent perspectives on social evolution may be recast as the question of when, if at all, the neighbour-modulated and inclusive fitness approaches are formally equivalent.

Box 5.1: *Two ways for altruism to pay*



Picture 1: Altruism pays due to indirect reproduction. Altruists (red) differentially confer fitness benefits on recipients who are disposed to transmit the genes for altruism. Recipients thereby provide actors with an indirect route to genetic representation in the next generation. Actor phenotypes may also correlate with recipient genotypes, but this is not assumed.



Picture 2: Altruism pays due to positive assortment. Altruists (red) differentially receive the benefits of altruism. They are therefore fitter, on average, than individuals who do not possess the altruistic genotype (blue). The recipient's offspring may also bear a genetic resemblance to the actor, but this is not assumed.

5.2 Neighbour-modulated and inclusive fitness

5.2.1 The conceptual contrast

The Price formalism describes the evolutionary change between two sets of entities connected by a mapping relation *R* (cf. Kerr and Godfrey-Smith 2009; Frank 2012). In Chapter 3 we noted that, while in biological contexts the salient mapping relation between the two sets is usually direct lineal descent (i.e., parenthood, if descendants and ancestors are separated by a single generation), we may in principle assign descendants to ancestors in alternative ways. The fundamental difference between neighbour-modulated and inclusive fitness is that, while the former prioritizes considerations of parenthood in the assignment of descendants to ancestors, the latter prioritizes considerations of social causation and control (cf. Frank 1998). This can lead to radically divergent measures of an individual's social fitness.

An individual's neighbour-modulated fitness (Figure 5.1) is a measure of its personal reproductive success: typically, it is the expected or realized number of offspring of which it is a parent. The qualifier 'neighbour-modulated' is merely an acknowledgement that in cases of social behaviour, an individual's personal reproductive success is influenced—often heavily influenced—by the properties of the individuals with which it interacts. Note that, although neighbour-modulated fitness can be used to analyse kin selection, the concept of relatedness does not feature in the *definition* of neighbour-modulated fitness. To evaluate the neighbour-modulated fitness of a particular individual, we need only look at the offspring it personally produces: we need not have any prior information about its relatedness to its social partners.

An individual's inclusive fitness (Figure 5.2) is a weighted sum of the causal contributions its behaviour makes to the reproductive success of individuals, including (but not limited

to) its causal contribution to its *own* reproductive success. Each contribution is weighted by a measure of its value to the actor as a route to genetic representation in the descendant-population, and the correct weights are measures of genetic correlation between the actor and the recipient's descendants (Frank 1997a,b; 1998). Note, therefore, that relatedness *does* feature explicitly in the definition of inclusive fitness, and that the relatedness that matters is strictly genetic. In W. D. Hamilton's original words:

> Inclusive fitness may be imagined as the personal fitness which an individual actually expresses in its production of adult offspring as it becomes after it has been stripped and augmented in a certain way. It is stripped of all components which can be considered as due to the individual's social environment, leaving the fitness he would express if not exposed to any of the harms or benefits of that environment. This quantity is then augmented by certain fractions of the quantities of harm and benefit which the individual himself causes to the fitness of his neighbours. The fractions in question are simply the coefficients of relationship. (Hamilton 1964, 8)

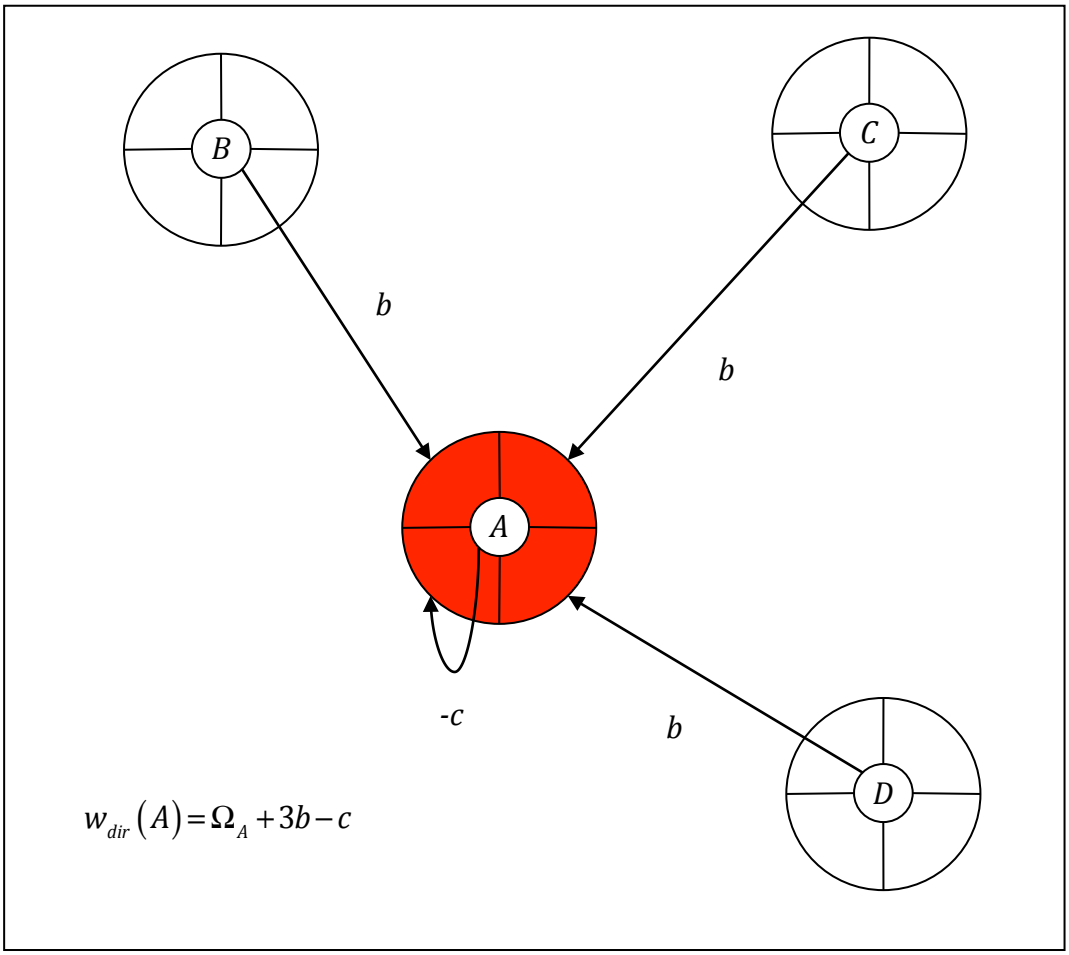


Figure 5.1: Neighbour-modulated fitness. In a neighbour-modulated fitness analysis, we ascribe to *A* all and only those fitness components that correspond to its personal reproductive success. Some of these components are influenced by the behaviour of *B*, *C* and *D* (as indicated by the arrows). *A*'s total neighbour modulated fitness is a straightforward, unweighted sum of these components ($3b$), plus a component corresponding to *A*'s *own* influence on its reproductive success via the character under study ($-c$), plus a *baseline* component (Ω_A) independent of that character.

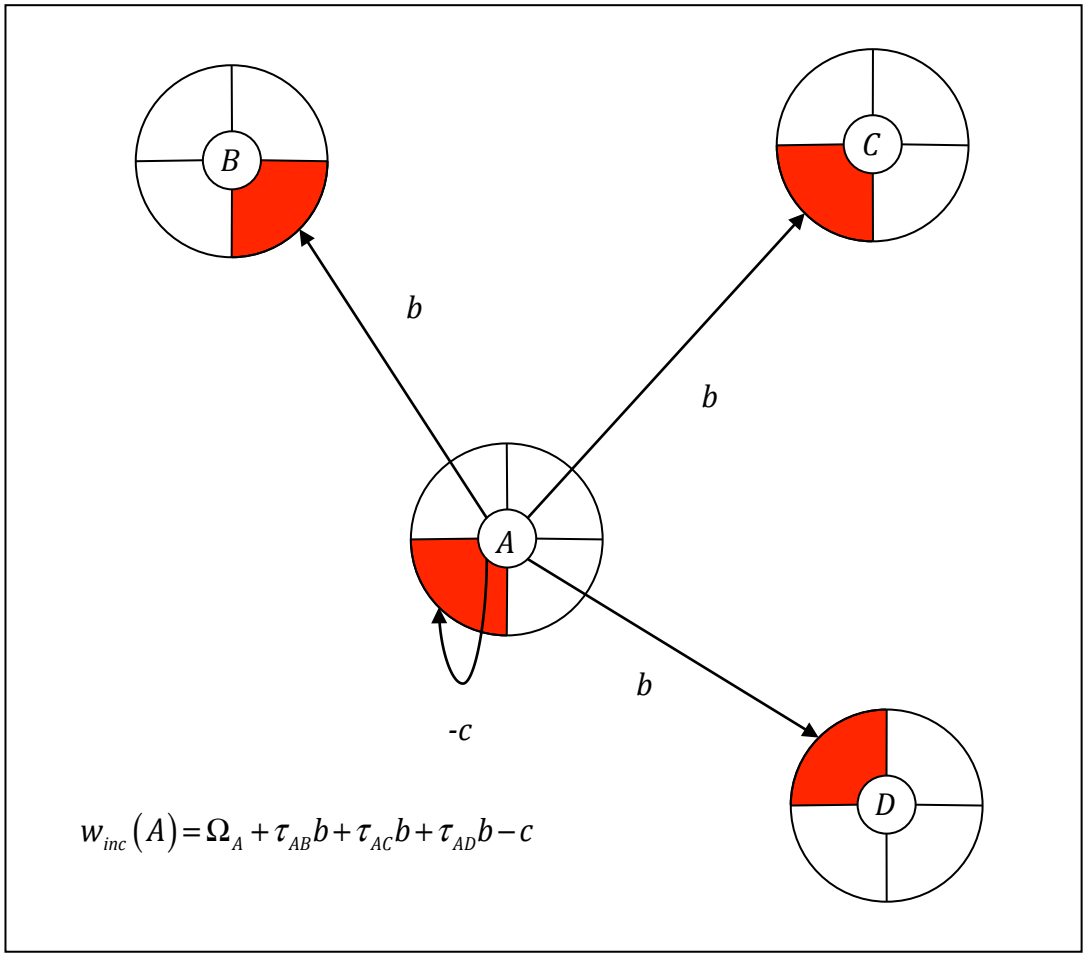


Figure 5.2: Inclusive fitness. In an inclusive fitness analysis, fitness effects are assigned to the actors whose behaviour was responsible for them. *A* therefore loses the slices of personal fitness it obtained through its interactions with *B*, *C*, and *D*. In compensation, it gains three new slices, taken from the personal fitness of *B*, *C* and *D*, which causally depend on its own behaviour. To calculate *A*'s inclusive fitness, these new slices must be weighted by a suitable measure of their value to *A* as routes to genetic representation in the next generation (τ), and this will be a measure of genetic relatedness. In short, therefore, *A*'s inclusive fitness is a relatedness-weighted sum of the fitness effects for which it is causally responsible.

5.2.2 Five subtleties of inclusive fitness

The notion of inclusive fitness is conceptually more challenging than that of neighbour-modulated fitness, for at least five reasons. Several of these arise from the counterintuitive consequences of the regression definition of relatedness (see Section 3.1.4), while others arise from the sensitivity of inclusive fitness to considerations of causal responsibility.

Inclusive fitness is an inherently causal notion

An individual's neighbour-modulated fitness functionally depends on its interactions with social partners. But the relationship between inclusive fitness and causation is more intimate than this, for the very notion of inclusive fitness is defined in explicitly causal terms. To calculate an individual's inclusive fitness, we need information about the reproductive success of all and only those individuals *with which it has causally interacted*. A failure to appreciate the causal nature of inclusive fitness leads to the widely held but inaccurate view that an individual's inclusive fitness depends on the fitness of all individuals to whom it is related, whether or not it has ever interacted with them. In fact, having a very successful relative will not increase one's inclusive fitness unless one is causally responsible for a portion of that relative's success (Grafen 1979, 1982, 1984; Dawkins 1982).

Inclusive fitness is character-relative

In an inclusive fitness analysis, fitness components are assigned to an actor on the basis of how it has influenced its own reproductive success and that of others *through expressing the character (or set of characters) under study*. This introduces a form of character-relativity, since an individual may have very high inclusive fitness with respect to one character (or set of characters), and yet have very low inclusive fitness with respect to another. One might suppose that we could define an organism's *overall* inclusive fitness as the sum of its inclusive fitness with respect to each of its characters, but this would quickly lead to problems of double counting. For instance, the founding of a new insect colony has many

downstream effects, including the production of new workers and all the work they do. If we are attempting to estimate the inclusive fitness effect on a foundress of founding a new nest (rather than staying in her mother's nest), we will need to take these effects into account and attribute them to the (distally responsible) foundress. But if we are attempting to estimate the inclusive effect on a worker of participating in an item of work, we will need to take some of these very same effects into account—but this time we will need to attribute them to the (proximally responsible) worker. Because causal responsibility is inescapably character-relative, so is inclusive fitness.

Inclusive fitness is independent of considerations of parenthood

Inclusive fitness replaces considerations of parenthood with considerations of causal responsibility, and this has occasionally counterintuitive consequences. On the one hand, an individual with no personal offspring at all can still have substantial inclusive fitness with respect to a particular character, if it confers fitness benefits on many related individuals. This is one reason to suspect that inclusive fitness is extremely important to the evolution of sterile workers in insect societies. These individuals typically have zero neighbour-modulated fitness, but, since they devote their lives to assisting the queen, their inclusive fitness is likely to be much higher. On the other hand, and less intuitively, an individual with high personal reproductive success could still have low or even zero inclusive fitness with respect to some social behaviour, should it be the case that (i) its personal fitness is heavily influenced by instances of the behaviour in other individuals, and yet (ii) it does not confer any fitness benefits on related individuals by expressing the behaviour itself. An inclusive fitness analysis will strip this individual of the fitness components it owes to the behaviour of others; and, since it does not confer any fitness benefits on others through its own behaviour, it will not gain any new components in return. The result is that its inclusive fitness will be much *lower* than its neighbour-modulated fitness.

Spiteful behaviours can increase inclusive fitness

In Section 4.1.4, we noted that, on the regression definition, relatedness can be negative, and that this allows for the evolution of spiteful behaviours which detract from the fitness of both actor and recipient. A further implication is that, if relatedness is sufficiently negative, and if the harm caused to the recipient is sufficiently greater than the harm incurred by the actor, then performing a spiteful behaviour can increase an organism's inclusive fitness. Even more counterintuitively, performing a mutually beneficial behaviour can *detract* from an actor's inclusive fitness, if relatedness is sufficiently negative and the benefit conferred on the recipient is sufficiently greater than the benefit obtained by the actor. This reflects the fact that inclusive fitness is ultimately a measure of how successful an actor has been in securing genetic representation in the next generation. Sometimes helping the wrong recipients is counterproductive in this regard, and the result is that the sign of the inclusive fitness effect differs from the sign of the actual fecundity payoff.

Inclusive fitness, like relatedness, is population-relative

In Section 4.1.4, we noted that relatedness is, strictly speaking, a property of a population, and that the choice of an appropriate reference population has significant consequences for the link between relatedness and altruism. Inclusive fitness, by contrast, is a property of a particular organism, and depends on the fitness effects for which the organism is personally responsible. Yet because these effects are weighted by coefficients of relatedness, and relatedness is population-relative, inclusive fitness is also population-relative: it is a property of a particular organism *relative to a reference population*. The upshot is that the choice of reference population may affect whether or not a given behaviour contributes positively or negatively to inclusive fitness.

Suppose (as in Section 4.1.4) that we are studying a viscous population in which relatedness between social partners is high relative to the global population mean, but low relative to the local subpopulation mean: kin cluster together on the whole, but organisms

do not differentially interact with the *closest* kin in their vicinity. In such a scenario, it may well be that altruistic behaviours contribute to an organism's inclusive fitness relative to the global population, yet detract from an organism's inclusive fitness relative to the local subpopulation. The implication is that, if we want to use inclusive fitness considerations to draw inferences about the kinds of social behaviour that are likely to evolve, we need to know where the competition is. If competition is mostly local (i.e., *within* subpopulations), then inclusive fitness should be evaluated relative to the local subpopulation; if competition is mostly global (i.e., *between* subpopulations), then inclusive fitness should be evaluated relative to the global population.

Creel's paradox

The importance of taking these various subtleties into consideration is aptly illustrated by 'Creel's paradox' (Creel 1990; Queller 1996). Scott R. Creel (1990) argues that, given how inclusive fitness is usually defined, the queen of a social insect colony seems to have virtually zero inclusive fitness. After all, her reproductive success owes little to her own behaviour, and she does nothing at all to aid the reproductive success of other individuals. She spends her life laying millions of eggs, all the while receiving a steady stream of fitness benefits from the minions who supply her with food and shelter, and who defend the colony with their lives. As a result, her personal (neighbour-modulated) fitness is undoubtedly extremely high, but her inclusive fitness, strictly speaking, must be negligible. The strange implication appears to be that, in a social insect colony, the workers have greater inclusive fitness than the queen! Creel considers this result sufficiently implausible to warrant a significant revision to the theory of inclusive fitness, for there is plenty of empirical evidence of workers fighting amongst themselves to replace or displace a queen (see, e.g, Queller and Strassmann 1998), but no evidence of queens and their daughters fighting amongst themselves for the right to be workers.

Queller (1996) shows, however, that the appearance of paradox (or at least, of a seriously counterintuitive result) is dispelled when we consider the causal and character-relative

nature of inclusive fitness. It is true enough that the queen has low inclusive fitness with respect to most of the behaviours routinely expressed by workers: foraging, nest construction, nest defence, and so on. She contributes little to these tasks, and gains a great deal from their completion. But the queen has very *high* inclusive fitness with respect to a different behaviour: the behaviour of founding a new colony and adopting the queen role. For in installing herself as queen, rather than choosing to work for another, the queen makes a vast causal contribution to her own reproductive success.

5.2.3 Frank's formalism for neighbour-modulated fitness

While the conceptual distinction between neighbour-modulated and inclusive fitness is easy enough to grasp, we can only understand their subtle relationship to each other by introducing a formal framework within which both conceptions of social fitness may be captured and compared. The formalism I introduce in this section is based on that of Frank (1997a,b, 1998), with a few simplifications made for ease of exposition.[2] Frank's formalism falls under the broader umbrella of the Price formalism (Chapter 3), and can also be regarded as an application of Queller's general method for the regression analysis of social evolution (Chapter 4).[3]

---

[2] In particular, Frank sorts organisms into both genotypic classes *and* developmental classes, whereas I only employ developmental classes. Sorting by genotype requires the introduction of class reproductive value weightings, since the fitness of a genotype in one developmental class will not in general equal the fitness of the same genotype in a different class. Because I only assign fitness values to individuals, not genotypes, I can avoid introducing class reproductive value weightings.

[3] Frank's formalism for kin selection theory has been influential, but it is not the only option. For a somewhat different way of formulating neighbour-modulated and inclusive fitness as partitions of the Price equation, see Grafen 2006a. It would be interesting to see whether results derived within Frank's formalism could be recovered within Grafen's framework; I suspect that any differences would turn out to be superficial.

The starting point

We start with the Price equation for the primary and secondary effects of natural selection, partitioned into components corresponding to distinct developmental classes (cf. Section 3.4.2, equation 3.4.6):[4]

$$\Delta_w \bar{g} = \mathrm{E}_m\left[\mathrm{Cov}^i\left(w, g'\right)\right] = \frac{1}{\bar{w}} \sum_i q_i \mathrm{Cov}^i\left(w, g'\right) \tag{5.2.1}$$

Following the notation introduced in Section 2.4, q_i represents the relative size of the i^{th} class. As a starting point for analysis, two features of equation (5.2.1) are worthy of comment. First, the equation explicitly accommodates developmental classes. This may involve grouping organisms by age, sex, morphological caste, or any combination of these. One might wonder whether this degree of complexity is needed, if the aim is simply to compare the neighbour-modulated and inclusive fitness frameworks rather than to apply them to particular biological problems. It *is* needed, however, because the interesting differences between the two frameworks mostly disappear if one considers an homogeneous population with no class structure. To see where the differences lie, it is essential to consider a population partitioned into classes (cf. Gardner et al. 2007; Wenseleers et al. 2010; Gardner et al. 2011; Queller 2011). Second, the equation considers the *secondary* effect of natural selection as well at the primary: we take $\mathrm{Cov}\left(w, g'\right)$, not $\mathrm{Cov}\left(w, g\right)$, as the target of analysis. This is an idiosyncratic feature of Frank's formalism that is rarely replicated elsewhere (though see Okasha 2006). It is optional when formulating the neighbour-modulated fitness approach, for this approach is mostly concerned with analysing assortment among actors and recipients in the ancestor-population, and this could equally be achieved by starting with $\mathrm{Cov}\left(w, g\right)$. It is necessary, however, if we want to formalize the *inclusive* fitness approach in a way that does justice to the intuitive idea of recipients providing actors with an indirect route to genetic

[4] As previously noted in Section 2.4, this equation assumes that there is no genetic variance between classes—otherwise a further term is needed to capture the between-class covariance (Frank 1997b, 1998).

representation in the next generation (cf. Box 4.1). The natural way to capture this idea is to weight fitness components by the correlation between the genotypes of actors and the genotypes of their recipients' *descendants*; and this requires that we take account of the recipients' values for g' as well as their values for g (cf. Frank 1997a).

Regression equations

Next, we introduce three separate regression analyses:

> ***Regression 1*:** For the i^{th} class, a regression model of all causally relevant phenotypic influences on fitness, including both intrinsic and extrinsic (i.e., 'neighbourhood') characters:
>
> $$w_{ik} = \sum_j \beta_{ij} z_{ijk} + \varepsilon$$
>
> Here, w_{ik} denotes the fitness of the k^{th} individual in the class, β_{ij} denotes the average effect of the j^{th} relevant phenotype, and z_{ijk} denotes the value of the j^{th} phenotype for the k^{th} individual. Note that this equation only analyses the relationship between phenotype and fitness *within* a particular class; each class thus requires a separate analysis. We thereby allow that different classes may be influenced by different phenotypes, and that the average fitness effect of a given phenotype may vary depending on the affected individual's class.
>
> ***Regression 2*:** For each relevant phenotype, a regression analysis of its statistical association with the g-value of the *affected individual*:
>
> $$z_{ijk} = \rho_{ij} g_{ik} + \varepsilon$$

Here, z_{ijk} denotes (as above) the value of the *j*th phenotype for the *k*th member of the *i*th class, g_{ik} denotes the *g*-value of that individual with respect to the character under investigation, and ρ_{ik} represents the simple regression of z_{ijk} on g_{ik}.

***Regression 3*:** A regression analysis of the fidelity of direct transmission from ancestors to their direct lineal descendants, averaging over all classes:

$$g' = \tau_0 g + \varepsilon$$

Here, τ_0 denotes the simple regression of descendant genotype on ancestor genotype, and this may be regarded as an overall measure of the fidelity of direct, vertical transmission.

Substitution

We now substitute all three ingredients into equation (5.2.1), making the (substantive) assumption that the residuals in the three different regression models are uncorrelated with each other. This yields:

$$\Delta_w \bar{g} = \frac{1}{\bar{w}} \left[\tau_0 \sum_i q_i \sum_j \rho_{ij} b_{ij} \, \mathrm{Var}^i \left(g_{ik} \right) \right]$$

Conditional on the (again, substantive) assumption that the within-class genetic variance, $\mathrm{Var}^i \left(g \right)$, is the same for all classes, we can exploit the fact that variance and $\bar{w}$ cannot be negative to obtain the following rule concerning the direction of partial change (Frank 1998):

$$\mathrm{sign}\left(\Delta_w \bar{g} \right) = \mathrm{sign}\left(\tau_0 \sum_i q_i \sum_j \rho_{ij} b_{ij} \right) \qquad (5.2.2)$$

We can think of the term on the right hand side of equation (5.2.2) as a measure of the overall extent to which transmitted differences in the character under study predict differences in neighbour-modulated fitness. We can call this the neighbour-modulated fitness increment for the character under study. The rule tells us that the (primary and secondary) effect of natural selection will be to drive the evolution of the character in the same direction as the neighbour-modulated fitness increment.

In the special case in which the evolution of a particular social behaviour is influenced *only* by the direct cost it imposes on the actor ($\beta_{ij} = -c, \rho_{ij} = 1$) and by the benefit actors tend to receive from social partners with the same trait ($\beta_{ij} = +b, \rho_{ij} = \rho_0$), and in which class structure is wholly absent, the general rule reduces to the following, much simpler rule:

$$\text{sign}\left(\Delta_w \bar{g}\right) = \text{sign}\left[\tau_0 \left(\rho_0 b - c\right)\right]$$

Conditional on the further assumption that the actor transmits to its own direct lineal descendants with perfect fidelity ($\tau_0 = 1$), we recover a two-term rule that bears a strong resemblance to Hamilton's rule in its traditional form (Frank 1997a):

$$\text{sign}\left(\Delta_w \bar{g}\right) = \text{sign}\left(\rho_0 b - c\right)$$

5.2.4 Frank's formalism for inclusive fitness

To formalize inclusive fitness, we start, as before, with equation (5.2.1):

$$\Delta_w \bar{g} = \frac{1}{\bar{w}} \sum_i q_i \text{Cov}^i \left(w, g'\right)$$

Regression equations

An inclusive fitness approach, much like a neighbour-modulated fitness approach, involves partitioning the within-class covariance through regression analysis. Indeed, the first regression is exactly the same: we write an individual's personal fitness as a weighted sum of correlated phenotypes (whether intrinsic or extrinsic), weighted by partial regression coefficients. But the second and third regression equations are different. Instead of relating the correlated phenotypes to the genotype of the recipient, we relate these phenotypes to the genotype of *the actor who controls the phenotype*. And instead of relating the genotypes of ancestors to those of their direct lineal descendants, we relate the genotypes of actors to those of their *recipients'* descendants, as if the recipient had provided the actor with an indirect channel of transmission.

***Regression 1*:** For the i^{th} class, a regression model of all causally relevant phenotypic influences on fitness, including both intrinsic and extrinsic (i.e., 'neighbourhood') characters:

$$w_{ik} = \sum_j \beta_{ij} z_{ijk} + \varepsilon$$

Here, z_{ijk} again denotes the value of the j^{th} phenotype for the k^{th} member of the k^{th} class, and β_{ij} again denotes the extent to which the j^{th} phenotype predicts recipient fitness, correcting for other relevant phenotypes.

***Regression 2*:** For each relevant phenotype, a regression analysis of its statistical association with the breeding value of the *controlling actor*:

$$z_{ijk} = d_{ij} g_{ijk} + \varepsilon$$

Here, z_{ijk} denotes (as above) the value of the j^{th} phenotype for the k^{th} member of the i^{th} class; g_{ijk} denotes the breeding value of the individual who controls

the character; and d_{ij} represents the simple regression of z_{ijk} on g_{ijk}. We can think of a *d*-coefficient as a measure of the extent to which the *j*th social phenotype is predicted by the genotype of the actor who controls it. I will refer to these *d*-coefficients as 'coefficients of control'.

***Regression 3*:** For each relevant phenotype, a regression of the breeding value of the *controlling actor* on the average breeding value of the descendants of the affected individual:

$$g_{ijk} = \tilde{\tau}_{ij} g'_{ik} + \varepsilon$$

Substitution

We now substitute all three ingredients into equation (5.2.1), again making the (substantive) assumption that residuals in the three regression models do not correlate with each other:

$$\Delta_w \bar{g} = \frac{1}{\bar{w}} \left[\sum_i q_i \sum_j d_{ij} \beta_{ij} \tilde{\tau}_{ij} \operatorname{Var}^i \left(g'_{ik} \right) \right]$$

As Frank (1997b, 1998) notes, this result does not yet admit of an intelligible interpretation in terms of inclusive fitness. We can obtain a result that *does* admit of such an interpretation by 'flipping' the direction of Regression 3, so that we regress descendant breeding values on controlling actor breeding values rather than the other way round. We can do this by noting that, for all *ij*, $\tilde{\tau}_{ij} \operatorname{Var}\left(g'_{ik}\right) = \tau_{ij} \operatorname{Var}\left(g_{ijk}\right)$, where τ_{ij} is the simple regression of the breeding values of the descendants of the *i*th recipient class on the breeding values of the actors who control the *j*th phenotype. This yields:

$$\Delta_w \bar{g} = \frac{1}{\bar{w}} \left[\sum_i q_i \sum_j d_{ij} \beta_{ij} \tau_{ij} \operatorname{Var}^i \left(g_{ijk} \right) \right]$$

Conditional on the further substantive assumption that the genetic variance among controlling actors (i.e, $\mathrm{Var}\left(g_{ijk}\right)$) is the same for every actor-class, we can exploit the fact that variance and $\bar{w}$ cannot be negative to obtain the following rule concerning the direction of partial change (Frank 1998):

$$\mathrm{sign}\left(\Delta_w \bar{g}\right) = \mathrm{sign}\left(\sum_i q_i \sum_j d_{ij} \beta_{ij} \tau_{ij} \right) \tag{5.2.3}$$

We can think of the term on the right hand side of equation (5.2.3) as a measure of the overall extent to which 'transmitted' differences in the genotypes of controlling actors predict differences in the fitness effects for which those actors are responsible, where 'transmitted' means that the genes reappear not in the direct lineal descendants of the actors, but rather in the descendants of the recipients it affects. We can interpret this as a measure of the overall extent to which an actor's genotype is associated with its *inclusive* fitness, where the inclusive fitness of a controlling actor is understood as the sum of fitness components for which its behaviour is responsible, weighted by τ-coefficients representing the 'transmission fidelity' of the actor's genes through each component. We can call this quantity the inclusive fitness increment for the character under study. The rule tells us that the (primary and secondary) effect of natural selection will be to drive social evolution in the same direction as the inclusive fitness increment.

In the special case in which the evolution of a particular social behaviour is influenced *only* by its direct effect on the actor ($d_{ij} = 1, \beta_{ij} = -c, \tau_{ij} = \tau_1$) and by its direct effect on a single recipient ($d_{ij} = 1, \beta_{ij} = +b, \tau_{ij} = \tau_2$), and in which class structure is wholly absent, the general rule reduces to the following, much simpler rule (Frank 1997a):

$$\mathrm{sign}\left(\Delta_w \bar{g}\right) = \mathrm{sign}\left(\tau_2 b - \tau_1 c\right)$$

Conditional on the further assumption that the actor transmits to its own direct offspring with perfect fidelity ($\tau_1 = 1$), we once again obtain a rule that bears a strong resemblance to Hamilton's rule in its most familiar form:

$$\text{sign}\left(\Delta_w \bar{g}\right) = \text{sign}\left(\tau_2 b - c\right)$$

5.2.5 *The two pictures revisited*

There is a close relationship between the two formal representations of kin selection outlined above and the two informal explanations for the evolution of altruism discussed in Section 5.1: neighbour-modulated fitness is the natural framework for analysing whether altruism pays due to positive assortment (i.e., Picture 1), while inclusive fitness is the natural framework for analysing whether altruism pays due to indirect reproduction (i.e., Picture 2). Because of this, the formal equivalence (or otherwise) of the two theoretical representations would reveal something of wider significance about the equivalence (or otherwise) of our informal explanations. I will briefly elaborate on these points, because they will be important later on.

Neighbour-modulated fitness analyses positive assortment

Each of the ρ-coefficients in the neighbour-modulated fitness framework can be interpreted as a measure of the 'relatedness' between recipients of the i^{th} class and the j^{th} influence on their fitness, in the sense of relatedness introduced in Chapter 4. Note, however, that 'relatedness' in this sense is not purely genetic. What these coefficients measure is the degree of association between an individual's breeding value and its phenotypic characters, where these are considered to include extrinsic characters that represent aspects of its social milieu. If possessing the genes for altruism makes a member of a particular class more likely to be surrounded by agents with a particular altruistic phenotype, ρ_{ij} will be positive for the relevant *i* and *j*. If the correlation is strong enough, the genes for altruism will be favoured by selection. Because the neighbour-modulated

fitness framework analyses patterns of genotype-phenotype assortment within the ancestor-population, it is naturally regarded as a formal treatment of the informal 'positive assortment' explanation for the evolution of altruism. The stronger the assortment between possessing the genes for altruism and receiving the benefits of altruism, the more likely it is that the neighbour-modulated fitness increment will be positive, potentially leading to a situation in which altruists have higher neighbour-modulated fitness, on average, than non-altruists.

Inclusive fitness analyses indirect reproduction

The τ-coefficients in Frank's inclusive fitness formalism can also be regarded as measures of 'relatedness' in some sense. But they differ from the ρ-coefficients of the neighbour-modulated fitness analysis in two important respects: they are *purely genetic*, and they concern *cross-generational* correlations between the ancestor- and descendant-populations. Specifically, each of the τ-coefficients measures the association between the genotypes of the descendants of the i^{th} class and those of the actors who controlled the j^{th} influence on the fitness of their direct lineal ancestors. As Frank (1997a,b; 1998) notes, these coefficients are naturally interpreted as measures of the 'transmission fidelity' of the actor's genes through each of the fitness components for which its behaviour is causally responsible. Of course, there is usually no *literal* process by means of which the actor's (token) genes are replicated and transmitted to the recipient's descendants. But, as our informal 'indirect reproduction' story notes, something broadly analogous to this does happen when genetic relatives interact: a related recipient affords the actor something broadly analogous to an indirect channel of transmission. If the 'fidelity' of this 'transmission' is sufficiently high, then altruism may be favoured by selection, for the genetic representation an altruistic gene earns via this 'indirect pathway' may outweigh what it sacrifices through the direct pathway.

This brings out the close relationship between inclusive fitness and the 'indirect reproduction' explanation for the evolution of altruism. For, in analysing patterns of τ-

correlation between the genotypes of actors and the genotypes of their recipients' descendants, the inclusive fitness approach proceeds just as if the recipient of a social effect provided the actor with an indirect channel of genetic transmission. The framework thus formally captures the sense in which the 'indirect reproduction' metaphor is justified, by showing in precise terms how the spread of a social gene depends not only on the sign and magnitude of its fitness effects, but also on its 'transmission fidelity' through the fitness components for which it is causally responsible.

5.3 When the frameworks are formally equivalent

5.3.1 Conditions for formal equivalence

With Frank's formalism in hand, we can now address the question of when the neighbour-modulated and inclusive fitness approaches are equivalent. For current purposes, I will assume that the neighbour-modulated and inclusive fitness frameworks are 'equivalent' if and only if they cannot disagree with regard to the direction of (the primary and secondary effect of) selection on the character under study. It follows that the frameworks are 'equivalent' under some conditions if and only if the neighbour-modulated fitness increment is guaranteed to have the same sign as the inclusive fitness increment under those conditions. A more stringent conception of 'equivalence' would require that they also agree on the magnitude of the change; but in practice the direction is often what we want to know.

One might think it obvious that the two approaches are equivalent in this sense. After all, both start with the same version of the Price equation, and both proceed to decompose that equation through regression analysis. Moreover, both derivations rely on a broadly similar assumption, namely an assumption of uncorrelated residuals: we assume that the

residuals in the relevant regression equations are uncorrelated with any other variable in the analysis. In effect, this amounts to the assumption that there is no unexplained residual covariance between g' and w once we take account of the statistical associations described by the relevant regression equations (this is equivalent to the assumption that Queller's 'separation condition' is satisfied, though with g replaced by g'; cf. Chapter 4). Note, however, that the relevant regression equations differ significantly between the two frameworks. The neighbour-modulated fitness approach regresses all social phenotypes on the genotype of the *recipient*, and considers only the fidelity of direct transmission between recipients and their descendants. The inclusive fitness approach, by contrast, regresses all social phenotypes on the genotype of the *controlling actor*, and separately considers the transmission fidelity of the actor's genotype through each fitness component. The implication is that, while both derivations require an assumption of uncorrelated residuals, the *content* of that assumption differs significantly between the two cases. As a result, we cannot simply assume that the neighbour-modulated and inclusive fitness increments will always have the same sign.

Given, then, that we cannot expect the frameworks to be equivalent in all possible cases, what are the conditions under which we *can* expect them to be equivalent? Here is the thought I want to develop. In Section 5.1, we noted that the 'indirect reproduction' and 'positive assortment' pictures invoke different kinds of correlation to explain the success of altruism. The 'indirect reproduction' picture invokes correlations between actor genotypes and those of their recipients' descendants (represented by τ-correlations in Frank's formalism), while the 'positive assortment' picture invokes correlations between recipient genotypes and actor phenotypes (represented by ρ-correlations in Frank's formalism). Both kinds of correlation can be glossed as measures of 'relatedness', but it is interesting to see the subtle difference between the two pictures with respect to the *kind* of relatedness they take to matter for the evolution of altruism. We also noted, however, that in practice the two kinds of correlation might often turn out to be generated by the same

causal mechanism. Plausible candidate mechanisms for both kinds of correlation include limited dispersal (i.e., kin stick together) and kin recognition (i.e., kin detect each other).

I submit that, when any of these causal mechanisms is at work, both τ- and ρ-correlations are likely to arise; and, moreover, they are both likely to arise as a by-product of underlying genetic correlation between actors and recipients. And I further submit that, when τ- and ρ-correlations arise wholly from these mechanisms, and hence can be seen wholly as by-products of underlying genetic correlation between actors and recipients, the neighbour-modulated and inclusive fitness frameworks will in general agree regarding the direction of social selection. If this claim is correct, it is significant. For it tells us, on the one hand, that there will be a large class of cases in which the frameworks can be relied upon to agree; but it also tells us, on the other hand, that there could also be cases in which they cannot be so relied upon. These will be cases in which τ- and ρ-correlations are independently influenced by distinct causal mechanisms, rather than arising together as by-products of underlying genetic correlation between actors and recipients.

Here is a formal argument for this claim. If a mechanism such as limited dispersal or kin recognition is at work, the result will be genetic correlation among actors and recipients. Let us first introduce two new regression coefficients, γ_{ij} and $\tilde{\gamma}_{ij}$, regressing recipient genotypes on actor genotypes and *vice versa*:

$$g_{ik} = \gamma_{ij} g_{ijk} + \varepsilon$$

$$g_{ijk} = \tilde{\gamma}_{ij} g_{ik} + \varepsilon$$

Verbally, γ_{ij} represents the regression of the breeding value of the k^{th} member of the i^{th} recipient class on the breeding value of the actor who controls its j^{th} neighbourhood phenotype; $\tilde{\gamma}_{ij}$ simply regresses the latter on the former. These new coefficients can also be glossed as measures of 'relatedness' between actors and recipients, though they differ

from both the τ- and ρ-coefficients in Frank's formalism. Unlike ρ-coefficients, γ-coefficients consider only *genetic* correlation between actors and recipients, ignoring actor phenotypes. And unlike τ-coefficients, γ-coefficients consider only *intra-generational* genetic correlation between actors and recipients, ignoring the genotypes of the recipients' descendants.

If, for all classes and all phenotypes (i.e., for all *ij*), the correlation between recipient genotypes and their neighbourhood phenotypes is *fully* explained (i.e., without correlated residuals) by underlying genotypic correlation between actors and recipients (in combination with the expression of the relevant social genes in the actor), then, for all *ij*, ρ_{ij} (the regression of actor phenotype on recipient genotype) will equal the product of $\tilde{\gamma}_{ij}$ (the regression of actor genotype on recipient genotype) and d_{ij} (the regression of actor phenotype on actor genotype):

$$\forall ij \left(\rho_{ij} = \tilde{\gamma}_{ij} d_{ij} \right) \tag{A}$$

On the assumption that condition (A) obtains, we can write a new expression for the neighbour-modulated fitness increment:

$$\forall ij \left(\rho_{ij} = \gamma_{ij} d_{ij} \right) \rightarrow \tau_0 \sum_i q_i \sum_j \rho_{ij} \beta_{ij} = \tau_0 \sum_i q_i \sum_j \tilde{\gamma}_{ij} d_{ij} \beta_{ij} \tag{5.3.1}$$

Meanwhile, if (for all *ij*) the correlation between the genotype of a particular actor and the genotype of its recipients' offspring is *fully* explained by underlying genotypic correlation (in combination with the direct, vertical transmission of the relevant genes from the recipient to its offspring), then (for all *ij*) τ_{ij} will equal the product of γ_{ij} (the intra-generational genetic correlation among actors and recipients) and τ_i, the fidelity of direct, 'vertical' transmission between recipients in the i^{th} class and their direct lineal descendants:

$$\forall ij \left(\tau_{ij} = \gamma_{ij} \tau_i \right) \tag{B}$$

On the assumption that condition (B) obtains, we can write a new expression for the inclusive fitness increment:

$$\forall ij \left(\tau_{ij} = \gamma_{ij} \tau_i \right) \rightarrow \sum_i q_i \sum_j d_{ij} \beta_{ij} \tau_{ij} = \sum_i q_i \tau_i \sum_j \gamma_{ij} d_{ij} \beta_{ij} \tag{5.3.2}$$

We can get from (5.3.1) and (5.3.2) to an equivalence result by invoking two further assumptions, both of which we previously made in deriving our original expressions for the neighbour-modulated and inclusive fitness increments. First, we assume (as we did in deriving our expression for the neighbour-modulated fitness increment) that the fidelity of vertical transmission is the same in all recipient classes. This allows us to take the τ coefficient outside the summation over *i*. Second, we assume (as we did in deriving our expression for the inclusive fitness increment) that the genetic variance is the same in all actor classes. This entitles us to substitute $\tilde{\gamma}_{ij}$ for γ_{ij} in (5.3.2) without any risk of changing the overall sign of the sum over classes. With this substitution made, we can derive the following result concerning the inclusive fitness increment:

$$\text{sign}\left(\sum_i q_i \sum_j d_{ij} \beta_{ij} \tau_{ij} \right) = \text{sign}\left(\tau_0 \sum_i q_i \sum_j \tilde{\gamma}_{ij} d_{ij} \beta_{ij} \right) \tag{5.3.3}$$

Comparing this to (5.3.1), we see that:

$$\text{sign}\left(\sum_i q_i \sum_j d_{ij} \beta_{ij} \tau_{ij} \right) = \text{sign}\left(\tau_0 \sum_i q_i \sum_j \rho_{ij} \beta_{ij} \right) \tag{5.3.4}$$

In plain terms, then: conditional on (A), (B) and two additional assumptions we already made when formulating the two frameworks, the neighbour-modulated and inclusive fitness increments will always have the same sign.

In some ways, (5.3.4) represents a stronger formal equivalence result than those previously obtained (e.g., Taylor et al. 2007), since it requires relatively few assumptions, and in particular does not invoke the assumptions of weak selection or fair meiosis. As we will see, however, it does not give defenders of formal equivalence everything they might want, for it also points to important classes of cases in which the two frameworks may come apart.

5.3.2 When they are formally equivalent, which should we use?

What our equivalence result shows, in a nutshell, is that the neighbour-modulated and inclusive fitness frameworks are formally equivalent (in the sense that they can never disagree regarding the direction of the response to natural selection) whenever both ρ-correlations (that is, correlations between recipient genotypes and their social neighbourhoods) and τ-correlations (that is, correlations between actor genotypes and those of their recipients' descendants) are fully explained as by-products of underlying genetic similarity between actors and recipients. Though this will not always be the case (Section 4.4), it seems reasonable to suppose that it often will be the case when correlations between social partners are generated by the mechanisms of kin recognition or limited dispersal. This is because these mechanisms are primarily sources of genetic similarity between actors and recipients: if they also happen to generate partly phenotypic ρ-correlations or intergenerational τ-correlations, these are merely by-products of the underlying γ-correlations. Hence, when correlations between social partners arise from kin recognition or limited dispersal, our choice regarding which framework to use will usually not be forced by considerations of accuracy.

Which framework should we prefer in such cases? When considerations of accuracy do not favour one framework over the other, what does? Hamilton (1964) and Maynard Smith (1982, 1983, 1987) both preferred the inclusive fitness framework on the grounds of

what they took to be its greater theoretical simplicity and ease of application. In recent years, however, this situation has largely reversed: theorists have increasingly come to favour the neighbour-modulated fitness framework, citing *its* greater simplicity and ease of application (Taylor and Frank 1996; Taylor et al. 2007; Gardner et al. 2007; West et al. 2011). There is no doubt that, as the theories are currently formulated, neighbour-modulated fitness theory is indeed the more straightforward of the two. This is because it neglects considerations of control. To perform a neighbour-modulated fitness analysis, we do not need to know how the neighbourhood phenotypes of recipients depend on the genotypes of controlling actors, and this is one less causal pathway to worry about.

Yet two considerations speak in favour of the inclusive fitness approach, in spite of its additional complexity. One is that, in this case, more complexity means more causal explanation: while neighbour-modulated fitness may be simpler because it neglects the causal pathways linking social phenotypes to controlling actors, this simplicity comes at the cost of its explanatory power. If we want to understand how social phenotypes are controlled, and how pathways of control affect the course of social evolution, an inclusive fitness approach is more informative. A second, related consideration is that inclusive fitness, unlike neighbour-modulated fitness, underwrites an intuitive 'maximizing agent' analogy (Dawkins 1982; Grafen 1984, 2006a). This too is ultimately due to the fact that inclusive fitness, unlike neighbour-modulated fitness, is sensitive to considerations of control. For recall that an actor's inclusive fitness is a τ-weighted sum of the fitness effects attributable to *the behaviours it controls*. If these effects are at least reasonably predictable, we can put ourselves in the position of the actor and ask: 'How should I behave, in order to maximize my inclusive fitness?' Since natural selection tends to favour traits that promote inclusive fitness on average, asking this question can serve as an informal route to predictions and explanations of social behaviour. This kind of agential thinking is commonplace in behavioural ecology, where it usually lies at the heart of informal inclusive fitness arguments. Its legitimacy will be explored in greater detail in Chapter 6.

By contrast, we cannot usefully ask the same question with regard to neighbour-modulated fitness, because in cases of social behaviour there are often substantial components of an individual's neighbour-modulated fitness over which it has little or no control.[5] All we can do, by way of an agential heuristic, is put ourselves in the position of a recipient and ask: 'What genotypes are "good news", as far as my neighbour-modulated fitness is concerned?' But this heuristic is considerably less intuitive, because considerations of causation and control are replaced by considerations of statistical auspiciousness.[6] A behavioural ecologist will often have a strong intuitive grip on whether a particular strategy would add to the inclusive fitness of the agent performing it, but she will typically be much less confident about whether having a particular genotype would correlate with occupying an advantageous social neighbourhood. Strictly speaking, all this shows is that inclusive fitness is a valuable notion for the purpose of making *informal* arguments about kin selection: it does not show that inclusive fitness is worth the trouble as a *formal* analytical approach. But while this distinction is important, I suspect that the two issues are closely related. For it is surely a virtue of any formal framework that it bears at least some resemblance to the informal arguments we use to generate hypotheses and interpret results within it. If our hypotheses are based on considerations of causation and control, and if our interpretation of results is also mindful of such considerations, then it is helpful if the framework we use for formalizing hypotheses and analysing results also takes them into account. There will always be a poor fit between formal neighbour-modulated fitness models and informal inclusive fitness arguments, because the considerations of control that lie at the heart of the latter are completely absent from the

---

[5] Rosas (2010) argues that social agents do control their neighbour-modulated fitness, because they control the mechanisms of assortment (i.e., they determine who is able to influence their fitness). This is probably true in some cases (most plausibly cases involving humans) but it is obviously not true in general. In the present context, where we are concerned with correlations generated by kin recognition or limited dispersal, it is very unlikely to be true.

[6] I am grateful to Johannes Martens for discussion of this issue. His PhD dissertation includes an in-depth discussion of the relationship between kin selection theory and decision theory.

former. When we formalize kin selection theory in inclusive fitness terms, the fit between intuitive thinking and formal theorizing is much closer.

In summary: when the two frameworks are equivalent we face a trade-off between theoretical simplicity and causal-explanatory power. Inclusive fitness analyses are in general harder to perform, but they provide explanations of kin selection that are causally richer and more intuitively intelligible than those supplied by the neighbour-modulated fitness approach.

5.4 When they are not

5.4.1 Losing control

The argument of the preceding subsection was that inclusive fitness theory, by virtue of prioritizing considerations of actor control, provides deeper causal explanations of kin selection than the neighbour-modulated fitness approach. This gives us a reason to prefer inclusive fitness when the conditions for equivalence are met. At the same time, however, the centrality of actor control to inclusive fitness theory also points to one important class of cases in which the conditions for equivalence are *not* met. These are cases in which the inclusive fitness approach cannot even begin, because it is not possible to assign fitness-relevant phenotypes to controlling actors.

A preliminary remark is in order here: what is the relevant sense of 'control' in this context? Frank does not offer an account of control as part of his formalism; indeed, in spite of the pivotal role it plays in the theory, the concept of control has received relatively little attention from inclusive fitness theorists. I take the relevant notion of control to be the

notion I presented in Section 1.3: that is, control as systematic counterfactual dependence: in this context, the systematic counterfactual dependence of phenotype on genotype.

Inclusive fitness theory requires that, for every fitness-relevant behavioural phenotype, a single, determinate class of controlling actors can be identified (e.g., males or females, workers or queens—the precise nature of the actor-classes will depend on the population and social phenomenon under investigation). This is not the same as the (much stronger) assumption that, for every *token instance* of a behavioural phenotype, a particular controlling actor can be identified. Because this stronger assumption is not required, the applicability of inclusive fitness theory to real ecological contexts is not imperilled by the admission that, in such contexts, we rarely possess fine-grained causal knowledge of token social behaviours. Yet, as we will see, even the weaker assumption that we can assign types of social behaviour to classes of controlling actor is enough to limit the scope of the (standard) inclusive fitness approach.

External control

In broad terms, two types of scenario can lead to a breakdown of determinate actor control. The first is a scenario in which fitness-relevant phenotypes are not controlled by *any* actor-class within the population under study, because they are controlled by an actor-class outside that population. For instance, Taylor et al. (2007) consider a model of host-parasite interaction in which a host, by tolerating the presence of the parasite, can induce the parasite to be less virulent. On the face of it, only a neighbour-modulated fitness approach can adequately analyse such a scenario, because parasite virulence is an influence on host fitness that is not controlled by any member of the host population.

The defender of inclusive fitness might object as follows: can we not simply extend our usual conception of a 'population' in this context to include both hosts and parasites? The answer is that, while nothing formally prevents us from doing this, it would not make the problem go away. For if we were to include parasites as members of the host population,

we would have to assign them a breeding value of zero for the host trait(s) under investigation, since they would not possess any of the relevant alleles. But if every individual in the actor-class has a breeding value of zero, then there is no variance in g within that class, and the d- and τ-coefficients we need to evaluate inclusive fitness are undefined.

Nevertheless, although this natural response on behalf of inclusive fitness is not successful, a more subtle response remains available: can we not, for the purposes of analysis, treat any relevant parasite phenotypes *as if* they were under the control of the host?[7] This may seem like a substantial distortion of reality; but it need not be, if control is understood as the systematic counterfactual dependence of phenotype on genotype. For, in the host-parasite model, the host's degree of parasite tolerance—which we assume to be under its control—has a reasonably fine-grained downstream effect on the degree of parasite virulence it experiences. Given this, it is not too much of a stretch to suggest that the degree of parasite virulence is controlled by the host to a significant degree, and that the host maximizes its inclusive fitness by minimizing this virulence. Note, however, that this response depends for its plausibility on a specific feature of Taylor and colleagues' simple model: to wit, that the parasite virulence experienced by a particular host depends in a fine-grained way on its own degree of tolerance, and is not influenced by any additional variables outside of its control (such as the behaviour of other classes of host). We can easily imagine models in which this assumption does not hold. For example, if hosts of different classes regularly exchanged parasites with each other, such that the virulence expressed by a parasite depended on the *average* tolerance of the hosts it experienced, then it would no longer be reasonable to regard the virulence experienced by a particular class of host as a trait over which it has sole control.

[7] Indeed, Taylor et al. (2007) themselves advocate this response.

Delocalized control

This brings us to the second, arguably more pervasive type of problem scenario. This is a scenario in which control of a fitness-relevant phenotype is not localized to any particular actor-class. Instead, control is spread across multiple actor-classes, with each controlling the phenotype to some imperfect but significant degree. While our second host-parasite case (in which virulence depends on the average tolerance of many host-classes) might be considered one example of this phenomenon, such scenarios can arise even if interactions occur among conspecifics only. I will refer to such cases as instances of delocalized control.

One particularly important source of delocalized control is collaborative, task-based cooperation in which actors of various classes participate (Section 1.2). Consider again, for instance, the case of *Pheidole pallidula*, in which minors and majors collaborate to pin down and decapitate intruders. As in any case of task-based cooperation, the fitness benefit is conferred by the completion of the task, and it would not be conferred by either of the contributory behaviours occurring in isolation. Task completion therefore counts as a fitness-relevant correlated phenotype of the affected recipients, and it should be included as a predictor in its own right in a causal analysis of kin selection. Yet the completion of the task is not controlled by any single actor-class. The successful decapitation of intruders cannot be attributed wholly to the minors or to the majors: it is controlled to some imperfect degree by each actor-class. Because there is no actor-class to whom this phenotype may be uniquely attributed, there is no way for an inclusive fitness analysis of the evolution of intruder-decapitation to get off the ground within the standard framework.

It is possible to foresee an objection to this line of argument: surely, whether or not the inclusive fitness formalism applies to task-based cooperation depends on how we individuate phenotypes in such cases, and we have some degree of flexibility in this regard. If we count task *completion* as a phenotypic character in its own right, then of course it cannot be assigned to a single controlling actor class, and this is problematic. But

if we consider each individual *contribution* to the task as a separate character, and take the partial regression of recipient fitness on each individual contribution, we can recover actor control, for each type of contribution to the task is controlled by a particular actor-class. To see what is wrong with this response, we need to return to the argument of Chapter 4. There we saw (in the context of simple synergy games) that a synergistic effect must be explicitly represented as a predictor in a regression analysis of fitness, on pain of potentially significant inaccuracy. For if we consider only the separate behaviours of individual actors and attempt to split any synergistic effect between these two predictors, our analysis will tend to overcompensate for synergy. The reason, in a nutshell, is that the partial regression coefficients in a two-predictor phenotypic regression only compensate for synergy on the basis of correlations between the two predictors, and do not consider underlying genetic correlations.

Even in very simple cases, this point is subtle and hard to see. Crucially, however, it is a *general* problem that arises from a general feature of partial regression coefficients. Hence, if the problem occurs in simple synergy games, it is likely to recur in much more complicated cases of synergistic interaction, such as cases of task-based cooperation. The solution to the problem is still the same in more complicated cases: explicitly include synergy-producing phenotypes in the predictor set. In Queller's model, the relevant predictor is simply $\hat{z}z$, the product of actor and recipient phenotypes. In cases of arbitrarily complex task-based cooperation, the natural phenotypic predictor to include is a dummy variable that is set to 1 if a particular task is completed and 0 otherwise. The key point is that, however we choose to represent synergistic effects in practice, representing them is not merely optional. If we ignore them, there is likely to be a component of $\mathrm{Cov}(w,g')$ that our predictor set fails to account for; and if this component is large, we may potentially get the direction of selection wrong.

Note that this is, in effect, an extension of an informal argument first made in Section 2.2. There, I argued that the fitness benefit conferred by task completion does not decompose,

in any straightforward way, into components attributable to individual contributions. Now, drawing on the considerations brought to bear in Chapter 3, I am further arguing that the fitness benefit conferred by task completion *still* does not decompose into components attributable to individual contributions, even when we define each 'component' as the partial regression of recipient fitness on actor phenotype. Because of the way partial regression coefficients are defined, we cannot reliably 'average out' synergistic effects in this way: we have to represent the synergistic effect explicitly; and, in many cases (though not in Queller's simple, classless synergy game), this will be an effect over which no single actor-class has control.

I therefore contend that there is no way for Frank's inclusive fitness formalism to accommodate cases of delocalized control, at least not without producing violations of the separation condition, and a concomitant loss of accuracy. I suspect, however, that the formalism can be extended to cover such cases. Roughly speaking, the best way to do this is to replace the simple regression of z on controlling actor genotype in the standard formalism with a *multivariate* regression of z on *all* relevant controlling actor genotypes (see Appendix D). This extended formalism for inclusive fitness still assigns fitness components by considerations of causal responsibility, but the conception of 'causal responsibility' it employs is more nuanced than that of the standard framework. Rather than assuming that every social phenotype is under the sole control of a single actor-class, we allow that control of a social phenotype can be distributed across any number of actor-classes (e.g., majors and minors). Then, rather than conceptualizing an actor's inclusive fitness as a τ-weighted sum of the fitness effects caused by the social phenotypes under its sole control, we conceptualize an actor's inclusive fitness as a τ-weighted sum of the fitness effects caused by the social phenotypes over which it has *some degree* of control, where the fitness effects are *also* weighted by a measure of the degree to which the actor controls the phenotype in question. In the limiting case in which every social phenotype is controlled by a single actor-class, the extended formalism becomes equivalent to the standard formalism. The upshot is that, although formalizations of inclusive fitness

routinely make an assumption of localized control, we may well be able to relax this assumption without having to abandon the notion of inclusive fitness altogether.

5.4.2 Sui generis ρ-*correlations*

In Section 5.4.1, I argued that the neighbour-modulated and inclusive fitness approaches are equivalent whenever ρ- and τ-correlations are fully explained by underlying genetic correlations between actors and recipients. This suggests that, if we want to find situations in which they are non-equivalent, we would be well advised to look for situations in which ρ- or τ-correlations are explained at least in part by something other than underlying genetic correlation between actors and recipients.

Let us consider ρ-correlations first. These, recall, are correlations between the genotypes of recipients and the social phenotypes by which their fitness is affected. Although such correlations may well arise most commonly from underlying correlations between actor and recipient genotypes, it is not hard to imagine various other possible sources. In one broad class of cases, *sui generis* ρ-correlations arise because recipient genotypes correlate with social phenotypes that are not controlled by any member of the population, but are instead controlled externally (e.g., by a member of a different species). These cases have already been considered in Section 5.4.1, under the heading of 'external control' (see also Frank 1997b; Fletcher and Zwick 2006; Taylor et al. 2007; Fletcher and Doebeli 2009).

In a second broad class of cases, *sui generis* ρ-correlations may arise because particular genotypes have 'extended' phenotypic effects on the phenotypes of other organisms within the same population (cf. Dawkins 1982). In recent years, such effects have been extensively studied under the banner of 'indirect genetic effects', or 'IGEs' (Moore et al. 1997; Wolf et al. 1998; Bijma and Wade 2008; McGlothlin et al. 2010). The classic examples of IGEs are non-social: for instance, maternal genes that influence a mother's production of milk can significantly affect the growth rate of her newborn offspring (Bijma 2006). But it

would not be surprising if comparable effects occur with respect to social traits. In this vein, Allen J. Moore and colleagues (1997) survey various examples where the expression of a social phenotype may plausibly be influenced by the genes of individuals other than the actor who expresses it, including examples of mating behaviour, aggression and social dominance. Joel W. McGlothlin and colleagues (2010) develop a detailed formal framework of their own for studying the impact of IGEs on social evolution, but I will not discuss this formalism here.[8] For, although I suspect that such a framework will prove invaluable for understanding how IGEs affect social-evolutionary processes, it is important to note that IGEs can, in principle, be accounted for within Frank's general framework for neighbour-modulated fitness. Indeed, since the neighbour-modulated fitness approach makes *no assumptions at all* about the pathways of control linking genotype to phenotype, it allows for arbitrarily complex networks of inter-organismal gene regulation. Of course, 'allow for' does not mean 'explain': because the framework simply *ignores* pathways of genetic control (be they direct or indirect), it will not provide any deep understanding of how such pathways affect the course of evolution.[9]

A third and final class—which should arguably be regarded as a subset of the second—comprises cases of strategic reciprocity, in which the expression of a social behaviour in one individual induces a social response from the individuals it affects. Jeffrey A. Fletcher and Martin Zwick (2006) consider an iterated Prisoner's Dilemma game (a model made famous by Axelrod and Hamilton 1981) in which pairs of agents interact randomly but

---

[8] One awkward feature of the McGlothlin et al. approach is that they describe 'relatedness' and 'IGEs' as separate effects on social selection. What they should say, I think, is that the usual mechanisms responsible for relatedness (kin recognition, limited dispersal, etc.) are sources of ρ-relatedness, but so are IGEs. IGEs affect social selection *via* their effect on ρ-relatedness.

[9] IGEs may also lead to a situation in which control of a social phenotype is substantially shared between the agent that expresses the behaviour and the agent whose genes affect it indirectly. If this occurs, there is a further problem for the (standard) inclusive fitness framework, viz., a problem of delocalized control. But here I am emphasizing a more obvious problem: even if IGEs do not lead to delocalized control, they still present a problem for the inclusive fitness framework by virtue of producing *sui generis* ρ-correlations.

repeatedly. Agents play one of two strategies: 'tit for tat' (TFT), whereby, on meeting a social partner they have met previously, agents copy the strategy this partner played on their previous meeting; or 'always defect' (ALLD), whereby agents always defect, regardless of what their social partner has done in the past. We stipulate[10] that $g = 1$ for TFT and $g = 0$ for ALLD. Since pairings are random, there can be no *genetic* correlation between social partners. Nevertheless, genotype *does* correlate positively with receiving the benefits of altruism. This is because *sui generis* ρ-correlations are generated by the TFT strategy. To see why, note that agents who play TFT will only behave altruistically towards a partner they have met before if that partner was previously altruistic. This ensures that agents who play ALLD are, on the whole, less likely to receive the benefits of altruism than agents who play ALLD.

Though the link is not immediately apparent, strategic reciprocity is still an 'indirect genetic effect' in a manner of speaking. It is simply that the effect is socially mediated: one agent's TFT genotype is expressed in the form of an altruistic behaviour, and this social behaviour then induces the expression of a similar behaviour in a TFT social partner. Having the TFT gene thus has differential effects on the behaviour of one's social partners, and this is what causes the *sui generis* ρ-correlation—which has downstream consequences for the course of social evolution. This is a special case of the general kind of phenomenon that the IGE research programme seeks to capture (cf. Bijma and Wade 2008). In all these scenarios, altruistic behaviours can potentially evolve if the ρ-correlation between possessing the genes for altruism and receiving the benefits of altruism is sufficiently strong. Yet none of these mechanisms requires τ-correlations between social partners, and hence none involves anything like the kind of 'indirect reproduction' often thought to be central to the concept of kin selection.

[10] These are not breeding values, though they are *p*-scores *sensu* Grafen 1985a (see Chapter 2).

5.4.3 Sui generis τ-*correlations*

The possibility of *sui generis* ρ-correlations lies at the heart of recent arguments for the general superiority of the neighbour-modulated fitness framework over the inclusive fitness alternative (see Frank 1997b; Fletcher and Doebeli 2006, 2009, 2010; Fletcher and Zwick 2006; Fletcher et al. 2006). I am broadly sympathetic to this line of argument: as the cases canvassed in the preceding section show, there is certainly a significant range of circumstances in which we cannot safely assume the neighbour-modulated and inclusive fitness frameworks to be equivalent, and in which we need to explicitly analyse ρ-correlations to understand why social selection proceeds as it does. Nevertheless, I think the Fletcher/Doebeli argument misses something important, namely that, just as it is possible for there to be *sui generis* ρ-correlations, it is *also* possible for there to be *sui generis* τ-correlations. These correlations, when they arise, will be represented in an inclusive fitness analysis but will be neglected by a neighbour-modulated fitness analysis.

Altruism causes biased transmission

Here is a far-fetched (but instructive) story about how *sui generis* τ-correlations might arise. Suppose that we have a population in which social partners' genotypes are completely uncorrelated. Some agents nevertheless act altruistically towards their social partners. Intuitively, we would expect this to be a scenario in which altruism simply does not pay: altruists will be, on average, less fit than non-altruists, and the genes for altruism will not spread. But suppose there is a curious twist: by virtue of performing the altruistic act, an altruist somehow confers on its partner a disposition to produce offspring with the genes for altruism. Let us say nothing (for now) about how this could possibly work in reality. The important point is simply that if the fitness benefits the altruist confers on the recipient are large enough and if the disposition to produce altruistic offspring it also confers is strong enough, then the genes for altruism might spread after all. But they would spread not because receiving the benefits of altruism co-varies with possessing the

genes for altruism, but rather because receiving the benefits of altruism co-varies with a disposition to produce offspring more altruistic than oneself.[11]

A neighbour-modulated fitness analysis would miss this, because (as the story stipulates) there are no ρ-correlations at all between social partners: bearers of the genes for altruism are *not* differentially likely to receive the benefits of altruism (cf. Picture 2). But there *are* intergenerational τ-correlations: the offspring of a recipient *do* bear a genetic resemblance to the actor, but this resemblance arises *sui generis*, unmediated by any pre-existing genetic resemblance between actors and recipients. The inclusive fitness framework will not miss these correlations, and so will be able to assess whether or not they are sufficiently strong for altruistic genes to spread.

What this story describes, in effect, is a case of 'indirect reproduction' without any pre-existing positive assortment. This is the general sort of case in which *sui generis* τ-correlations will arise, and in which the inclusive fitness approach will enjoy greater accuracy than the neighbour-modulated fitness approach.

Altruism and mobile genetic elements

The above story may seem fanciful. How could behaving altruistically towards a selfish individual cause it to produce offspring more altruistic than itself? Strange as it sounds, however, a mechanism not too far removed from this possibility may occur for real in microbial populations. Over the past decade, the application of social evolution theory to microbial populations has developed into a lively and fruitful research programme (for reviews, see Crespi 2001; Velicer 2003; West et al. 2007b; Velicer and Vos 2009; Damore and Gore 2012). One major respect in which social evolution in microbial populations differs from social evolution in populations of multicellular organisms is that patterns of

[11] In other words, the trait spreads because $\mathrm{Cov}(w, \Delta g) > 0$, even though $\mathrm{Cov}(w, g) < 0$. In the terminology of Chapter 2, it spreads because of the *secondary* effect of natural selection.

transmission are greatly complicated by horizontal gene transfer, in which genes are copied from one organism to another via independently replicating 'mobile genetic elements' (MGEs) (see, e.g., Frost et al. 2005). It has recently been observed that MGEs carry a greater number of cooperative traits than one might expect, if their phenotypic effects were typical of the phenotypic effects of a randomly sampled region of functional DNA (Rankin et al. 2011a). What could explain this over-representation of cooperative traits on MGEs? As Daniel Rankin and colleagues (2011a) note, one possible explanation is that gene mobility actually facilitates the evolution of cooperation.

Here is one way in which this might work. Microbes are restricted to a particular patch, and interact only with the other inhabitants of that patch. Competition, however, is global: it mostly occurs between patches rather than within patches. Moreover, within a patch, there is no pre-existing positive assortment: altruists are no more likely than average to interact with other altruists. At time *t*, an altruistic MGE causes its bearers to confer fitness benefits on the other inhabitants of their patch, at a cost to themselves. In absolute terms, the total benefits outweigh the total costs—but at first glance these benefits appear worthless from the point of view of the focal gene, because they fall on individuals who are no more likely than average to share that gene. But then, something else happens: at time $t + \delta$, the MGE spreads outwards from any surviving altruists to other members of the patch—that is, it spreads to individuals who previously received the fitness benefit at *t*. If this MGE spreads effectively enough throughout its patch, and if the benefits it conferred at *t* were sufficiently large, then it can spread through the global population as a consequence of the altruistic behaviour it causes. But it spreads not because receiving the benefits of altruism (at *t*) co-varies with possessing the genes for altruism (at *t*), but rather because receiving the benefits of altruism (at *t*) co-varies with a tendency to subsequently acquire the genes for altruism (at $t + \delta$).[12]

[12] I discuss this possible mechanism in greater detail in Birch 2013b. It is slightly different from the mechanism Rankin et al. (2011a) themselves propose: in their version, horizontal gene transfer occurs *prior* to

As in the 'altruism causes biased transmission' story, a neighbour-modulated fitness approach is liable to miss the correlations that make this process work. For it considers only ρ-correlations (i.e., measures of pre-existing positive assortment); and, by hypothesis, there are no ρ-correlations at the time of interaction. What matters are the diachronic τ-correlations between the genotype of the actor at t and the genotype of the recipient at $t + \delta$ (or, alternatively, the genotype of the recipients' lineal descendants at some even later time). Because we once again have a case of *sui generis* τ-correlation, we again have a process of social selection that the inclusive fitness framework is more likely to analyse accurately.

This 'ship-jumping' mechanism for the evolution of altruism is a possible explanation for the disproportionate representation of cooperative traits on MGEs. It is highly conjectural, and it is only one of several credible hypotheses in the mix (for various alternatives, see Rankin et al. 2011a,b; Giraud and Shykoff 2011). For current purposes, however, what matters is merely that it is empirically *credible*, given what we currently know—and yet it is not too far away from the 'altruism causes biased transmission' story that initially seemed so outré. The wider moral is that, while *sui generis* τ-correlations may be less common than *sui generis* ρ-correlations, they deserve to be taken seriously as a live empirical possibility, particularly in microbial populations. If there are cases in which such correlations arise, then there are cases in which we have 'indirect reproduction' without positive assortment. These are cases that inclusive fitness can handle and neighbour-modulated fitness cannot.

---

social interaction, so that social partners are positively assorted at the time of interaction. This version is less convincing as an example of *sui generis* τ-correlation.

5.4.4 Review

I began this chapter by identifying two rival camps in contemporary kin selection theory: those who think the neighbour-modulated and inclusive fitness frameworks are two equivalent perspectives on the same process; and those who think that they come apart, and that the neighbour-modulated fitness approach is the more general of the two. My analysis shows that the issue is more complex than either camp often seems to realize.

The 'equivalence' camp gets something right, for the two theories are indeed equivalent in a wide range of cases, particularly when correlations between social partners are caused by kin recognition or limited dispersal. Yet predictive equivalence does not imply explanatory equivalence; and there are reasons why the inclusive fitness approach, despite (or more accurately, because of) its additional complexity, provides deeper explanations of the direction of social selection. By representing pathways of genetic control, it provides an explicit analysis of an important causal process that the neighbour-modulated fitness approach ignores completely; and this has the added bonus of enabling it to underwrite a maximizing-agent analogy (cf. Chapter 6).

The 'neighbour-modulated fitness is more general' camp also gets something right, for there are several important classes of case in which the two frameworks do come apart, and in which the neighbour-modulated fitness approach (by virtue of neglecting pathways of genetic control) turns out to fare better than the inclusive fitness alternative. Yet there are also empirically credible cases in which this situation is reversed. These are cases in which horizontal transmission generates diachronic genetic correlation between social actors and their recipients' descendants, even though there is no positive assortment when they actually interact. Though such cases may be very rare, we cannot take their rarity for granted, particularly given that horizontal transmission is such a widespread feature of microbial populations. The implication is that there may be no 'simple and general explanation for the evolution of altruism' (cf. Fletcher and Doebeli 2009): only two

alternative theories that are substantially overlapping but non-coextensive in their domains of application.

I will close this chapter by relating these considerations to the two informal pictures of the evolution of altruism introduced in Section 5.1. At the time, I suggested that the 'positive assortment' picture and the 'indirect reproduction' picture were not obviously equivalent, and that the question of their equivalence would have to be settled by a more formal treatment. We can now answer that question, though the answer is not as straightforward as we might have hoped. When altruism pays due to kin recognition or limited dispersal, we can tell the story in terms of 'positive assortment' or in terms of 'indirect reproduction', because the kinds of correlation central to both stories will obtain. But in other cases, only one story is valid. When altruism pays due to *sui generis* ρ-correlations between possessing the genes for altruism and receiving its benefits, the positive assortment story will be correct, but it will be misleading to talk of 'indirect reproduction'. By contrast, when altruism pays due to *sui generis* τ-correlations between the genotypes of actors and the genotypes of their recipients' descendants, the opposite is true: we can reasonably gloss the overall process as a kind of 'indirect reproduction', but it will be misleading to talk of positive assortment.

SIX

Do Organisms Maximize Their Inclusive Fitness?

The notion that natural selection is a process of fitness-maximization gets a bad press in population genetics, and understandably so. Yet in other areas of biology, the view that organisms behave as if maximizing their fitness (or, in cases of social behaviour, their *inclusive* fitness) remains widespread. In a series of recent papers, the Oxford geneticist Alan Grafen has sought to reconcile population genetics with fitness-maximization through a research programme he terms 'Formal Darwinism' (Grafen 1999, 2002, 2003, 2006a,b, 2007a,b,c, 2008, 2009). In this chapter, I explain and ultimately criticize this attempted rapprochement.

In Section 6.1, I distinguish, in abstract terms, four varieties of maximization in evolutionary theory and examine the relations between them, withholding judgement as to which varieties (if any) are theoretically defensible. In Section 6.2, I consider where the most famous maximization principle in population genetics, Ronald A. Fisher's (1930) 'fundamental theorem of natural selection', fits with respect to this four-part distinction. I emphasize in particular that the fundamental theorem, even if correct, tells us nothing regarding what, if anything, individual organisms act as if attempting to maximize by means of their behaviour. In other words, the theorem falls short of underwriting an 'individual-as-maximizing-agent' analogy of the sort routinely employed in behavioural ecology (Grafen 1999, 2002, 2008). Grafen's ongoing 'Formal Darwinism' project can be regarded as an attempt to close this somewhat troubling theoretical lacuna. In broad terms, Grafen's strategy is to provide a solid theoretical basis for an 'individual-as-

maximizing-agent' analogy by proving formal links between evolutionary models and optimization programmes. In Sections 6.3 and 6.4, I critically examine Grafen's approach and argue that, ingenious though it is, it falls short of vindicating the maximization principle he intends it to vindicate.

6.1 Four varieties of maximization

6.1.1 Total versus *partial change*

For many authors (including Fisher 1930) the allure of a maximization principle for evolutionary theory arises from the hope that there might exist a biological analogue of the second law of thermodynamics (Price 1972b; Edwards 2007). Thermodynamics tells us that spontaneous physical or chemical transformations are such that the entropy of the universe always increases, and that the free energy of the system thus transformed always decreases; this principle is utterly fundamental to modern chemistry. A biological analogue of such a principle, capable of playing a similar foundational role, might be characterized by the following abstract schema:

> **MAX-A *(Total change)*:** For some property π, evolution is such that the population mean of π (i.e., $\overline{\pi}$) never decreases between earlier and later time-slices of the same population.

The statement deliberately leaves open the nature of the property with a non-decreasing mean. The most obvious candidate for the π-placeholder is individual fitness, *w*.

For reasons we will encounter in Section 6.2, however, it is very implausible to suggest that the mean fitness of an evolving population can never decrease over time. This has led some theorists to formulate more modest maximization theses. One influential idea is to formulate a maximization thesis not in terms of the direction of the *overall* evolutionary

change, but rather in terms of the direction of the *partial* evolutionary change for which natural selection is directly responsible, bracketing off the effects of processes such as genetic mutation and environmental change. In broad terms, we can think of this 'partial change' as the change that *would* occur in a population if the selective regime were held fixed and all other influences on the evolutionary change were nullified. We saw in Chapter 3 that quantifying this partial change is a subtle business (see also Section 6.2), but the qualitative idea is easy enough to grasp. In abstract terms, then, we might seek to defend a maximization thesis along the following lines:

> **MAX-B *(Partial change)*:** For some property π, the *partial change attributable to natural selection* is such that, if it were the only partial change relevant to the direction of evolution, $\bar{\pi}$ would never decrease between earlier and later time-slices of the same population.

As we will see in Section 6.2, however, even this weakened type of maximization thesis faces significant theoretical challenges.

One might also wonder what the *biological significance* of a MAX-B-type maximization thesis would be, were it to be vindicated. After all, it concerns only what *would* happen in a hypothetical scenario in which all evolutionary processes distinct from selection are abolished. Real evolutionary processes are not like this; and so, on the face of it, MAX-B tells us nothing about real evolutionary processes. In my view, however, such scepticism is misplaced: a MAX-B-type maximization thesis *would* still represent a valuable result, chiefly because it would bring new questions into focus. Once we knew that some variable *would* be maximized by selection under specified conditions, we could proceed to ask how closely an actual population would need to *approximate* these conditions for the same outcome to ensue. In particular, we might ask: how strong would selection need to be in comparison to the other 'partial changes' for the upward trend to occur? Conversely, how weak would the other 'partial changes' have to be before we could safely neglect their

influence on the direction of evolution—or could we never safely neglect them? MAX-B, if it turned out to be defensible, would matter not because it would finish the debate concerning the relationship between selection and maximization, but rather because it would steer our attention towards these further issues.

6.1.2 What is doing the maximizing?

In MAX-A and MAX-B, the variable that is maximized (i.e., invariably increased until variation disappears) is a *population mean* with respect to some property—paradigmatically, the population mean with respect to individual fitness. In a sense, therefore, the entity that is 'doing the maximizing' is the *whole population*, rather than any of the *individual organisms* it contains at any given moment. After all, the individual organisms live and die with the genes they were born with: only the population persists over evolutionary time, so only the population changes its properties in response to evolutionary processes (cf. Nowak 2006a; Godfrey-Smith 2009). Alternatively, we might say that selection or evolution or 'Mother Nature' is 'doing the maximizing' by acting on the population (cf. Dennett 1995). Of course, nothing is *literally* doing the maximizing: the point here is that the process of maximization occurs at the population level.

MAX-A and MAX-B capture the senses of 'maximization' that are most commonly at stake when the issue of maximization is raised in population genetics and its philosophy (e.g., Fisher 1930; Price 1972b; Ewens 1989, 2004, 2010; Frank and Slatkin 1992; Edwards 1994, 2000, 2007; Okasha 2008). But they are not the only senses of 'maximization' that matter in evolutionary biology. In behavioural ecology, it is common to regard *individual organisms* as behaving as if trying to maximize some evolutionarily significant variable, typically their fitness or their inclusive fitness. Ecologists use this principle as the basis for an agential heuristic: we hypothetically suppose an individual organism to be a rational agent seeking to maximize its fitness or inclusive fitness, and we ask: 'what strategy, from the set of options available to it, would it rationally adopt?'. The answer to this question is

then used as a means of predicting and explaining the strategy the organism has in fact evolved. As Alan Grafen (2007a) observes:

> Empirical biologists in many fields have routinely assumed since the 1970s that natural selection leads organisms to act as if (more or less) maximizing a quantity often called fitness, intended to be roughly the lifetime number of offspring, and base research projects on that foundation. (Grafen 2007a, 1243)

We can say that ecologists who think in this way are employing an 'individual-as-maximizing-agent' analogy (Grafen 1984, 1999, 2002, 2006a, 2007a, 2008). Agential thinking of this sort is rife in many areas of theoretical sociobiology, including evolutionary game theory and optimality modelling, and it appears explicitly in informal arguments that appeal to the 'inclusive fitness interests' of an organism (e.g., Bourke and Franks 1995; Bourke 2011).

On this sense of 'maximization' it is individual organisms, not evolving populations, that 'do the maximizing'. Yet this individual-as-maximizing-agent analogy remains closely wedded to evolutionary considerations, since it is usually assumed that organisms behave as if maximizing some variable only because the variable is evolutionarily significant, and only because the organisms have evolved by natural selection to maximize it.[1] In the abstract, therefore, our formulations of the individual-as-maximizing-agent analogy will bear some resemblance to the principles of population-level maximization given in MAX-

[1] This presupposes that it makes sense to talk of an individual organism behaving in a certain way because of past natural selection. There is, however, a tradition in the philosophy of biology of denying that natural selection can ever help explain why an individual organism has the traits it has, or (by implication) why it behaves the way it does (Sober 1984, 1995; Walsh 1998; Pust 2002, 2004). The thought is that selection can explain the *distribution* of traits in a population, and perhaps the *origin* of trait *types*, but never the *instantiation* of trait *tokens* by specific individuals. I have argued elsewhere that this 'negative view' of natural selection relies on an implausibly narrow conception of causal explanation (Birch 2012c).

A and MAX-B: they will appeal to the same evolutionary processes, but will focus on their consequences for individual behaviour rather than on their consequences for population means.

We might, for instance, attempt to spell out the individual-as-maximizing-agent analogy as follows:

> **MAX-C *(Individual as maximizing agent, Total change)*:** For some property π, evolution is such that (given sufficient time and sufficient genetic variation in π) it reliably causes individual organisms to act as if maximizing π.

For completeness, we should note that a weakening move similar to that from MAX-A to MAX-B is available (and might also be useful; see Section 6.4) in the context of individuals as maximizing agents. That is, by focussing purely on the partial change attributable to natural selection (bracketing off the effects of all other evolutionary processes, including cultural evolution) we might formulate a weaker maximization principle along the following lines:

> **MAX-D *(Individual as maximizing agent, Partial change)*:** For some property π, the partial change attributable to natural selection is such that, if it were the only partial change relevant to the direction of evolution, then (given sufficient time and sufficient genetic variation in π) it would reliably cause individual organisms to act as if maximizing π.

As with MAX-B, the correctness of MAX-D is open to debate, and its correctness ultimately turns on subtle theoretical and philosophical considerations. For now, I merely want to note that MAX-C and MAX-D are *prima facie* distinct, and that both of them are *prima facie* distinct from MAX-A and MAX-B. To assert that natural selection would cause

individual organisms to maximize π in the absence of countervailing influences is not to assert that evolution as a whole causes individual organisms to maximize π, and neither of these assertions is equivalent to the claim that selection or evolution maximizes $\overline{\pi}$.

6.1.3 Relations between the varieties

We have now distinguished four varieties of maximization, summarized below:

	Total change	Partial change due to natural selection
Populations maximize $\overline{\pi}$	**MAX-A**	**MAX-B**
Individual organisms maximize π	**MAX-C**	**MAX-D**

It is worth dwelling briefly on the relations between these varieties. In particular, it is worth emphasizing just how different the four varieties are. For I submit that, despite their superficial similarities, none of them entails any of the others. I will argue for this claim piecemeal, by separately considering each of the rows and columns.[2]

Row 1: MAX-A does not entail MAX-B, nor vice versa

The first non-entailment I want to consider concerns the first row. MAX-B, the claim that selection *acting alone* would maximize the population-level property $\overline{z}$, does not logically entail MAX-A, the claim that evolution *as a whole* maximizes the same quantity. For, as Fisher emphasizes, 'natural selection is not evolution' (1930, vii): even if MAX-B is correct, other evolutionary influences may counteract any maximizing tendency on the part of selection.

[2] I assume that, if there is no entailment across the rows or down the columns, then there is no serious prospect of entailment across the diagonals.

Less obviously, MAX-A does not logically entail MAX-B. For even if evolution was found to maximize some quantity, we would not be entitled to infer from this that natural selection would also maximize this quantity (or any other quantity) in the absence of other evolutionary processes. The reason is that we would not be able to rule out *a priori* the possibility that a process other than natural selection was responsible for the systematic bias in the direction of evolution. For instance, we would not be able to rule out *a priori* the possibility that the observed directionality arose from adaptively biased mutation. We might doubt this on empirical grounds, but it is clearly a logical possibility.

Row 2: MAX-C does not entail MAX-D, nor vice versa

Parallel considerations show that MAX-C and MAX-D are also logically independent of one another. MAX-D does not entail MAX-C, because the fact that selection *would* lead organisms to behave as if maximizing some quantity in the absence of countervailing partial changes does not entail that evolution as a whole, in which other partial changes are usually significant, will actually lead organisms to behave in this way. And MAX-C does not entail MAX-D because, even if it did turn out that evolution leads organisms to behave like maximizing agents, we could not infer *a priori* that *natural selection*—as opposed to some other process—was responsible for this behaviour.

It might be objected that, although MAX-D does not *logically entail* MAX-C, it still provides some degree of *inductive support* for MAX-C. The thought would be that, generally speaking, natural selection is by far the strongest influence on the evolution of behavioural phenotypes; and so, generally speaking, the outcomes of behavioural evolution are unlikely to depart significantly from the hypothetical outcomes of 'pure' natural selection. Naturally, if we have specific evidence that natural selection was *not* the strongest influence in a particular case (for instance, evidence that non-genetic transmission was important), then we should not infer MAX-C from MAX-D in that case, but in the absence of such evidence this inference seems reasonable. Even this, however, may be doubtful. The assumption that, generally speaking, natural selection is by far the strongest influence

on the direction of evolution amounts to a form of empirical adaptationism—one that Tim Lewens (2009) has termed 'pan-selectionism'—and it is far from uncontroversial (Orzack and Sober 1994, 2001; Sober 2008).

Column 1: MAX-A does not entail MAX-C, nor vice versa

Let us now turn to the relations between population-level and individual-level versions of the maximization thesis. Let us first go up the first column: does MAX-C entail MAX-A? In other words, if evolution reliably caused individual organisms to act as if they were maximizing their personal π-value, would it follow that evolution maximizes $\bar{\pi}$? It would not. For it is a familiar idea in classical game theory that a network of interacting agents, each trying to maximize the same quantity, may nevertheless arrive at an equilibrium that fails to maximize the total (or average) amount of that quantity they receive. The same is true in evolutionary dynamics, where the 'desired' quantity is usually assumed to be fitness (Nowak 2006a). Much theoretical work in recent years (reviewed by Rankin et al. 2007) has focussed on evolutionary 'tragedy of the commons' scenarios (*sensu* Hardin 1968), in which a population of individual maximizers arrives at an outcome that is disastrous for all of them, canonically through their overexploitation of a shared resource.

Conversely, if evolution maximizes $\bar{\pi}$, does it follow that it will reliably cause individual organisms to act as if they were maximizing their personal π-value? It does not, because MAX-A by itself implies nothing about how an individual's π-value is *controlled* (Grafen 2002, 2006a). Suppose, for example, that selection acts to maximize, but that there is nothing whatsoever any individual can do *by means of its own behaviour* to alter its π-value (perhaps because its π-value is wholly determined by non-behavioural traits). In such a scenario, organisms would not evolve to 'act as if maximizing π' in any meaningful sense: the agential heuristic would fail to yield any useful predictions or explanations, for there is no sense in asking 'what would an organism rationally choose, from the options available, in order to maximize π?' when *none* of the available options has even the slightest

relevance to its π-value. This is a scenario in which MAX-C does not apply, even though (by hypothesis) MAX-A does.

This may seem like a pedantic point, but it turns out to be extremely important. For it shows that, even if we could establish that evolution maximizes the average individual fitness of a population, we would not thereby be entitled to infer that evolution causes individual organisms to behave as if maximizing their individual fitness. Of course, it would be plainly untrue to suggest that an organism can *never* alter its personal fitness by means of its behaviour. But there are still *components* of its personal fitness that it usually does not control. These are the 'neighbour-modulated' components of its fitness, which arise through the social behaviour of other organisms (cf. Chapter 5). The implication is that, even though the mean personal (neighbour-modulated) fitness of a population might in principle be maximized by evolution, we should not expect social evolution to cause organisms to behave as if maximizing this quantity. As we will see (and as Grafen 2006a forcefully emphasizes) the situation is more promising when we consider *inclusive* fitness, because this often *is* a property over which its bearer has full control.

Column 2: MAX-B does not entail MAX-D, nor vice versa

MAX-B does not entail MAX-D since, as we have already seen, it is possible in principle for selection to maximize the mean of a quantity (e.g., neighbour-modulated fitness) over which an individual organism does not control by means of its own behaviour. Conversely, MAX-D does not entail MAX-B, because a 'tragedy of the commons' scenario in which individual maximizers produce a catastrophic collapse in mean fitness is no less of a possibility when natural selection is the only evolutionary process at work.

6.2 The status of Fisher's 'fundamental theorem of natural selection' (FTNS)

The most recognizable maximization principle in evolutionary theory is Fisher's (1930) 'fundamental theorem of natural selection' (FTNS), which states, in broad terms, that 'the increase of average fitness of the population ascribable to changes in gene frequency... is equal to the [additive] genetic variance in fitness' (Fisher 1941, 377). Since variance cannot be negative, the implication is that 'the increase of average fitness of the population ascribable to changes in gene frequency' can never decrease. It will prove helpful to consider how this influential but perennially controversial theorem relates to our four varieties of maximization.

6.2.1 Old and new interpretations

In the decades from its original publication until 1972, Fisher's theorem was interpreted almost without exception as an attempt to establish a form of MAX-A, with the population mean for individual fitness playing the role of the maximand $\bar{\pi}$. That is, early commentators took Fisher to be claiming that the mean fitness of an evolving population cannot decrease; and moreover that, as long as there is additive genetic variance in fitness (i.e., variation in individual breeding values for fitness), the mean fitness will increase at a rate equal to this variance (cf. Price 1972b; Ewens 1989; Edwards 1994, 2007; Grafen 2003; Okasha 2008).

This interpretation was understandable, given Fisher's verbal presentation of the theorem, but problematic. For, on this interpretation, the theorem is undoubtedly false, as these early commentators were quick to point out. It is not true that mean fitness always increases whenever there is additive genetic variance in fitness. The easiest way to see this is to consider cases of evolution in which processes other than natural selection are the dominant influences on the direction of evolution. Consider, for example, a population influenced primarily by mutation: it could easily be that the mutations which arise in some

particular generation happen to be uniformly detrimental to fitness, and that this brings about an overall decrease in the population mean. For another example, consider a population in which the mean fitness is influenced primarily by environmental change: a population of lake-dwelling fish, perhaps, in which mean fitness is greatly decreased by the sudden drying-up of the lake. Moreover, Fisher was well aware of the problems for MAX-A-type maximization theses, having himself constructed an example in which the mean fitness of an evolving population fails to increase (Edwards 1994; Okasha 2008).

In 1972, George R. Price (1972b) suggested a novel interpretation of FTNS, on which the theorem is regarded (in our terminology) not as a failed attempt at establishing MAX-A but rather as a more promising attempt at establishing MAX-B. On this 'new' interpretation, introduced to mainstream genetics by Warren Ewens (1989), the theorem is concerned not with the *overall* direction of genetic evolution but only with the direction of the *partial* change attributable to natural selection. It should thus be read as stating that (in Price's terms):

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

The motivation for the new interpretation is that it appears to make FTNS far more defensible on theoretical grounds. On the new interpretation, the theorem is far from obviously false, and possibly even true.

6.2.2 FTNS and the Price formalism

Though Price (1972b) himself re-derived FTNS using Fisher's original notation, his own (1970, 1972a) formalism allows us to see more clearly why FTNS, glossed as a claim about the *partial* change attributable to natural selection, might be justified after all (Frank 1997a,

1998; Grafen 2002). We start with the full genetic Price equation for the evolution of some unspecified phenotypic property *z*. We start, that is, with equation (3.2.3):

$$\Delta\bar{z} = \frac{1}{\bar{w}}\left[\mathrm{Cov}(w, g) + \mathrm{E}(w\Delta g)\right]$$

We then note that individual fitness, *w*, is itself an evaluable phenotypic property of individual organisms, and so is quite capable of occupying the '*z*' placeholder in the Price equation. Moreover, differences in individual fitness are partly explained by allelic differences, so any individual will have an evaluable breeding value for fitness, g_w. This quantity represents its fitness as predicted by the linear combination of its allelic values, where each such value is weighted by the average effect of the allele on fitness. Hence, setting $z = w$ and correspondingly setting $g = g_w$, we obtain:

$$\Delta\bar{w} = \frac{1}{\bar{w}}\left[\mathrm{Cov}(w, g_w) + \mathrm{E}(w\Delta g_w)\right]$$

By definition, the covariance between a phenotypic character and its breeding value is equal to the variance in the breeding value, implying that:

$$\Delta\bar{w} = \frac{1}{\bar{w}}\left[\mathrm{Var}(g_w) + \mathrm{E}(w\Delta g_w)\right] \qquad (6.2.1)$$

Equation (6.2.1) is a straightforward implication of the Price equation, and it shows clearly why FTNS cannot be correct as a general statement about the *overall* change in mean fitness over evolutionary time. For it shows that, in general, the overall change in mean fitness between ancestor- and descendant-populations depends not only on the additive genetic variance in fitness in the ancestor-population but also on a *second* term, $\mathrm{E}(w\Delta g_w)$, of which FTNS takes no account. Yet the equation also shows why FTNS might be rather more promising if interpreted as a statement about *partial* change.

In Chapter 3, I argued that the covariance term in the Price equation could legitimately be identified with the partial change attributable to the *primary effect* of natural selection. This was a conceptual identification rather than a substantive empirical thesis. The thought behind it was that natural selection causes covariance between breeding values and fitness, and that this—*regardless of whatever else happens*—mathematically implies a partial change in the mean value of the character under selection. The effect is 'primary' only in so far as it is the most obvious and probably the most powerful way in which selection issues in change; it need not be *temporally* prior to other effects. The partial change attributable to the primary effect of selection may ultimately be cancelled out or augmented by other partial changes, but this does not prevent us separating out this particular partial change and employing it to draw inferences as to what the overall change *would* have been, if all other partial changes had been nullified.

For the special case in which the character of interest is fitness itself, we can identify the primary effect of selection with the covariance term in equation (6.2.1):

$$\Delta_{1^\circ}\bar{w} = \frac{1}{\bar{w}}\left[\mathrm{Var}\left(g_w\right)\right] \qquad \text{(FTNS*)}$$

This result (henceforth: FTNS*) has the surface form of FTNS. Interpreted in the terminology of Chapter 3, it amounts to the claim that the primary effect of natural selection is responsible for the additive genetic variance in fitness[3], and that this effect—regardless of whatever else happens—mathematically implies a partial change in the mean population fitness proportional to this variance.[4] This partial change may be

[3] It therefore neglects the effects of random drift on the covariance term in the Price equation, though drift can be accounted for by extending the Price equation to accommodate stochasticity (Grafen 2000, 2002, 2006). I do not discuss this issue here.

[4] The '$1/\bar{w}$' is included for normalization. We can move it inside the variance term if we want to, yielding the result that the partial change in mean fitness equals the additive genetic variance in *relative* fitness. This explains why Fisher talks of the change in fitness being 'equal to', rather than merely 'proportional to', the additive genetic variance.

cancelled out or augmented by other partial changes, but that does not impugn the correctness of the identification. It follows, significantly, that the partial change in mean fitness implied by the primary effect of natural selection cannot be negative.

6.2.3 FTNS and MAX-B

There can be little doubt as to the correctness of FTNS*, since it is true more or less by definition.[5] But one may question whether it is truly what Fisher had in mind when he formulated his original FTNS. I will not explore this complex exegetical issue here: I will simply consider FTNS* at face value—as a true statement that is formally close to, if not identical to, Fisher's FTNS—and move on to the question of what it actually tells us.[6] For, even though FTNS* is correct, one may still question the *biological significance* of a 'maximization' result that concerns only partial change, as Price and others have done (Price 1972b; Ewens 1989; Okasha 2008).

In particular, one might ask: does FTNS* successfully establish MAX-B? MAX-B, recall, is the thesis that there is some variable such that it *would* never decrease over evolutionary time, if all partial changes other than the partial change attributable to natural selection were absent. Does FTNS* imply MAX-B, where mean population fitness is the variable in question? I am sceptical on this score. The reason is that FTNS* strictly speaking concerns only the *primary* effect of natural selection, while neglecting the *secondary* and *tertiary* effects (cf. Chapter 3). That is, it says nothing about how selection on genetic variation in *transmission biases with respect to* w impacts on the change in $\bar{w}$ (the secondary effect), and it says nothing about how the effects of selection on the *average effects of alleles in the descendant-population* impacts on the change in $\bar{w}$ (the tertiary effect).

---

[5] Subject to the qualification in footnote 3.

[6] Perhaps it would be more apt to refer to FTNS* as the 'Fisher-Price theorem', though this would have the unfortunate side effect of making it sound rather childish.

Both effects are likely to make a difference to the overall direction of evolution at least some of the time. The tertiary effect—change in the average effects of alleles brought about by change in gene frequencies—is undoubtedly extremely important in cases of frequency-dependent selection. Consider again the 'tragedy of the commons' scenario, in which the increase in frequency of an initially advantageous allele beyond a critical threshold leads to a catastrophic collapse in the mean fitness of the population. As we noted above, the collapse is indirectly attributable to selection, because it was caused by changes in gene frequency for which selection was directly responsible. But this collapse in the population mean is not directly attributable to the *primary effect* of selection. What happens instead is that, as the critical threshold is crossed, the average effect of the resource-exploitation allele on fitness collapses between the ancestor- and descendant-populations: an allele that was previously a positive predictor of fitness (because resource-exploitation was advantageous) suddenly becomes—once the resource has all gone—at best neutral with respect to fitness, and possibly disadvantageous. This is a tertiary effect, in the terminology of Chapter 3.

We therefore have good reason to conclude that FTNS* fails to imply MAX-B, because the partial change attributable to natural selection *simpliciter* is *not* equivalent to the partial change solely attributable to the primary effect of natural selection. FTNS* concerns only this latter quantity, and therefore fails to take account of two other potentially important pathways through which natural selection influences the direction of evolution.

Could we reformulate MAX-B so that it *is* implied by FTNS*? A reasonable move here would be to change the wording of MAX-B so that talk of 'the partial change attributable to natural selection' gives way to talk of 'the partial change attributable solely to the *primary effect* of natural selection (and hence not to the secondary or tertiary effects)':

> **MAX-B':** For some property π, the partial change solely attributable to the *primary effect* of natural selection is such that, if it were the only partial change

> relevant to the direction of evolution, $\bar{\pi}$ would never decrease between earlier and later time-slices of the same population.

The drawback to such a move is that, on this proposed rewording, the biological significance of the reformulated principle is obscure: what is the special significance of the primary effect, given that natural selection may also have secondary and tertiary effects? It is not clear that separating out the 'primary effect' of natural selection from the 'secondary' and 'tertiary' effects is anything more than an *ad hoc* attempt to recover something that looks like a maximization principle from the ruins of MAX-B, which already amounted to a significant weakening of MAX-A. It is doubtful whether what is left still constitutes a maximization thesis worthy of the name.

A subtler reformulation of MAX-B (based on Okasha 2008, and indirectly on Price 1972b) replaces talk of 'the partial change attributable to natural selection' with talk of 'the partial change attributable to natural selection *acting in a constant environment*':

> **MAX-B'' *(Partial change)*:** For some property π, the partial change attributable to natural selection *in a constant environment* is such that, if it were the only partial change relevant to the direction of evolution, $\bar{\pi}$ would never decrease between earlier and later time-slices of the same population.

Might FTNS* imply MAX-B''? After all, FTNS* fails to imply MAX-B partly because it ignores what I have called the 'tertiary effect' of selection (i.e., changes in the average effects of alleles on fitness due to changes in gene frequency). Arguably, however, it is legitimate to ignore this 'tertiary effect' for the purposes of evaluating MAX-B''.

The reason for this is that a change in the average effects of alleles arguably requires a *change in the environment* between the ancestor- and descendant-populations (cf. Chapter 3; Price 1972b; Okasha 2008). This need not amount to an environmental change in the

ordinary (i.e., ecological) sense of the word, and indeed it need not be mediated by any change in the fitness of genotypes. It may simply be that genes interact non-additively (i.e., with dominance or epistasis), so that their average effects on fitness depend on the relative frequencies of the different genotypic contexts in which they might find themselves. For instance, in a case of heterozygote advantage, the average effects of the relevant alleles will depend on the frequency of heterozygous individuals. When the frequencies of genotypic contexts change in a way that has a knock-on effect on the average effects of alleles, we can think of this change as a change in the 'genic environment' the alleles experience (Price 1972b; Sterelny and Kitcher 1988; Okasha 2008).

If we accept the premise that the average effects of alleles cannot change without a change in the (ecological or genic) environment, then the tertiary effect of selection would disappear if selection were to take place in a constant environment. This leads to the thought that, while FTNS* may not be sufficient to imply MAX-B, it may be sufficient to imply MAX-B''; and while MAX-B'' is weaker than MAX-B as originally stated, it still appears to constitute a contentful, non-*ad-hoc* maximization thesis. In my view, this line of argument (which is approximately that of Okasha 2008, transposed into the new terminology I have introduced here and in Chapter 3) is almost correct: the difficulty that remains is that FTNS* also neglects the *secondary* effect of natural selection (i.e., in this context, covariance between fitness and individual transmission biases with respect to fitness), and this secondary effect is *not* guaranteed to disappear when we 'hold fixed' the average effects of alleles. In fact, the only sure-fire way to rule out the possibility of covariance between fitness and individual transmission biases with respect to fitness is to assume there is no variance at all in the latter quantity (i.e., $\mathrm{Var}\left(\Delta g_w\right)=0$). FTNS* would imply MAX-B'' in conjunction with this assumption, but the assumption is plainly a substantive one. It might be rendered false by the presence of meiotic drive, gametic selection, or any other variable disruptive to the perfect transmission of genes from parents to offspring.

I conclude, therefore, that FTNS* by itself implies neither MAX-B nor MAX-B''. The strongest result it does imply appears to be MAX-B'. As we noted above, however, this maximization thesis is so hedged that its biological significance remains unclear.

6.2.4 FTNS and individuals as maximizing agents

Let us put these interpretative issues to one side, so as to consider the relationship between FTNS* and the *second* row in our table. For even if we accept that FTNS* does imply MAX-B, it is important to note that FTNS* implies nothing at all regarding what, if anything, *individual organisms* will act if maximizing by means of their own behaviour. FTNS* certainly does not imply MAX-C, because it concerns only the partial change attributable to the primary effect of natural selection, whereas MAX-C concerns the outcome of evolution as a whole. Less obviously, FTNS* equally fails to imply MAX-D, because *FTNS* is insensitive to considerations of control* (cf. Grafen 2002). The claim that the partial change in mean fitness attributable to the primary effect of natural selection equals the additive genetic variance in fitness is quite compatible with an individual's fitness being substantially or even wholly independent of its own behaviour—and it is therefore quite compatible with the idea that organisms do not 'act as if maximizing' anything.

There is a more general point here. For in fact, no result in 20th Century population genetics formally establishes MAX-C or MAX-D, even though principles of this nature are routinely assumed in behavioural ecology (Grafen 2002). In other words, there is no equivalent of FTNS* that even purports to do at the level of individual organisms what the fundamental theorem purports to do at the level of populations: there is nothing that even purports to establish the claim that evolution or natural selection leads to individual organisms behaving as if maximizing anything. Behavioural ecologists often *assume* that organisms act as if maximizing their fitness or inclusive fitness—at least to the extent that their behaviour has been shaped by natural selection—but there is no formal result on which they can legitimately draw in support of this assumption.

This somewhat alarming observation—that behavioural ecologists typically assume the validity of an individual-as-maximizing-agent analogy without any formal justification for doing so—is part of the motivation for the 'Formal Darwinism' project, pursued in recent years by the Oxford geneticist Alan Grafen and colleagues. The goal of the project, in Grafen's terms, is to provide a 'secure logical foundation for the commonplace biological principle that natural selection leads organisms to act as if maximizing their 'fitness'' (Grafen 2002, 75). While Grafen does not employ the terminology introduced here, we can plausibly interpret his project as an attempt to vindicate a version of MAX-C or MAX-D[7], with personal fitness or (in cases of social behaviour) inclusive fitness occupying the role of maximand.[8] In a substantial series of papers, Grafen and colleagues have sought to achieve this end by proving formal analogies between evolutionary models and optimization programmes (see Grafen 1999, 2000, 2002, 2003, 2006a,b, 2007a,b,c, 2008, 2009; Gardner and Grafen 2009; Gardner and Welch 2011). The nature of these analogies is explored in the subsequent sections , along with the question of what they actually show.

6.3 The 'Formal Darwinism' project

This section provides a brief overview of Grafen's project and the results he has obtained so far. This will prepare us for the next section, which subjects the project to philosophical scrutiny. I should note at the outset that, although I will present Grafen's results and proofs verbally, Grafen himself presents them within a dauntingly complex formalism.

---

[7] I revisit the question of which type of maximization thesis he is trying to establish in Section 6.4.

[8] Grafen (2002) argues that personal fitness is the appropriate maximand only when social behaviour is absent. Inclusive fitness is the more general maximand, since it is maximized in cases of social behaviour in which individuals do not maximize their personal fitness. Hence Grafen (2006a) states in a later paper that his aim is to provide a 'fully explicit argument ... that broadly supports a widespread belief among whole-organism biologists that natural selection tends to lead to organisms acting as if maximizing their *inclusive* fitness' (2006a, 541; my italics). I discuss Grafen's extension of his argument to social behaviour in Section 6.3.4.

This high degree of formality is understandable: after all, the whole point of the 'Formal Darwinism' project is to give expression to (and provide a vindication of) an 'individual-as-maximizing-agent' analogy *without* falling back (as previous authors have done) on purely verbal arguments. The downside to this laudable rigour, however, is that it renders the arguments all but unintelligible without very careful reading. An informal précis of Formal Darwinism may sound like a contradiction in terms—and it is certainly no substitute for the real thing—but it will do for our purposes.

6.3.1 Ingredients

The Price equation for p-scores

The goal of Formal Darwinism is to forge links between population genetics and optimization programmes. The 'population genetics' half is provided by the Price equation, formulated in terms of *p*-scores:[9]

$$\Delta\bar{p} = \frac{1}{\bar{w}}\left[\mathrm{Cov}\left(w,p\right)+\mathrm{E}\left(w\Delta p\right)\right] \tag{3.2.2}$$

As discussed in Chapter 3, the '*p*-scores' version of the genetic Price equation has merits and demerits in comparison to the more commonly seen 'breeding values' version. The major drawback is that the '*p*-scores' version, when taken in isolation, tells us nothing about *phenotypic* evolution; it does so only if conjoined with some auxiliary result showing that that the direction of evolution in some particular phenotype will reliably equal the direction of evolution in some particular *p*-score. Accordingly Grafen, in order to derive results concerning phenotypic change, introduces a *p*-score-to-phenotype mapping function separate from the Price equation itself. By contrast, the definition of 'breeding value' implies that the change in the population mean for a phenotypic character *always* equals the change in population mean for the corresponding breeding value, befitting the

[9] In order to separate out the effects of random drift from those of non-random evolutionary processes, Grafen (2000) reformulates the Price equation in terms of *expected* change. Though this move is important, I omit discussion of it here, because it is tangential to the arguments I want to make.

theoretical role of breeding values as a formal 'bridge' between a phenotype and its transmissible basis.

The main advantage to using *p*-scores is that the notion is more minimal and consequently more general than that of a breeding value. A breeding value is a set of allelic values aggregated using a specific set of weights, namely the Fisherian average effects (i.e., partial regression coefficients) of allelic predictors on the phenotype of interest. A *p*-score, by contrast, can be any quantity 'that an offspring inherits by averaging together the gametic contributions of its parents' (Grafen 1985a, 33)[10]**.** There is thus no *formal* requirement that a *p*-score is interpretable as a weighted sum of allelic values, even though such weighted sums no doubt constitute an important subset of *p*-scores.[11] The notion is in fact so broad that we could, in principle, define a formally permissible *p*-score simply by assigning numbers to individuals arbitrarily and equating an individual's *p*-score with the number it has been assigned, subject to the sole constraint that the number assigned to each individual must equal the average of the numbers assigned to its parents. Grafen explicitly allows for such 'hypothetical *p*-scores, in which we assign a number to each individual, without asking whether there is an actual set of allelic weights that does produce that set of numbers' (2006a, 553). Indeed, as we will see presently, this extremely minimal conception of a *p*-score turns out to be indispensable to some of his formal arguments.

Grafen assumes unbiased transmission of *p*-scores from ancestors to descendants, allowing him to work with the covariance term of the Price equation only:

---

[10] An offspring's breeding value is unlikely to be an average of its parents' breeding values if the average effects of alleles change across generations. Hence I argued in Chapter 3 that, strictly speaking, we should not regard breeding values as a subset of *p*-scores unless the constancy of average allelic effects is assumed.

[11] As noted in the preceding footnote, breeding values fall into this subset only if average allelic effects are constant across generations.

$$\Delta\bar{p} = \frac{1}{\bar{w}}\left[\mathrm{Cov}\left(w,p\right)\right]$$

One might well ask: what does it mean to assume 'unbiased transmission of *p*-scores', given that the notion of a *p*-score is so broad and so abstract? Grafen (2002, 2006a) suggests that this assumption is guaranteed by fair meiosis and the absence of gametic selection and mutation.[12] This would be correct if all *p*-scores were interpretable as linear combinations of allelic values with constant weights, but we have already seen that *p*-scores are *not* constrained to be thus interpretable. I suggest, therefore, that Grafen is too quick here: he takes for granted that the assumption of unbiased transmission of *p*-scores can be glossed in genetic terms, despite acknowledging that *p*-scores themselves need not admit of a biological interpretation. In practice, it is quite unclear what biological conditions, if any, would ensure the unbiased transmission of *any possible* *p*-score, including those without a biological interpretation.[13]

Optimization programmes

The Price equation is, in essence, a means of representing formally a cluster of fairly intuitive ideas about the evolutionary process: (i) that evolutionary change is change in population means of properties, (ii) that it depends on selection and transmission of those properties, (iii) that selection involves covariance between a property and fitness, and (iv) that biased transmission involves a change in the value of the property between ancestors and descendants. By expressing these truisms in a single, rigorous statement, the equation shows us precisely *how* the evolutionary change in some character depends on its

---

[12] The question of whether it is reasonable to ignore the effects of mutation in an attempt to formalize Darwinism will be considered in Section 6.4.

[13] The conjunction of fair meiosis and no gametic selection is also insufficient for the unbiased transmission of *breeding values*. This additionally requires an assumption of the constancy of average effects between the ancestor- and descendant-populations (cf. Chapter 3). So switching from *p*-scores to breeding values would not make Grafen's claim correct.

covariance with fitness and on its pattern of transmission. As far as 'Formal Darwinism' is concerned, however, this is only half the story. In order to forge formal links between concepts of selection and concepts of optimization, we also need some comparable means of formalizing our intuitive ideas about the latter. This, Grafen (1999, 2002, 2006a) suggests, is best accomplished by borrowing from economics the formal apparatus of an *optimization programme* (Mas-Colell et al. 1995).

The two central concepts that define an optimization programme are the *strategy set* and the *maximand*. The 'strategy set', X, is a set each member of which corresponds to a distinct possible phenotype. If we are interested in behaviour, these phenotypes will usually be strategies (on the definition of strategy outlined in Chapter 2), but the phenotypes need not admit of a behavioural interpretation. In principle, the set of possible phenotypes can be as inclusive as we like. Grafen's formalism is deliberately permissive on this score, since his aim is to prove links that hold regardless of the choice of X. One consequence of this is that X could contain phenotypes very far removed in 'phenotypic space' from the phenotypes actually realized in the population: if we wanted to, we could include pigs with wings, dolphins with arms, and so on. As we will see, Grafen's liberal conception of a strategy set combines with his liberal conception of a *p*-score to yield some curious results.

The 'maximand', π, is formally defined as a function that maps X on to $\mathbb{R}$, the set of real numbers. Ideally, of course, we would like a maximand that is *biologically interpretable*, in the sense that its value can be identified with the value of a property that individual organisms actually possess. This maximand-property could in principle be any biological property such that (i) we can assign a real-numbered value to every individual in the population and (ii) an individual's value for the property functionally depends on its phenotype. Many properties could satisfy these minimal criteria, and we could formally define an optimization programme for any of them; but, of course, there would be no *a priori* guarantee that organisms would solve or even approximately solve any of these optimization programmes. The challenge is to find a biological maximand such that the

corresponding optimization programme *is* actually solved (or at least approximately solved) by organisms whose behaviour has been shaped by natural selection.

In the abstract, an optimization programme takes the following form:[14]

$$x \max\left[\pi\left(x\right)\right], x \in X \tag{6.3.1}$$

A solution to the optimization programme, x^*, is any phenotype that satisfies the two constraints embodied in (6.3.1): it is a member of X and, among the members of X, it maximizes (i.e., maps on to the highest value of) π. Ideally, we would like to find a biological maximand such that individual organisms reliably solve (or more or less solve) the corresponding optimization programme.

'Scope for selection' and 'potential for positive selection'

The Price equation for *p*-scores formally captures our intuitive ideas about evolution by natural selection, while the notion of an 'optimization programme' formally captures our intuitive ideas about the kind of 'optimization' or 'maximization' that is exhibited in the behaviour of individual organisms. Grafen's strategy is to prove material conditionals linking claims about natural selection (formulated in terms of the Price equation) to claims about optimization (formulated in terms of an optimization programme). The links themselves are formulated in terms of two crucial concepts, which may thus be thought of as 'bridging' notions. These are 'scope for selection', and 'potential for positive selection in relation to X'.

[14] Grafen (2002) adds several layers of complexity to his optimization programme that we do not need to consider here; (6.3.1) adequately captures the basic idea. Grafen also works with a more complex version of the Price equation; see footnote 9.

'Scope for selection' is informally defined as follows:

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

The word *possible* is significant here. We noted above that, although the notion of a *p*-score was originally introduced by Grafen (1985a) in part as a means of aggregating an organism's allelic values into a single number, the only *formal* requirement on a *p*-score is that an offspring's value is an average of its parents' gametic contributions. There is thus no formal requirement that a *p*-score can be interpreted as a weighted sum of allelic values: *p*-scores may be merely 'hypothetical', in the sense that they could be expressed for each individual as a (consistently) weighted sum of its allelic values only if the allelic composition of the population were different from its actual composition (Grafen 2006a, 553).

Crucially, these merely 'hypothetical' (i.e., biologically non-interpretable) *p*-scores still count for the purposes of determining whether there is 'scope for selection' in Grafen's sense of the term. The implication is that there can be 'scope for selection' in a population even when no actual change in allele frequencies takes place. Indeed, if a population contains fitness differences then there is *guaranteed* to be some formally permissible *p*-score that co-varies with fitness in that population, irrespective of whether fitness co-varies with any biologically interpretable *p*-score (Okasha and Paternotte 2012). It follows that there is technically 'scope for selection' whenever a population contains fitness differences.

The second bridging notion is that of 'potential for positive selection in relation to *X*'. Unlike 'scope for selection', this notion explicitly blends the language of population

genetics ('positive selection') with the language of optimization programmes ('in relation to *X*'). Informally, it is defined as follows:

> ***'Potential for positive selection in relation to X':*** There is a phenotype in *X* that would have been favoured by selection had that phenotype been present.

Again, the strikingly liberal nature of this definition is worthy of comment. As we have already noted, the set of possible phenotypes, *X*, can be as inclusive as we like: there are no 'plausibility constraints' built into the formalism. The upshot is that 'potential for positive selection in relation to *X*' does not imply that there is *serious* potential for positive selection, where the qualifier 'serious' indicates that the trait that would be selected if it were to arise has a non-negligible probability of actually arising.

6.3.2 Formal links

With the necessary conceptual and formal machinery in place, it is relatively straightforward to prove links between the Price equation and an optimization programme in which *individual fitness* is the maximand, conditional on a few substantive assumptions (discussed below). Grafen (2002) proves four such links:

> **LINK 1:** If each individual acts optimally, then there is no scope for selection and no potential for positive selection.

> **LINK 2:** If each individual acts sub-optimally, but equally so, then there is no scope for selection but there is potential for positive selection.

> **LINK 3:** If individuals vary in the value of the maximand they attain, then there is scope for selection, and the change in every gene frequency and in the additive genetic value of every character equals its covariance across individuals with the value of the maximand.

> **LINK 4:** If there is no scope for selection and no potential for positive selection, then each individual in the population acts optimally.

The links are certainly plausible at face value; indeed, they may appear at first sight to be little more than banal truisms about natural selection. Yet, in Grafen's view at least, the conjunction of these four links implies something far from banal, for the links jointly provide 'a secure logical foundation for the commonplace biological principle that natural selection leads organisms to act as if maximizing their "fitness"' (Grafen 2002, 75). In other words, Grafen's view appears to be that the conjunction of Links 1-4 implies something close to our MAX-C or MAX-D; I return in Section 6.4 to the question of whether the links imply any such thing.

6.3.3 Assumptions required for the links to obtain

The idea that the four links are indeed contentful and revealing (rather than trivial and unilluminating) gains some support from the fact that substantive assumptions are required if they are to hold. One important assumption is that all individuals face the same optimization programme: that is, the strategy set X and the mapping function $\pi\left(x\right)$ are the same for every organism, so that if we were to swap the phenotypes of any pair of organisms we would also swap their attained values of the maximand. Grafen introduces this assumption under the name 'pairwise exchangeability'. If the 'maximand' in question is individual fitness, it amounts to the assumption that the (expected) fitness of every organism functionally depends on its phenotype in the same way; this in turn requires that there are no expected fitness-relevant differences in the local environments organisms may occupy in their lifetime. In essence, we are assuming that 'all individuals face the same environmental challenges, and so are having to solve the same problems' (Grafen 2002, 79).

It is debatable whether this is a reasonable assumption to make of real populations; its reasonableness ultimately turns, I think, on how populations are to be individuated. There is no doubt that organisms of the same species may face quite different environmental challenges, depending on the local circumstances in which they find themselves. The challenges a fox faces in a forest, for instance, are not the same as those faced by its urban conspecifics. If we count rural and urban foxes as members of the same population even though they occupy qualitatively different environments, then pairwise exchangeability will not be a reasonable assumption for this population. One might object, however, that if organisms supposedly within the same population *predictably* face very different environmental challenges, then we have individuated populations at too coarse a grain. If a population can be subdivided into subpopulations, each of which faces its own characteristic set of environmental challenges, then these subpopulations deserve to be regarded as populations in their own right: *they* (not the overarching meta-population) are the appropriate targets for evolutionary analysis in the first instance. One might consequently be tempted to recast the assumption of 'pairwise exchangeability' (i.e., the absence of predictable environmental differences significant enough to matter to the phenotype-fitness mapping function) as formally capturing part of what it *is* for a population to be paradigmatically Darwinian (cf. the '*S*' parameter in Godfrey-Smith 2009).

More seriously, Grafen's (2002) arguments also assume the absence of frequency-dependent selection and of social interactions. The rationale is broadly the same in both cases: if individual fitness is to be an appropriate maximand, an organism must be able to *control* its fitness by means of its own behaviour. The relevant sense of 'control' here is that outlined in Chapter 2: an organism's fitness must exhibit systematic counterfactual dependence on its phenotype, allowing a function-like mapping from phenotypic states to attained maximand values in a given environment. This sort of control is central to the very idea of an optimization programme. Yet frequency-dependence and social interaction both generate problems in this regard, because both introduce ways in which an

organism's fitness in a given environment can depend on something other than its own phenotype. Grafen thus assumes their absence. Of course, there can be no doubt that both phenomena are very often present in real populations, so the necessity of assuming their absence severely limits the generality of Grafen's (2002) arguments.

6.3.4 Allowing social behaviour

We have seen that, when the putative 'maximand' is individual fitness, Grafen's links hold only when social behaviour is absent. The reason is that, when social behaviour is present, there is a component of an individual's fitness over which it has no control, and which it is therefore unable to 'maximize' by means of its own behaviour. If we want to prove Grafen's links in populations where social behaviour is present, we need to find a different maximand: a maximand over which a social actor *does* have full control even when its personal (neighbour-modulated) fitness is partially under the control of others. This is where inclusive fitness (a sum of the fitness components a social actor controls, weighted by the transmission fidelity of its genes through each component) comes into its own (Chapter 5). In a more recent paper, Grafen (2006a) proves that his four links do indeed obtain in cases of social interaction when inclusive fitness is taken as the putative maximand.

In the special case in which social interactions are absent, an individual's inclusive fitness is equal to its personal fitness, so we recover the non-social version of Grafen's links as a special case of the social version. This suggests that, to the extent that individual organisms can be said to maximize anything *in general*, the quantity that they maximize is their *inclusive* fitness, not their personal fitness. Substantive assumptions are still required for these arguments to go through, however. 'Pairwise exchangeability' is still invoked (albeit under the new name of 'universal strategic equivalence'): that is, we must still assume that all individuals face the same optimization programme. Moreover, the social interactions must satisfy two further important assumptions. The first of these, 'actor's control', amounts to the assumption that an actor fully controls (in the sense of Chapter 2)

the sign and magnitude of the fitness effects it confers on other individuals. In Grafen's terms, 'the nature and quantitative effects of one individual's action *depend only on the phenotype of that individual*, and not, for example, on some capacity of the recipient to use the help provided' (Grafen 2006a, 543, my italics). This is a strong assumption, since it implies the absence of synergistic interaction (see Chapter 4) or any other source of frequency-dependent fitness effects.

The second assumption is 'additivity', which in this context has a special meaning: it does not denote dominance or epistasis, nor does it denote synergistic interaction (the absence of synergistic interaction having already been entailed by the assumption of actor's control). It refers rather to the idea that an individual's lifetime personal fitness can be equated with a baseline, non-social component plus a *sum* of all the social fitness effects it has experienced during its lifetime. That is, the social effects an individual experiences must *add up* to yield a measure of the social component of its lifetime fitness, ruling out multiplicative effects such as diminishing returns from repeated interactions. This amounts to a further assumption above and beyond actor's control, for even if an actor fully controls the sign and magnitude of the fitness effects it confers, it might still be that these effects combine multiplicatively with other fitness components to yield the lifetime personal fitness of the recipient.

6.3.5 Allowing frequency-dependence

Grafen (2006a) does not relax the assumption of frequency-independence: as noted above, it is implicit in his assumption of actor's control. Andy Gardner and J. J. Welch (2011), in a paper primarily aimed at extending the 'individual-as-maximizing-agent' analogy at the level of individual genes (an issue I do not discuss here), suggest one way in which the assumption of frequency-independence might be relaxed. Their suggestion, in informal terms, is that we regard the actual phenotypic composition in a population as part of the *environment* that all agents share. By making this move in cases of frequency-dependence, we can recover a function-like mapping from phenotypic states to attained maximand

values *in a given environment*, because trait frequencies are now considered to be part of this environment. Note the close affinity between this proposal and the thought we encountered in Section 6.2 that changes in the average effects of alleles due to changes in gene frequency can be attributed to 'changes in the genic environment'.

One oddity of the Gardner and Welch approach is that, in order to evaluate the hypothetical value of the maximand that would be attained by a non-actual phenotype, we are supposed to assess its value for the maximand relative to the *actual* composition of the population. That is, we need to assess its value for the maximand *relative to a distribution of phenotypes in which its frequency is zero*. It is unclear how we should go about this: after all, it is one thing to ask what the inclusive fitness of a winged pig would be, if it existed; but it is quite another thing to ask what the inclusive fitness of a winged pig would be, if it existed in a population in which there are no winged pigs. Gardner and Welch (2011) suggest that we should imagine introducing a non-actual phenotype at a vanishingly small frequency, so that its introduction makes no significant difference to the composition of the population, and assess the value of the maximand it would attain in these circumstances. If the actual population is finite, however, it is something of a leap of faith to suppose that it is always possible to introduce a new phenotype in this way. In some cases, it could be that introducing even a *single individual* would alter the population's composition in a way that makes a significant difference to the optimization programme. Gardner and Welch might reply that such cases will be rare and can probably be ignored for most practical purposes. But appealing to pragmatic considerations seems questionable in this context, given the nature of the Formal Darwinism project. The goal, after all, is not to develop a framework that is instrumentally useful for pragmatic ends, but rather to construct a 'secure logical foundation' for the intuitive connection between concepts of selection and concepts of optimization.

6.4 What do the links actually show?

I now turn to the question of the biological significance of Grafen's links. First of all, let us consider how the links relate the MAX-C/MAX-D distinction outlined in Section 6.1. An analogy with Fisher's fundamental theorem may prove useful in this context. A common gut-reaction to FTNS is that it simply cannot be true because, if it were true, it would show too much. After all, the theorem appears to show that the mean fitness of an evolving population cannot decrease, when we know full well that it can decrease. In Section 6.3, we encountered Price's (1972b) deflationary response to this worry: on his interpretation, the theorem implies nothing at all about the overall direction of evolution, because it concerns only the partial change attributable to natural selection in a constant environment.

At first, one may react in a similar way to Grafen's four links. If they are supposed to show that *evolution* reliably leads to organisms acting as maximizing their fitness, then they surely cannot have succeeded in doing so, because it is absurd to think we could ever have such assurances about the overall outcomes of evolution. Evolution is, after all, an historical process, and its the long-run trajectory is sensitive to all manner of chance events (Gould 2002; Beatty and Desjardins 2009). But a version of Price's deflationary response applies here too: Grafen should not be read as even attempting to prove results regarding the *overall* outcomes of evolution. He is not purporting to have established MAX-C. Instead, much like Fisher, he can more plausibly and more charitably be seen as attempting to prove results regarding the *partial change* attributable to natural selection in a constant environment. When Grafen talks of vindicating the idea that 'natural selection leads organisms to act as if (more or less) maximizing … fitness', I suspect that his use of the phrase *natural selection* in place of *evolution* is quite deliberate, and that we should interpret the statement as a thesis about partial change, not total change.

On this reading, the goal of the project is neither a population-level maximization thesis nor an unqualified vindication of the 'individual-as-maximizing-agent' analogy. The goal is rather a *qualified* vindication of such an analogy as a reliable heuristic for predicting the phenotypes that *would* be produced by the partial change caused by natural selection in a constant environment, if this were the *only* influence on the direction of evolution. Naturally, there will then be a further empirical question as to how often these hypothetical optimal outcomes are attained by actual evolving populations, given that evolutionary processes other than natural selection may impede the maximization of inclusive fitness. Settling this empirical question, however, is a task beyond the scope of the project, since no amount of purely formal work could tell us the relative importance of different evolutionary processes in the actual world.

Accordingly, I suggest that, to the extent that Grafen's project is intended to vindicate the idea that fitness or inclusive fitness is 'maximized' in a process of natural selection, the relevant sense of 'maximization' is that captured by MAX-D, the bottom-right cell in our table. Yet I am sceptical as to whether the four formal links Grafen proves are sufficient to establish MAX-D—a thesis that, for all its qualifications, undoubtedly remains ambitious and controversial. In the rest of this section, I offer two reasons for this scepticism.

6.4.1 The surprisingly weak nature of the links

There are well-known cases in which natural selection does not lead to optimal distributions of phenotypes, even if the ecological environment is unchanging and all other evolutionary processes are presumed absent. Among them are cases of *heterozygote advantage*, including the famous case of sickle-cell anaemia and malarial resistance. The population at equilibrium contains an allele (S) which causes sickle-cell anaemia when present in a homozygote (SS), but which confers resistance against malaria without adverse side effects when present in a heterozygote (SN) with the 'normal' allele (N). The easiest way to see informally why the equilibrium is polymorphic and suboptimal (i.e., contains a significant number of individuals with sickle-cell anaemia) is to imagine what

would happen if the population were composed entirely of heterozygotes (100% *SN*), rendering it both monomorphic and optimal in its genotypic composition. This optimal state would be fleeting, since, under fair meiosis, homozygous individuals would be virtually certain to arise in the next generation. The *NN* homozygotes would then outperform the *SS* homozygotes, leading the overall frequency of the *S* allele to drop below 50%. The *N* allele would not go to all the way to fixation, however, because at low frequencies an *S* allele is much less likely to find itself in a homozygote, and so much more likely to confer a benefit on its bearer. The true (polymorphic) equilibrium lies between these unstable extremes, and depends on the size of the advantage the heterozygote confers (which in turn depends on the local incidence rate of malaria).

What happens to Grafen's links in such a case? One might initially suppose that such a scenario must constitute a counterexample to at least one of them. For can it really be true in a case of heterozygote advantage that, 'if each individual acts optimally, then there is no scope for selection and no potential for positive selection' (Link 1)? A moment ago, we pictured a monomorphic population entirely composed of heterozygous (*SN*) individuals carrying a sickle-cell anaemia allele. In this scenario, each individual is phenotypically optimal,[15] because all individuals are *SN*. Yet it seems intuitively as though there *is* scope for selection, since the monomorphic distribution is not an equilibrium: *N* will spread at the expense of *S* as soon as homozygous *SS* genotypes arise. And can it really be true that 'if individuals vary in the value of the maximand they attain, then there is scope for selection, and the change in every gene frequency and in the additive genetic value of every character equals its covariance across individuals with the value of the maximand' (Link 3)? After all, at the polymorphic equilibrium (in which *SN*, *NN* and *SS* genotypes all coexist) individuals clearly vary in the value of the maximand they obtain (i.e., their fitness), and yet no selection occurs.

[15] In relation to the relevant set of phenotypes: {malarial resistance, sickle-cell anaemia, neither}.

In fact, Grafen's links *do* hold in cases of heterozygote advantage, but only owing to the somewhat abstruse technical meaning he assigns to the phrases 'potential for positive selection' and 'scope for selection' (Grafen 2002; Grafen 2007a; Okasha and Paternotte 2012). To see why, let us first consider the monomorphic distribution in which all individuals are *SN*. In this population, there is technically no 'scope for selection', because there is no variation in fitness. This is true even though the current distribution of genotypes is highly unstable, and even though selection will kick into action as soon as homozygous individuals arise. And there is technically 'no potential for positive selection', because (since all individuals are phenotypically optimal) there is no alternative *phenotype* that would be favoured by selection over malarial resistance were it to arise. This is true even though, if a mutant allele (*S**) were to arise that caused malarial resistance in heterozygous individuals *without* causing sickle-cell anaemia in homozygous individuals, it would spread in the long run at the expense of both *N* and *S*. The result is that Link 1 holds in this population. Link 4 also holds substantively, while Links 2 and 3 hold vacuously, since their antecedents are false (and, in classical logic, a material conditional with a false antecedent is true).

Now let us consider the actual polymorphic equilibrium. Here too, it turns out that Grafen's links do hold. Links 1, 2 and 4 hold vacuously (Grafen 2007a). Link 3, meanwhile, holds substantively, even though the population is at equilibrium (Grafen 2007a). For even though the equilibrium is a stable one—i.e., the gene frequencies do not change over time—there is nonetheless 'scope for selection' *sensu Grafen*. After all, the population contains fitness differences, and this implies that there exists a *possible p*-score that co-varies with fitness (Okasha and Paternotte 2012). The reason we do not observe any *actual* changes in gene frequency is that the *p*-scores that co-vary with fitness cannot be expressed as linear combinations of the *S* and *N* alleles, and so lack any meaningful biological interpretation.

Grafen's links, then, hold after all in cases of heterozygote advantage, despite appearances to the contrary. What should we make of this? Grafen (2007a) suggests that we should see this as a 'reassuring' feature of his formal framework: if the links turn out to hold even in cases when it seems they must fail, this shows that they embody impressively general truths about the evolutionary process. I would draw a rather different moral. What the case of heterozygote advantage shows, I suggest, is that Grafen's links are surprisingly weak—considerably weaker than they initially appear to be.[16] In particular, Link 1 may appear to imply that, if all individuals are phenotypically optimal, then this optimal distribution of phenotypes will be retained in subsequent generations. The case of heterozygote advantage shows that Link 1 implies no such thing: a monomorphic population of *SN* individuals is phenotypically optimal and yet highly unstable. Link 3, meanwhile, may appear to imply that, if individuals vary in their optimality, then selection will alter actual gene frequencies. Again, however, the case of heterozygote advantage shows that Link 3 has no such implication. As long as there are fitness differences in the population there will be 'selection' on some *possible* *p*-scores, but these *p*-scores need not have any meaningful biological interpretation.

I contend that, as a consequence, the links are in fact *too* weak to imply any interesting maximization thesis. In particular, the links surely do not establish what Grafen sets out to prove: viz. that 'natural selection tends to lead to organisms acting as if maximizing their inclusive fitness' (2006a, 541). For there is no escaping the fact that, in cases of heterozygote advantage, selection simply cannot be relied upon to produce optimal *or even approximately optimal* distributions of phenotypes: malarial resistance would not go to fixation and sickle-anaemia would not be eradicated even if all other influences on the

[16] Grafen openly admits that 'the results [i.e., links] do not imply that the outcome of natural selection is inevitably that each individual in the population has an optimal phenotype' (2002, 90). I am arguing that the links are still surprisingly weak, even once we concede that they do not entail the 'inevitable' attainment of optimality.

direction of evolution were absent. If Grafen's links failed in such cases, we might have been led to conclude that they support a maximization thesis that holds 'ceteris paribus' but that happens to fail in cases of heterozygote advantage. Yet because the links *do* hold in such cases, we cannot say even this; the only conclusion we can draw is that the links do not imply a maximization thesis at all. The surprising generality of Grafen's links therefore comes at a heavy cost: they turn out to be too weak to imply the claim he sets out to defend.

One might object here that I have been too quick to endorse the premise that 'selection simply cannot be relied upon to produce optimal or even approximately distributions of phenotypes' in cases of heterozygote advantage. For one might argue that, *given sufficient time*, an allele would eventually arise that produced the same beneficial effect as the *S* allele without the need for heterozygosity, and this allele would spread by selection at *S*'s expense, thereby removing the barrier to optimality imposed by heterozygote advantage (Grafen 1984). If this is correct, then perhaps selection *would* reliably produce optimality in the long run. Note, however, that this reply invokes an empirical assumption about the malleability of the genetic architecture underlying the relevant phenotypes—an assumption that goes well beyond anything contained in Grafen's four, purely formal links. The consequence is that even if we grant that this reply is correct—and thereby concede that some form of MAX-D-type maximization thesis may be legitimate after all in cases of heterozygote advantage—we should still be sceptical of the idea that any such thesis is implied by Grafen's links alone. These links are completely insensitive to considerations about the long-run malleability of genetic systems: they would still hold in a case of heterozygote advantage even if the genetic architecture could not be changed and optimality could never be achieved or approximated.

I conclude, then, that Grafen's four links fall short of implying that 'natural selection tends to lead to organisms acting as if maximizing their inclusive fitness'. Cases of heterozygote advantage show that, even when all four links obtain and all other evolutionary processes

are absent, selection cannot be relied upon to produce organisms that even approximately solve the applicable optimization programme. Indeed, I cannot see any interesting maximization principle that the links could establish, given that all four hold even in populations stuck at suboptimal equilibria. There is, after all, no limit in principle to just how suboptimal such equilibria might be.

6.4.2 A concern about timescale

My second concern is less technical, but no less serious. It concerns the *timescale* with which Grafen's formalism is concerned. As the cases discussed above make plain, Grafen's links connect optimization programmes to *short-term* evolutionary dynamics: indeed, in the discrete-generations models Grafen employs in his (2002, 2006a) papers, the links concern the partial change attributable to natural selection between *two consecutive generations* only. For instance, Link 1 tells us that if all individuals are behaving optimally, then there is no scope for selection and no potential for positive selection; but, as we have seen, the relevant notions of 'scope' and 'potential' both pertain to *immediate* gene frequency change between the initial generation and the following generation. As a result, they imply little about the long-run evolutionary dynamics. Indeed, in the hypothetical case of a 100% heterozygous (*SN*) population, Link 1 technically holds even though selection is sure to occur as soon as a homozygote arises, implying that the currently optimal distribution of phenotypes is highly unstable in the long run.

One may reasonably question whether this is the correct timescale on which to think about the 'individual-as-maximizing-agent' analogy. The analogy, recall, concerns the *long-run outcome* of natural selection: the idea is that, in the absence of countervailing influences on the direction of evolution, natural selection will cause organisms to *act as if maximizing* their fitness or inclusive fitness. This 'as if maximization' (which occurs at the level of the individual organism, in contrast to the form of population-level maximization implied by FTNS*) refers to a characteristic product of natural selection, not to a characteristic of the process itself. Assuming, then, that the operation of natural selection does tend to produce

'as if maximization' on the part of individual organisms, on roughly what timescale is this process likely to take place? Truly striking examples of 'as if maximization' (be they in social insects, mammals or microbes) are products of *cumulative* adaptive evolution, in which natural selection favours incremental improvements on existing structures over an extended time period. At each stage in the process, the fixation in the population of the previous improvement creates new opportunities for a further incremental improvement to arise, allowing the population to gradually ascend 'Mount Improbable' (Dawkins 1996). Godfrey-Smith and Wilkins (2009) suggest the name 'mesoevolution' for this gradual cumulative change, since a contrast with 'microevolution' seems appropriate (and the term 'macroevolution' is usually reserved for very long-term changes in organismal bodyplans; see Gould 2002, Sterelny 2007).

Formal theories of this 'mesoevolutionary' process are conspicuously absent from mainstream population genetics. The result is a curious gulf between informal presentations of Darwinism, which tend to give pride of place to the creative role played by natural selection in cumulatively assembling adaptive complexity over protracted timescales[17], and the formal models of neo-Darwinian population genetics, which tend to concern only short-term gene frequency change. Grafen is to be applauded for noticing this gap between informal and formal presentations of Darwinism, and for endeavouring to bridge it. Crucially, however, he attempts to do so *without* shifting his theoretical gaze from short-term microevolution to gradual cumulative change. His four links, whatever they may tell us about the former, tell us little if anything about the latter. Indeed, the models in which he derives the links assume the absence of mutation, implying that gradual cumulative change is *impossible* in these models!

This focus on short timescales would be unproblematic if the goal of the 'Formal Darwinism' project were merely to vindicate a version of MAX-A or MAX-B: that is, to

[17] See especially Dawkins 1986, 1996; Dennett 1995.

vindicate a claim about the directionality of the short-term total or partial change in mean fitness. This, however, was Fisher's goal; it is not Grafen's. The goal of Grafen's project is that of explaining a particular form of adaptive complexity, viz., the 'as-if maximization' exhibited in the behavioural strategies of individual organisms. The question of whether we should expect natural selection to produce such 'as-if maximization' is not a question about short-term microevolution, and is therefore not a question that any number of formal results concerning the directionality of short-term microevolution could establish. It is a question regarding the complex behavioural strategies we should expect natural selection to assemble *cumulatively* over extended timescales. As such, it is a question that we cannot address while ignoring the processes—notably the process of mutation—that are responsible for the generation of incremental improvements on existing structures.

With this in mind, I suggest that, if we truly hope to vindicate the notion that natural selection leads to organisms that act as if (more or less) maximizing their inclusive fitness, what stands most urgently in need of formal vindication is the informal thought that, because the fixation of an incremental improvement on some trait generates a large number of sites at which further incremental improvement can arise, natural selection makes the evolution of complex adaptations much more likely than it otherwise would be (Dawkins 1986, 1996; Dennett 1995; Neander 1996; Godfrey-Smith and Wilkins 2008; Godfrey-Smith 2009). This is the essence of 'Darwin's dangerous idea', and it is routinely (and plausibly) assumed in informal theorizing about the likely outcomes of mesoevolution. Yet it also remains—to my knowledge—mathematically unproven, by Grafen or anyone else.

In raising this concern about timescales, I do not mean to imply that Grafen's formal links are irrelevant or unimportant to the project of formalizing Darwinism. An expanded 'Formal Darwinism' project aimed at underwriting the 'individual-as-maximizing agent' analogy by capturing the creative power of cumulative selection over protracted timescales might well employ optimization programmes as a means of specifying whether

or not some new variant constitutes an 'improvement' on an existing structure. But the project would not focus *solely* on forging links between these programmes and models of short-term microevolution. Moreover, it certainly could not afford to assume, as Grafen's models do, the absence of mutation. For mutation, being the source of each incremental improvement, is an indispensable component of the process.

SEVEN

Conclusion

The six chapters of this dissertation have covered considerable ground. We have examined conceptual classifications of social behaviour, formal representations of natural selection, alternative versions of Hamilton's rule, neighbour-modulated and inclusive fitness, and arguments for inclusive fitness maximization. The result, I hope, is a comprehensive examination of the conceptual and theoretical foundations of contemporary kin selection theory. I will close by highlighting several themes that have recurred throughout the chapters, and that embody the main morals I want to draw from the discussion.

Qualified pluralism

This has been, in essence, a dissertation about representations of social evolution. Philosophical and theoretical disputes in sociobiology often arise because alternative representations of the same evolutionary phenomena are available, and it is often unclear whether these alternative representations are compatible with one another. Of course, if they turn out to be incompatible, there is a further question as to which is correct. Even if they turn out to be compatible, however, there is still a further question as to which representation we ought to prefer for explanatory purposes, given the choice.

The three central chapters of this dissertation (3, 4 and 5) all confronted issues of this general form. In Chapter 3 (Section 3.2), I compared and contrasted genetic and phenotypic formulations of the Price equation for evolutionary change. Later in the same chapter (Section 3.4), I examined kin selectionist and group selectionist methods of sorting populations of organisms into equivalence classes. Chapter 4 explored the differences in

scope and accuracy between genetic and phenotypic formulations of Hamilton's rule, while Chapter 5 investigated the relationship between the 'inclusive' and 'neighbour-modulated' conceptions of social fitness. In all four instances, we were faced with alternative representations of the same evolutionary process. We therefore faced a question as to the compatibility of the alternatives; and, regardless of the answer to that question, we faced a further question as to which is theoretically preferable.

In all four instances, I have defended a qualified form of pluralism. On the one hand, I do not think any of these disputes are settled unequivocally by considerations of accuracy. There is no reason (for example) to say that kin selectionist methods of analysis are always more accurate that group selectionist methods, or that neighbour-modulated fitness is always more accurate than inclusive fitness. Indeed, we have seen that, in many cases, 'formal equivalence' results are available, demonstrating that under specified conditions the alternative representations make the same predictions regarding the direction of natural selection. On the other hand, I have resisted the move from formal equivalence to *explanatory* equivalence. For even when two approaches agree about the direction of selection, they may differ in the depth of *explanation* they provide as to why selection has the direction it does. Moreover, I have repeatedly emphasized that formal equivalence *under specified conditions* does not imply formal equivalence *under all conditions*, and we have often seen that representations claimed to be 'formally equivalent' in fact diverge in certain important cases. So my pluralism is doubly qualified: predictive equivalence under specified conditions does not imply explanatory equivalence, nor does it imply predictive equivalence under *all* conditions.

In some ways, this qualified pluralism is *more* seriously and substantively pluralistic than an unqualified pluralism on which we take our alternative representations to be equivalent under all conditions and in all senses of the word. For, in practice, unqualified pluralism is often invoked as a licence to *ignore* one mode of representation altogether, because we can safely assume that no predictive or explanatory content is lost when we

favour the alternative mode. Hence (for example) the presumed formal equivalence of kin and group selection is often invoked as a licence to ignore the latter for serious explanatory purposes, and the presumed formal equivalence of neighbour-modulated and inclusive fitness is often invoked as a licence to model social phenomena predominantly in terms of the former while continuing to gloss the results verbally in terms of the latter. My qualified pluralism provides no such licence. On the contrary, it implies that we ought to actively develop and employ *both* alternative formal representations in each case, because a global preference for one over the other will sometimes result in a loss of explanatory content or predictive accuracy.

Coping with causal complexity

In Chapter 1, I suggested that one reason social evolution theory is deserving of greater philosophical scrutiny—and deserving of it *now* more than ever—is that it has come to occupy a central rather than peripheral role in evolutionary biology. Year after year, we are discovering incredible social phenomena in places we never thought to look before (see Chapter 2). The microbial world in particular turns out to be richly social, leading to the suggestion that the multicellular organism itself may be regarded as an example of microbial sociality taken to extremes (Queller 1997, 2000; Queller and Strassmann 2009; Bourke 2011a). At the same time, we are increasingly coming to see that even in more familiar cases, social behaviour can be much more complex than we previously thought. Cooperation in advanced eusocial insect colonies, for example, often turns out to have a sophisticated task structure, requiring impressive feats of coordination on the part of the workers (see Chapter 2; see also Anderson et al. 2001). This has led to a revival of the controversial suggestion that eusocial insect colonies constitute organisms in their own right (Hölldobler and Wilson 2009, 2011; Queller and Strassmann 2009; Wilson 2012).

In both trends, one sees a growing realization on the part of social evolution theorists that complex causal relationships between individual behaviours and their downstream fitness consequences are the norm in nature, not the exception. Some theorists take such causal

complexity to pose a problem for kin selection theory as traditionally formulated, since the fitness payoffs in Hamilton's original (1964) models are assumed to depend only on the behaviour of a single actor (e.g., Queller 1992a; smith et al. 2010; Cornforth et al. 2012; Damore and Gore 2012). Others maintain that the theory as originally formulated already accommodates more complex cases (e.g., Grafen 1985b; Gardner et al. 2007, 2011). Many of the chapters in this dissertation have been concerned, directly or indirectly, with the question of how (if at all) the conceptual and formal foundations of Hamilton's theory need to be extended to accommodate causal complexity.

In Chapter 2, for example, I argued that a classification of social behaviours purely on the basis of the sign of their fitness effects misses some of the most interesting causal distinctions between social behaviours in nature: in many cases, it is more illuminating to categorize behaviours by considerations of task structure, or by considerations of control. In Chapter 3, I argued that we should be cautious about interpreting any particular term in the Price equation as 'the change due to natural selection', since the relationship between natural selection and changes in gene frequency is more complex than it may at first appear to be. In Chapter 4, I considered whether the standard version of Hamilton's rule still holds when relatives interact synergistically. Developing a line of argument presented informally by Queller (1992a, 2011), I argued that it does not (at least if the '*b*' and '*c*' terms are interpreted as average effects of phenotypes); and I discussed two ways to repair the rule to accommodate synergy. In Chapter 5, I considered whether the notion of inclusive fitness still makes sense when fitness effects depend in causally complex ways on behavioural phenotypes, and I argued that it (probably) does. Importantly, however, Grafen's arguments for inclusive fitness optimization (discussed in Chapter 6) assume a simple relationship between fitness effects and individual behaviours, and it is doubtful whether his arguments still work in the absence of this assumption. The bottom line (to the extent that there is one) is that kin selection theory does not *easily* accommodate causal complexity, but that it can be extended in various ways that allow it to do so.

Devils in the details

This dissertation blends the theory and philosophy of evolutionary genetics to such a degree that there can be no clean separation of the 'philosophical' parts from the 'theoretical' parts. To the extent that the dissertation is a defence of any single idea, it is a defence of this way of doing philosophy. In general, I have endeavoured to avoid superfluous technical detail, so as to make my arguments accessible to a general philosophical audience. Repeatedly, however, I have found myself attending to details I had originally hoped to avoid, because they turned out not to be superfluous after all.

In several of the chapters, a close engagement with the formal details of social evolution theory has led us to conclusions that a verbal treatment of the same questions would have missed. In Chapter 3 (Section 3.3), for example, I argued that natural selection can potentially affect the evolution of a trait in three different ways: firstly by generating covariance between the trait and fitness, secondly by generating covariance between transmission biases with respect to the trait and fitness, and thirdly by changing the average effects of alleles on the trait. The conceptual distinction between these three effects is difficult to grasp, as is the precise way in which they interact to yield the overall change in the population mean; but matters become much clearer when we distinguish these effects using the Price equation. This result subsequently played an important role in Chapter 6, where I used it to cast doubt on the claim that Fisher's 'fundamental theorem of natural selection' succeeds in capturing a biologically significant truth about the partial change attributable to natural selection.

In Chapter 5, I addressed the question of when (if at all) neighbour-modulated and inclusive fitness models of social evolution are formally equivalent. The devil in the detail here is that class structure makes a difference: if we only consider populations without class structure, it is hard to see how the two frameworks could ever disagree, because social actors and recipients are equivalent in every sense that matters. But if (following Peter D. Taylor, Steven A. Frank and others) we allow that actors and recipients may

belong to different behavioural or morphological classes, we start to see how the frameworks might come apart. And I argued that they do indeed come apart, in certain interesting cases. This is one example of what may at first seem like an irrelevant detail turning out to make a big difference: if we ignored class structure altogether, we would fail to see interesting and substantive differences between the inclusive and neighbour-modulated fitness perspectives.

This dissertation is not 'one long argument', but these recurring themes lend it a substantial degree of unity. While the chapters can (to some extent) be read and understood separately, the arguments they contain are complementary and mutually reinforcing, and they exemplify the interdisciplinary methodology I advocated in Chapter 1. The result, I hope, is an examination of the concept of kin selection that is more than the sum of its parts.

APPENDICES

Appendix A: Partitioning covariance into between- and within-subset components

Price (1972a) presents a form of the Price equation partitioned into between- and within-subset components, and suggests that this provides a useful starting point for the analysis of group selection. The procedure for partitioning (the covariance term of) the Price equation is fairly straightforward but is mostly elided by Price, generating scope for doubt about the generality of the result. In this appendix, I show that the partitioning of covariance in this way requires no substantive assumptions about the biological nature of the 'subsets' with which we are concerned.

Let us first introduce some appropriate notation. Let G_{ij} represent the local group mean of the j^{th} individual of the i^{th} group with respect to breeding value; let W_{ij} represent the local group mean of the j^{th} individual of the i^{th} group with respect to fitness. Let G_i represent the group mean of the i^{th} group with respect to breeding value; let W_i represent the group mean of the i^{th} group with respect to fitness. Trivially, an individual's local group mean is always equal to the mean of the group of which it is a member, i.e., $G_i = G_{ij}$ and $W_i = W_{ij}$ for all i, j. But it is initially important to treat local group means as contextual properties of the group members, and to index them accordingly.

We start the derivation with equation (3.4.2) in the main text:

$$\mathrm{Cov}(w,g) = \sum_{ij} \frac{q_i}{m_i} \left(w_{ij} - \bar{w}\right)\left(g_{ij} - \bar{g}\right)$$

We rewrite this equation as follows, in order to introduce local group means with respect to *breeding value*:

$$\mathrm{Cov}(w,g)=\sum_{ij}\frac{q_i}{m_i}\left(w_{ij}-\bar{w}\right)\left(g_{ij}-G_{ij}+G_{ij}-\bar{g}\right)$$

We then bifurcate the summand as follows:

$$\mathrm{Cov}(w,g)=\sum_{ij}\frac{q_i}{m_i}\left(w_{ij}-\bar{w}\right)\left(g_{ij}-G_{ij}\right)+\sum_{ij}\frac{q_i}{m_i}\left(w_{ij}-\bar{w}\right)\left(G_{ij}-\bar{g}\right)$$

We rewrite the above equation, in order to introduce local group means with respect to *fitness*:

$$\mathrm{Cov}(w,g)=\sum_{i}\frac{q_i}{m_i}\left(w_{ij}-W_{ij}+W_{ij}-\bar{w}\right)\left(g_{ij}-G_{ij}\right)+\sum_{ij}\frac{q_i}{m_i}\left(w_{ij}-W_{ij}+W_{ij}-\bar{w}\right)\left(G_{ij}-\bar{g}\right)$$

We then further partition the summands:

$$\begin{aligned}\mathrm{Cov}(w,g)=&\sum_{i}\frac{q_i}{m_i}\left(w_{ij}-W_{ij}\right)\left(g_{ij}-G_{ij}\right)+\sum_{ij}\frac{q_i}{m_i}\left(W_{ij}-\bar{w}\right)\left(g_{ij}-G_{ij}\right)\\&+\sum_{ij}\frac{q_i}{m_i}\left(w_{ij}-W_{ij}\right)\left(G_{ij}-\bar{g}\right)+\sum_{ij}\frac{q_i}{m_i}\left(W_{ij}-\bar{w}\right)\left(G_{ij}-\bar{g}\right)\end{aligned}$$

We exploit the fact that $G_i=G_{ij}$ and $W_i=W_{ij}$ for all *i, j* to rewrite the above expression as follows:

$$\begin{aligned}\mathrm{Cov}(w,g)=&\sum_{ij}\frac{q_i}{m_i}\left(w_{ij}-W_{ij}\right)\left(g_{ij}-G_{ij}\right)+\sum_{i}q_i\left(W_i-\bar{w}\right)\sum_{j}\left(g_{ij}-G_{ij}\right)\\&+\sum_{i}q_i\left(G_i-\bar{g}\right)\sum_{j}\left(w_{ij}-W_{ij}\right)+\sum_{ij}\frac{q_i}{m_i}\left(W_{ij}-\bar{w}\right)\left(G_{ij}-\bar{g}\right)\end{aligned}$$

Next, we can exploit the fact that $\sum_j \left(w_{ij} - W_{ij}\right) = 0$ and $\sum_j \left(g_{ij} - G_{ij}\right) = 0$ for all i (i.e., the average absolute deviation [*not* squared deviation] of an individual from its local group mean is zero for all groups; this is a general property of arithmetic means) to eliminate the second and third terms, yielding:

$$\mathrm{Cov}(w,g) = \sum_{ij} \frac{q_i}{m_i}\left(w_{ij} - W_{ij}\right)\left(g_{ij} - G_{ij}\right) + \sum_{ij} \frac{q_i}{m_i}\left(W_{ij} - \bar{w}\right)\left(G_{ij} - \bar{g}\right)$$

Finally, because, $G_i = G_{ij}$ and $W_i = W_{ij}$ for all i, j, a sum of $\left(W_{ij} - \bar{w}\right)\left(G_{ij} - \bar{g}\right)$ over i, j is equivalent to a sum of $m_i\left(W_i - \bar{w}\right)\left(G_i - \bar{g}\right)$ over i. This allows us to simplify the second term as follows:

$$\mathrm{Cov}(w,g) = \sum_{ij} \frac{q_i}{m_i}\left(w_{ij} - W_{ij}\right)\left(g_{ij} - G_{ij}\right) + \sum_{i} q_i\left(W_i - \bar{w}\right)\left(G_i - \bar{g}\right)$$

This is the partition of covariance deployed in the main text, and originally derived by Price (1972a).

Appendix B: The Taylor-Frank method

When differences between individuals with respect to the social behaviour under study are very small, and when the effects on fitness of these differences are very weak, we can reasonably approximate *differences* with *differentials*, and replace *partial regression coefficients* with *partial derivatives*. In the case of Hamilton's rule, this amounts to making the following approximations:

$$\beta_{wz|\hat{z}} \approx \frac{\partial w}{\partial z} \qquad \beta_{w\hat{z}|z} \approx \frac{\partial w}{\partial \hat{z}} \qquad \beta_{\hat{z},g} \approx \frac{d\hat{z}}{dg}$$

If we substitute these approximations into Hamilton's rule, we obtain a 'marginal' version that holds for very small behavioural deviations from the mean and very weak fitness effects:

$$\Delta_{1^o}\bar{g} > 0 \text{ iff } \frac{\partial w}{\partial z} + \frac{d\hat{z}}{dz} \cdot \frac{\partial w}{\partial \hat{z}} > 0$$

By turning the inequality into an equation, and by further assuming that the primary effect of selection is the only effect that influences the evolution of the trait under study, we can obtain a useful sufficient condition[1] for evolutionary equilibrium:

$$\Delta\bar{g} = 0 \text{ if } \frac{\partial w}{\partial z} + \frac{d\hat{z}}{dz} \cdot \frac{\partial w}{\partial \hat{z}} = 0$$

For a given model of the functional relationships between w, z, and $\hat{z}$, we can solve this differential equation to find the evolutionarily stable values for z. This is the basis for Peter

[1] This condition is not *necessary* because, trivially, the effect of selection will also be zero if there is no genetic variance in the relevant character.

D. Taylor and Steven A. Frank's (1996) influential method for the prediction of evolutionarily stable strategies (see also Frank 1995, 1997b, 1998).

I include an explanation of Taylor and Frank's method here only so as to leave it aside in the main text. For it is sometimes supposed that *Hamilton's rule itself* requires that there be only very small deviations from the mean in the behaviour under study, and that these deviations issue in only very small fitness differences. Crucially, however, these assumptions are only required to *approximate* Hamilton's rule using partial derivatives. The fundamental version of the rule requires no such assumptions (cf. Gardner et al. 2011).

There is a general moral here. In order to understand the foundations of contemporary social evolution theory, it is vital to distinguish between those assumptions which are *essential* to the theory's formulation and those which are only *methodologically convenient* for its application to particular cases (cf. Gardner et al. 2011). The marginal version of Hamilton's rule provides a convenient route to theoretical predictions of evolutionary stability. But the fundamental version from which it is derived uses partial regression coefficients, not differentials, and makes no assumption that the former may be approximated by the latter. Since my concern is with general theory rather than its application to particular problems, I focus exclusively on maximally general, regression-based formulations of kin selection theory in the main text.

Appendix C: Regression analysis of synergy games

Part I: The simple synergy game

Individuals interact in pairs. They may either cooperate ($z=1$) or defect ($z=0$), and their strategy is fully determined by their allelic value at the focal locus, x (i.e., $x=0\rightarrow z=0$ and $x=1\rightarrow z=1$). Note that, because allelic value determines phenotype, allelic values are equivalent to breeding values (i.e., $x=0\rightarrow g=0$ and $x=1\rightarrow g=1$).

A fraction a of individuals are assigned social partners with allelic values guaranteed to be identical to their own, while a fraction $\left(1-a\right)$ are assigned a social partner at random from the (infinite) population. The background frequency of the allele ($x=1$) is p. As stated in the main text, the payoff matrix is as follows (where $\hat{z}$ denotes the phenotype of the focal individual's social partner):

	COOPERATE ($\hat{z}=1$)	**DEFECT** ($\hat{z}=0$)
COOPERATE ($z=1$)	$B-C+D$	$-C$
DEFECT ($z=0$)	B	0

The frequencies of the various possible character combinations are given by the following 2 x 2 table:

	$\hat{z}=0$	$\hat{z}=1$
$z=0$	$(1-p)\left(a+(1-a)(1-p)\right)$	$p(1-p)(1-a)$
$z=1$	$p(1-p)(1-a)$	$p\left(a+(1-a)p\right)$

To apply HRP, we analyse fitness using the following, two-predictor regression model (where, as in the main text, $-c=\beta_{w,z|\hat{z}}$ and $b=\beta_{w,\hat{z}|z}$):

$$w=-cz+b\hat{z}+\varepsilon_w$$

This model of fitness fully accounts for the overall *w*-*g* covariance if and only if Queller's 'separation condition' is satisfied:

$$\textbf{Separation condition}: \mathrm{Cov}\left(g,\varepsilon_w\right)=0$$

In the regression model under consideration, the separation condition is equivalent to the following condition, which will be more useful for current purposes:

$$\textbf{Separation condition:}\ \beta_{w,g}=rb-c$$

As in the main text, '*r*', is the regression of one's social partner's phenotype on one's breeding value (i.e., $r=\beta_{\hat{z},g}$).

To ascertain beyond any doubt whether or not the separation condition is satisfied by the two-predictor regression in this game—and hence whether or not HRP is a reliable guide to the direction and magnitude of selection—we can compute the left- and right-hand

sides of the above expression and see if they differ. We can do this using the information contained in the payoff and frequency tables. In this instance, I will go through the computation step by step to illustrate how it works.

Let us start with *r*. As in the main text, this is formally defined as the regression of one's social partner's phenotype on one's *breeding value*. In this game, because allelic value determines strategy, this quantity is equal to the regression of one's social partner's phenotype on one's *phenotype*, and both quantities are equal to the *a* parameter:

$$r = \beta_{\hat{z},g} = \beta_{\hat{z},z} = a$$

To calculate the partial regression coefficient corresponding to the '*-c*' term in HRP from the simple regression coefficients, we start by calculating the relevant conditional expected values:

$$\mathrm{E}\left(w \middle| z=1\right) = \mathrm{E}\left(w \middle| g=1\right) = -C + \left(a + \left(1+a\right)p\right)\left(B+D\right)$$

$$\mathrm{E}\left(w \middle| z=0\right) = \mathrm{E}\left(w \middle| g=0\right) = B\left(1-a\right)p$$

$$\mathrm{E}\left(w \middle| \hat{z}=1\right) = B + \left(a + \left(1+a\right)p\right)\left(-C+D\right)$$

$$\mathrm{E}\left(w \middle| \hat{z}=0\right) = -C\left(1-a\right)p$$

From these, we compute the following simple (i.e., not partial) regressions:

$$\beta_{w,z} = \beta_{w,g} = \mathrm{E}\left(w \middle| z=1\right) - \mathrm{E}\left(w \middle| z=0\right) = -C + aB + \left(a + \left(1-a\right)p\right)D$$

$$\beta_{w,z'} = \mathrm{E}\left(w \middle| \hat{z}=1\right) - \mathrm{E}\left(w \middle| \hat{z}=0\right) = B - aC + \left(a + \left(1-a\right)p\right)D$$

To compute the '*c*' term in HRP from these simple regressions, we apply the general formula for partial regression coefficients (see Section 4.1.1):

$$-c = \beta_{w,z|\hat{z}} = \frac{\beta_{w,z} - \beta_{w,\hat{z}} \cdot \beta_{\hat{z},z}}{1 - \rho^2_{\hat{z},z}}$$

$\rho_{\hat{z},z}$ denotes the Pearson correlation coefficient between z and $\hat{z}$. In this case, because z and $\hat{z}$ can only adopt the values 0 or 1, this quantity is equal to $\beta_{\hat{z},z}$.

Substituting our simple regressions into the formula, we obtain:

$$-c = \frac{-C + aB + \left(a + \left(1-a\right)p\right)D - a\left[B - aC + \left(a + \left(1-a\right)p\right)D\right]}{1 - a^2}$$

Which simplifies to the following:

$$c = C - \frac{\left(a + \left(1-a\right)p\right)D}{\left(1+a\right)}$$

This recovers the expression Gardner et al. (2007) derive for the partial regression coefficient corresponding to the '*c*' term in Hamilton's rule. An exactly parallel procedure yields the expression for the partial regression coefficient corresponding to the '*b*' term:

$$b = B + \frac{\left(a + \left(1-a\right)p\right)D}{\left(1+a\right)}$$

Substituting our expressions for *r*, *b*, *c* and $\beta_{w,g}$ back into (the more conveniently expressed version of) the separation condition, we obtain:

$$-C+aB+\left(a+\left(1-a\right)p\right)D=a\left[B+\frac{\left(a+\left(1-a\right)p\right)D}{1+a}\right]-\left[C-\frac{\left(a+\left(1-a\right)p\right)D}{1+a}\right]$$

The left- and right-hand sides of this expressions are equivalent, showing that $\beta_{w,g}=rb-c$ for any values of the parameters B, C, D, a and p. The separation condition is satisfied, and $rb-c$ is a reliable guide to the direction of selection.

Part II: The extended synergy game

HRP holds unproblematically in the simple synergy game, in which allelic value fully determines phenotype. We see a different story, however, when we relax this assumption. As before, a useful statement of Queller's separation condition is the following:

$$\beta_{w,g}=rb-c$$

If the condition is satisfied, the two-predictor model of fitness employed by HRP fully accounts for the overall *w-g* covariance, and can therefore be trusted as a guide to the direction and magnitude of selection. Queller (1992a) argues informally that, when (i) genotypes are imperfect predictors of phenotypes and (ii) synergistic interactions are at work, it is typically *not* the case that the separation condition is satisfied, so HRP should *not* be trusted as a guide to the overall direction and magnitude of selection. We can use an extended synergy game to make a formal argument for the same conclusion.

The setup for the extended synergy game is as follows. As before, a fraction a of individuals are assigned a social partner with an allelic value guaranteed to be identical to their own. A fraction $\left(1-a\right)$ have their social partner drawn randomly from the population. Of individuals with the allele ($x=1$), a fraction k express the cooperative phenotype ($z=1$). Individuals who do not possess the allele ($x=0$) never express the cooperative phenotype (this somewhat simplifies the calculations).

The payoff matrix is as before. The frequencies of the various possible character combinations are given in the following 3 x 3 table (where $\hat{x}$ denotes the allelic value of the focal individual's social partner, which may now differ from its phenotypic value):

	$\hat{x}=0, \hat{z}=0$	$\hat{x}=1, \hat{z}=0$	$\hat{x}=1, \hat{z}=1$
$x=0, z=0$	$\left(a+\left(1-a\right)\left(1-p\right)\right)\left(1-p\right)$	$p\left(1-p\right)\left(1-a\right)\left(1-k\right)$	$p\left(1-p\right)\left(1-a\right)k$
$x=1, z=0$	$p\left(1-p\right)\left(1-a\right)\left(1-k\right)$	$\left(a+\left(1-a\right)p\right)p\left(1-k\right)^2$	$\left(a+\left(1-a\right)p\right)pk\left(1-k\right)$
$x=1, z=1$	$p\left(1-p\right)\left(1-a\right)k$	$\left(a+\left(1-a\right)p\right)pk\left(1-k\right)$	$\left(a+\left(1-a\right)p\right)pk^2$

All the results that follow can be computed (tediously) from the relevant frequencies and payoffs.

To evaluate whether the separation condition obtains, we need to evaluate $\beta_{w,g}$ and $rb-c$ and see if they are equal. As a preliminary, we need to introduce breeding values. In the simple synergy game, breeding values can simply be equated with allelic values because $\beta_{z,x}=1$, but that is not the case here: in the extended game, $\beta_{z,x}=k$. Accordingly, individuals with the allele have a breeding value for z equal to k ($x=1\rightarrow g=k$), while individuals without the allele have a breeding value for z equal to 0 ($x=0\rightarrow g=0$).

The coefficient of relatedness, *r*, is once again equal to *a*:

$$r=\beta_{\hat{z},g}=a$$

Note that the coefficient of relatedness is defined (as in the main text, and in Queller's original article) strictly in terms of *breeding values*, not allelic values. If relatedness had

been defined as a regression of one's partner's phenotype on one's allelic value, the coefficient of relatedness would have been *ka*, reflecting the fact that one's partner may not express the social allele even if it has it. But this is already taken into account implicitly when we switch from allelic values to breeding values.

Social partners are also phenotypically correlated. Unlike in the simple synergy game, however, the phenotypic association between social partners is *not* equal to the coefficient of relatedness. Because the coefficient of relatedness regresses one's social partner's phenotype on one's *breeding value*, it takes no account of whether or not one's own genes for z are actually expressed. In some cases, social partners will share genes for the cooperative behaviour and yet one will fail to express the trait, so we should expect to find that the overall phenotypic association between social partners is weaker than the coefficient of relatedness. Formally, the regression of one's partner's phenotype on one's *phenotype* in this game is equal to:

$$\pi = \beta_{\hat{z},z} = kR - \frac{1}{1-kp}\left[\left(1-p\right)\left(1-R\right)kp + k\left(1-k\right)\left(R+\left(1-R\right)p\right)p - \left(1-kp\right)\left(1-R\right)kp\right]$$

Assuming that k, a and p all assume values greater than 0 and less than 1, the second term in this expression is sure to be positive, implying that $r > \pi$, as we expected. In the special case of $k = 1$ (perfect prediction of phenotype by genotype), the second term goes to zero and $\pi = r = a$; this takes us back to the simple synergy game.

The simple regression of fitness on breeding value in the extended game is given by:

$$\beta_{w,g} = -C + aB + k\left(a + \left(1-a\right)p\right)D$$

Applying the formula for partial regression coefficients, we find that the c term in HRP is given by:

$$c = -\beta_{w,z|\hat{z}} = C - \frac{k}{1+\pi}\left(a + \left(1-a\right)p\right)D$$

The b term in HRP (the partial regression of fitness on one's social partner's phenotype, controlling for one's own phenotype) is given by:

$$b = \beta_{w,\hat{z}|z} = B + \frac{k}{1+\pi}\left(a + \left(1-a\right)p\right)D$$

As one would expect, in the special case of $k = 1$ (perfect prediction of phenotype by genotype), these coefficients collapse into those derived by Gardner et al. (2007) for the simple synergy game.

Putting the pieces together, we obtain:

$$rb - c = aB - C + k\left(\frac{1+a}{1+\pi}\right)\left(a + \left(1-a\right)p\right)D$$

Comparing our expressions for $\beta_{w,g}$ and $rb - c$, we see that Queller's intuition was correct: when synergy is present and genotype imperfectly predicts phenotype, these quantities are *not* equal. If genotype is a perfect predictor of phenotype (such that k=1) or if synergy is absent (such that D=0), then they are. But in the general case they are not. In short:

$$\beta_{w,g} = rb - c \text{ if and only if } D = 0 \vee k = 1$$

When the two quantities differ, the exact difference between them is given by:

$$\beta_{w,g} - \left(rb - c\right) = k\left(1 - \frac{1+a}{1+\pi}\right)\left(a + \left(1-a\right)p\right)D$$

If D is positive, then (given that k, a and p all assume values greater than 0 and less than 1) then this difference will be negative, implying that $rb - c > \beta_{w,g}$. If D is negative, the sign will be reversed, implying that $rb - c < \beta_{w,g}$. In short, we can see that HRP systematically *overcompensates* for the effects of synergy, regardless of whether the synergistic effect is positive or negative. Moreover, we cannot safely assume that the difference will be small, or irrelevant to the sign of the predicted change. If D is large, the difference between the true magnitude of selection and the magnitude predicted by HRP may be large too, and there is no formal guarantee that they will have the same sign.

In a nutshell, the discrepancy arises because, in the computation of the b and c coefficients, we divide the total expected synergistic effect by a factor $1+\pi$, where π is a measure of the association between social partner *phenotypes*. Yet when we combine the b and c coefficients in HRP, we multiply the expected synergistic effect by a factor $1+r$, where r is the regression of one's partner's phenotype on one's *breeding value*. Because in general $r > \pi$, the outcome is that we systematically overweight the synergistic effect. The problem disappears only in the special case of genetically determined strategies, in which case $r = \pi$. Two solutions to the problem are considered in the main text.

Appendix D: Inclusive fitness with delocalized control

In this appendix, I propose a means of formally extending the notion of inclusive fitness to cases of delocalized phenotypic control. The proposal is somewhat conjectural and certainly stands in need of further development: its main role in the present discussion is to make plausible the claim that such an extension is *possible in principle*, though work undoubtedly remains to be done.

The extension of the inclusive fitness formalism to cases of delocalized control requires the rather unfortunate practice of quadruple indexing. We need to introduce a new index, *l*, labelling all the relevant controlling actor classes (e.g., majors and minors) for the j^{th} neighbourhood phenotype (e.g., the completion of a collaborative task, such as intruder-decapitation) of the k^{th} member of the i^{th} recipient class. We then regress the value of the j^{th} social phenotype on the genotypes of *all* relevant controlling actor classes:

$$z_{ijk} = \sum_{l} d_{ijl} g_{ijkl} + \varepsilon$$

The *d*-coefficients in this analysis are naturally interpreted as measuring the degree to which the l^{th} actor class controls the j^{th} phenotype. Note that, while these play the same role as the 'coefficients of control' in the standard formalism, and can still be glossed in these terms, they allow for the fact that a social phenotype need not be under the *sole* control of a single actor class. Control of the phenotype can be delocalized across any number of actor classes.

Next, we perform a separate $\tilde{\tau}$-regression (one for each of the controlling actor classes) of the genotype of the actor on the genotype of the recipient's descendants.[1] Generically, for the l^{th} controlling actor-class:

$$g_{ijkl} = \tilde{\tau}_{ijl} g'_{ik} + \varepsilon$$

Substituting all these regressions into (4.2.1) we obtain:

$$\Delta_w \bar{g} = \frac{1}{\bar{w}} \left[\sum_i q_i \sum_j \beta_{ij} \sum_l d_{ijl} \tilde{\tau}_{ijl} \operatorname{Var}^i \left(g'_{ik} \right) \right]$$

To get an expression that admits of an intelligible interpretation in terms of inclusive fitness, we (as in the standard formalism) need to flip the direction of the $\tilde{\tau}$-regression, so that we are regressing the recipient's descendants' genotype on the actor genotype, rather than the other way round. We can do this by noting that, for all *ijl*, $\tilde{\tau}_{ijl} \operatorname{Var}\left(g'_{ik} \right) = \tau_{ijl} \operatorname{Var}\left(g_{ijkl} \right)$. This yields:

$$\Delta_w \bar{g} = \frac{1}{\bar{w}} \left[\sum_i q_i \sum_j \beta_{ij} \sum_l d_{ijl} \tau_{ijl} \operatorname{Var}^i \left(g_{ijkl} \right) \right]$$

Conditional on the assumption that the genetic variance is the same for every relevant actor class, this entails the following principle concerning the direction of social selection:

---

[1] Note that, if the same phenotype also affects multiple classes of recipient, we simply have to repeat the same procedure for every recipient class—the standard inclusive fitness formalism already involves summing over *i*.

$$\operatorname{sign}\left(\Delta_w \bar{g}\right) = \operatorname{sign}\left(\sum_i q_i \sum_j \beta_{ij} \sum_l d_{ijl} \tau_{ijl} \right)$$

This principle does admit of an interpretation in inclusive fitness terms: it says that selection will favour actor genotypes that are positively associated with inclusive fitness, where inclusive fitness is once again conceptualized as a τ-weighted sum of fitness effects for which a particular actor is causally responsible. But the extended formalism embodies a more nuanced conception of 'causal responsibility' than the standard formalism. We are no longer assuming that all social phenotypes are under the *sole* control of a particular actor, so that their effects count towards the inclusive fitness of only that actor. Instead, we allow that control of a social phenotype may be distributed among multiple actors, and may therefore contribute to the inclusive fitness of all these actors. The magnitude of the inclusive fitness effect of the j^{th} phenotype on an actor of the l^{th} controlling class will depend on (i) the size of the effect on the personal fitness of recipients of the i^{th} class, (ii) the τ-relatedness between actors and recipients, and (iii) the relevant coefficient of control (d_{ijl}). Accordingly, the inclusive fitness effect on a particular actor of its involvement in a collaborative task will be increased (all else being equal) by (i) bigger benefits if the task is completed, (ii) stronger τ-relatedness to the recipients, and (iii) a greater degree of control over the outcome.

In keeping with my overall project, my emphasis here is on general theory rather than the application of that theory to particular problems. But I assume that my extended framework for inclusive fitness with delocalized control could be applied to particular problems in the same way as the standard formalism: approximate differences with differentials, write a differential equation for the trait value at equilibrium, and solve to derive a prediction of the ESS (Taylor and Frank 1996; Frank 1998).

The extension of the inclusive fitness framework presented here is best viewed as the outline of a programme for future work, rather than as a completed project. Open

questions remain regarding the relationship between the 'standard' inclusive fitness formalism—which assumes phenotypes to be under the sole control of a particular actor-class—and my extended formalism, which allows for delocalized control. In particular, it is unclear whether the extended inclusive fitness framework still underwrites a useful maximizing-agent analogy. Existing defences of the maximizing-agent analogy explicitly invoke an assumption of 'actor control', i.e., they assume that an actor is solely responsible for its own inclusive fitness (see Grafen 2006a; Gardner and Welch 2011). This involves not merely an assumption of localized control of social phenotypes, but also an assumption that the fitness effects of a social phenotype are fully determined by the value of that phenotype (and are not, for instance, frequency-dependent). We can show that, conditional on various substantive assumptions including actor control, organisms whose behaviour has been optimized by selection will tend to act as if trying to maximize their standard inclusive fitness. But if we allow for delocalized control, is it still the case that organisms whose behaviour has been optimized by selection will tend to act as if trying to maximize their *extended* inclusive fitness? I do not know.

BIBLIOGRAPHY

Abbot, P., et al. 2011. Inclusive fitness theory and eusociality. *Nature* 471:E1-E2.

Allen, Colin, Marc Bekoff and George V. Lauder. 1998. *Nature's purposes: analyses of function and design in biology.* Cambridge, MA: Bradford Books.

Anderson, Carl and Nigel Franks. 2001. Teams in animal societies. *Behavioral Ecology* 12:534-540.

Anderson, Carl and Daniel W. McShea. 2001. Individual *versus* social complexity, with particular reference to ant colonies. *Biological Reviews* 76:211-237.

Anderson, Carl and Nigel R. Franks and Daniel W. McShea. 2001. The complexity and hierarchical structure of tasks in insect societies. *Animal Behaviour* 62:643-651.

Alexander, Richard D. 1974. The evolution of social behavior. *Annual Review of Ecology and Systematics* 5:325-383.

Ariew, André, Mark Perlman and Robert C. Cummins. 2002. *Functions: new essays in the philosophy of psychology and biology.* New York: Oxford University Press.

Ayers, Janelle S. and Russell E. Vance. 2012. Cellular teamwork in antibacterial innate immunity. *Nature Immunology* 13:115-117.

Beatty, John and Eric Cyr Desjardins. 2009. Natural selection and history. *Biology and Philosophy* 24:231-246.

Bekoff, Marc, Colin Allen and Gordon M. Burghardt (eds.). *The cognitive animal: empirical and theoretical perspectives on animal cognition*. Cambridge, MA: MIT Press.

Bekoff, Marc and Dale Jamieson. 1995. *Readings in animal cognition.* Cambridge, MA: MIT Press.

Berleman, James E. and John R. Kirby. 2009. Deciphering the hunting strategy of a bacterial wolfpack. *FEMS Microbiology Reviews* 33:942-957.

Biernaski, Jay M., Stuart A. West and Andy Gardner. 2011. Are greenbeards intragenomic outlaws? *Evolution* 65:2729-2742.

Bijma, P. 2006. Estimating maternal genetic effects in livestock. *Journal of Animal Science* 84:800-806.

Bijma, P. and M. J. Wade. 2008. The joint effects of kin, multilevel selection and indirect genetic effects on response to genetic selection. *Journal of Evolutionary Biology* 21:1175-1188.

Birch, Jonathan. 2012a. Collective action in the fraternal transitions. *Biology and Philosophy* 27:363-380.

———. 2012b. Social revolution (Review essay on Andrew F. G. Bourke: Principles of social evolution). *Biology and Philosophy* 27:571-581.

———. 2012c. The negative view of natural selection. *Studies in History and Philosophy of Biological and Biomedical Sciences* 43:569-573.

———. 2013a. Organismality as extreme sociality. Unpublished manuscript.

———. 2013b. Gene mobility and the concept of relatedness. Unpublished manuscript.

———. forthcoming. Hamilton's rule and its discontents. *British Journal for the Philosophy of Science*.

Bonner, John Tyler. 1959. *The cellular slime molds*. Princeton, NJ: Princeton University Press.

Bourke, Andrew F. G. 2011a. *Principles of social evolution*. Oxford: Oxford University Press.

———. 2011b. The validity and value of inclusive fitness theory. *Proceedings of the Royal Society B: Biological Sciences* 278:3313-3320.

——— and Nigel R. Franks. 1995. *Social evolution in ants*. Princeton, NJ: Princeton University Press.

Borrello, Mark E. 2010. *Evolutionary restraints: the contentious history of group selection.* Chicago, IL: University of Chicago Press.

Brandon, Robert N. 1982. The levels of selection. *Proceedings of the Philosophy of Science Association* 1:315-323.

———. 1988. Levels of selection: a hierarchy of interactors. In H. C. Plotkin (ed.), *The role of behaviour in evolution*. Cambridge, MA: MIT Press, 51-71

Briggs, Rachael. 2012. Interventionist counterfactuals. *Philosophical Studies* 160:139-166.

Buller, David J. 1999. *Function, selection and design*. Albany, NY: SUNY Press.

Buss, Leo W. 1987. *The evolution of individuality*. Princeton, NJ: Princeton University Press.

Calcott, Brett. 2006. Transitions in biological organization. PhD thesis, Australian National University.

———. 2008. The *other* cooperation problem: generating benefit. *Biology and Philosophy* 23:173-203.

Cavalli-Sforza, L. L. and M. W. Feldman. 1978. Darwinian, selection and 'altruism'. *Theoretical Population Biology* 12:268-280.

Chang, Hasok. 2004. *Inventing temperature: measurement and scientific progress*. New York: Oxford University Press.

Charlesworth, Brian. 1980. Models of kin selection. In H. Markl (ed.), *Evolution of social behaviour: hypotheses and empirical tests*. Weinheim: Verlag, 11-26.

Charnov, Eric L. 1977. An elementary treatment of the genetical theory of kin selection. *Journal of Theoretical Biology* 66:541–50.

Cheverud, James M. 1984. Evolution by kin selection: a quantitative genetic model illustrated by maternal performance in mice. *Evolution* 38:766-777.

Choe, Jae C. and Bernard C. Crespi (eds.). 1997. *The evolution of social behaviour in insects and arachnids*. Cambridge: Cambridge University Press.

Clarke, Ellen. 2011. Plant individuality and multilevel selection theory. In Brett Calcott and Kim Sterelny (eds.), *The major transitions in evolution revisited*. Cambridge, MA: MIT Press, 227-250.

Cornforth, Daniel M., Sumpter, David J., Brown, Sam P. and Åke Brännström. 2012. Synergy and group size in microbial cooperation. *American Naturalist* 180:296-305.

Creel, Scott. 1990. How to measure inclusive fitness. *Proceedings of the Royal Society B: Biological Sciences* 241:229-231.

Crespi, Bernard J. 2001. The evolution of social behaviour in microorganisms. *Trends in Ecology and Evolution* 16:178-183.

Cronk, Lee. 1991. Human behavioral ecology. *Annual Review of Anthropology* 20:25-53.

Damore, James A., and Gore, Jeff. 2012. Understanding microbial cooperation. *Journal of Theoretical Biology* 299:31-41.

Damuth, John and I. Lorraine Heisler. 1988. Alternative formulations of multilevel selection. *Biology and Philosophy* 3:407-430.

Dawkins, Richard. 1976. *The selfish gene*. New York: W. W. Norton & Company.

———. 1979. Twelve misunderstandings of kin selection. *Zietschrift für Tierpsychologie* 51:184-200.

———. 1982. *The extended phenotype: the gene as the unit of selection*. New York: W. W. Norton & Company.

———. 1986. *The blind watchmaker: why the evidence of evolution reveals a universe without design*. W. W. Norton & Company.

———. 1996. *Climbing mount improbable*. W. W. Norton & Company.

Dennett, Daniel C. 1995. *Darwin's dangerous idea: evolution and the meanings of life*. New York: Simon & Schuster.

Detrain, C. and J. M. Pasteels. 1992. Caste polyethism and collective defense in the ant *Pheidole pallidula*: the outcome of quantitative differences in recruitment. *Behavioral ecology and sociobiology*. 29:405-412.

Edwards, A. W. F. 1994. The fundamental theorem of natural selection. *Biological Reviews* 69:443-474.

———. 2000. *Foundations of mathematical genetics (2nd edition)*. Cambridge: Cambridge University Press.

———. 2007. Maximisation principles in evolutionary biology. In Mohan Matthen and Christopher Stephens (eds.) *Handbook of the Philosophy of Science: Philosophy of Biology*. Amsterdam: North-Holland, 335-347.

El Mouden, Claire and Andy Gardner. 2008. Nice natives and mean migrants: the evolution of dispersal-dependent behaviour in viscous populations. *Journal of Evolutionary Biology* 21: 1480-1491.

Eldakar, Omar Tonsi and David Sloan Wilson. 2011. Eight criticisms not to make about group selection. *Evolution* 65:1523-1526.

Ewens, Warren J. 1989. An interpretation and proof of the fundamental theorem of natural selection. *Theoretical population biology* 36:167-180.

———. 2004. *Mathematical population genetics*. New York: Springer.

———. 2011. What is the gene trying to do? *British Journal for the Philosophy of Science* 62:155-176.

Falconer, Douglas S. and Trudy F. C. Mackay. 1996. *Introduction to quantitative genetics (4th edition)*. London: Longman.

Fisher, Daniel C. 1985. Evolutionary morphology: beyond the analogous, the anecdotal and the ad hoc. *Paleobiology* 11:120-138.

Fisher, Ronald A. 1930. *The genetical theory of natural selection*. Oxford: Clarendon Press.

———. 1941. Average excess and average effect of a gene substitution. *Annals of Human Genetics* 11:53-63.

Fletcher, David J. C. and Kenneth G. Ross. 1985. Regulation of reproduction in eusocial Hymenoptera. *Annual Review of Entomology* 30:319-343.

Fletcher, Jeffrey A. and Michael Doebeli. 2006. How altruism evolves: assortment and synergy. *Journal of Evolutionary Biology* 18:1389-1393

———. 2009. A simple and general explanation for the evolution of altruism. *Proceedings of the Royal Society B: Biological Sciences* 276:13-19.

———. 2010. Assortment is a more fundamental explanation for the evolution of altruism than inclusive fitness or multilevel selection. *Proceedings of the Royal Society B: Biological Sciences* 277:677-678.

Fletcher, Jeffrey A. and Martin Zwick 2006. Unifying the theories of inclusive fitness and reciprocal altruism. *American Naturalist* 168:252-262.

Fletcher, Jeffrey A., Martin Zwick, Michael Doebeli and David Sloan Wilson. 2006. What's wrong with inclusive fitness? *Trends in Ecology and Evolution* 21:597-598.

Foster, Kevin R. 2009. A defense of sociobiology. *Cold Spring Harbor Symposium on Quantitative Biology*, 74:403-418.

Foster, Kevin R., Gad Shaulsky, Joan E. Strassmann, David C. Queller, Chris R. L. Thompson. 2004. Pleiotropy as a mechanism to stabilize cooperation. *Nature* 431:693-696.

Foster, Kevin R., Katie Parkinson and Christopher R. L. Thompson. 2007. What can microbial genetics teach sociobiology? *Trends in Genetics* 23:73-80

Frank, Steven A. 1995. George Price's contributions to evolutionary genetics. *Journal of Theoretical Biology* 175:373-88.

———. 1997a. The Price equation, Fisher's fundamental theorem, kin selection, and causal analysis. *Evolution* 51:1712-1729.

———. 1997b. Multivariate analysis of correlated selection and kin selection, with an ESS maximization method. *Journal of Theoretical Biology* 189:307-316.

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

———. 2006. Social selection. In C. W. Fox and J. B. Wolf (eds.), *Evolutionary genetics: concepts and case studies*. Oxford University Press, 350-363.

———. 2012. Natural selection. IV. The Price equation. *Journal of Evolutionary Biology* 25:1002-1019.

Frank, Steven A. and Montgomery Slatkin. 1992. Fisher's fundamental theorem of natural selection. *Trends in Ecology and Evolution* 7:92-95.

Franks, Nigel R. 1986. Teams in insect societies: group retrieval of prey by army ants (*Eciton burchelli*, Hymenoptera: Formicidae). *Behavioral ecology and sociobiology* 18:425-429.

———. 1987. The organization of working teams in insect societies. *Trends in Ecology and Evolution* 2:72-75.

Gadagkar, Raghavendra. 2001. *The social biology of* Ropalidia marginata*: towards understanding the evolution of eusociality*. Cambridge, MA: Harvard University Press.

Gadagkar, Raghavendra and John Tyler Bonner. 1994. Social insects and social amoebae. *Journal of Biosciences* 2:219-245.

Gardner, Andy. 2008. The Price equation. *Current Biology* 18:R198-R202.

———. 2009. Adaptation as organism design. *Biology Letters* 5:861-864.

———. 2011. Kin selection under blending inheritance. *Journal of Evolutionary Biology* 284:125-129.

Gardner, Andy and Kevin R. Foster. 2008. The evolution and ecology of cooperation—history and concepts. In J. Korb and J. Heinze (eds), *Ecology of social evolution*. Heidelberg: Springer-Verlag, 1-36.

Gardner, Andy and Alan Grafen. 2009. Capturing the superorganism: a formal theory of group adaptation. *Journal of Evolutionary Biology* 22:659-671.

Gardner, Andy and J. J. Welch. 2011. A formal theory of the selfish gene. *Journal of Evolutionary Biology* 24:1020-1043.

Gardner, Andy and Stuart A. West. 2004a. Spite and the scale of competition. *Journal of Evolutionary Biology* 17:1195-1203.

———. 2004b. Spite among siblings. *Science* 305:1413-1414.

Gardner, Andy, Stuart A. West and Nicholas H. Barton. 2007. The relation between multilocus population genetics and social evolution theory. *American Naturalist* 169:207-26.

Gardner, Andy, Stuart A. West and Geoff Wild. 2011. The genetical theory of kin selection. *Journal of Evolutionary Biology* 24:1020-1043.

Giraud, T. and J. A. Shykoff. 2011. Bacterial controlled by mobile elements: kin selection versus infectivity. *Heredity* 107:277-278.

Glymour, Bruce. 2008. Correlated Interaction and Group Selection. *British Journal for the Philosophy of Science* 59:835-855.

Godfrey-Smith, Peter. 2000. Information, arbitrariness and selection: comments on Maynard Smith. *Philosophy of Science* 67:202-207.

———. 2006. Local interaction, multilevel selection, and evolutionary transitions. *Biological Theory* 1:372-380.

———. 2007a. Conditions for evolution by natural selection. *Journal of Philosophy* 104:489-516.

———. 2007b. Information in biology. In David L. Hull and Michael Ruse (eds.), *The Cambridge companion to the philosophy of biology*. Cambridge: Cambridge University Press, pp. 103-113.

———. 2008. Varieties of population structure and the levels of selection. *British Journal for the Philosophy of Science* 59:25-50.

———. 2009a. Darwinian populations and natural selection. Oxford: Oxford University Press.

———. 2009b. Abstraction, idealizations, and evolutionary biology. In Anouk Barberousse, Michel Morange and Thomas Pradeu (eds.), *Mapping the future of biology: evolving concepts and theories*. Dordrecht: Springer, 47-56.

Godfrey-Smith, Peter and Benjamin Kerr. 2002. Individualist and multi-level perspectives on selection in structured populations. *Biology and Philosophy* 17:477-517.

———. 2009. Selection in ephemeral networks. *American Naturalist* 174:906-911.

———. forthcoming. Gestalt-switching and the evolutionary transitions. *British Journal for the Philosophy of Science*.

Godfrey-Smith, Peter and Jon F. Wilkins. 2009. Adaptationism and the adaptive landscape. *Biology and Philosophy* 24: 199-214.

Goodnight, Charles J., James M. Schwartz and Lori Stevens. 1992. Contextual analysis of models of group selection, soft selection, hard selection, and the evolution of altruism. *American Naturalist* 140:743-761.

Gorelick, Root, Susan M. Bertram, Peter R. Killeen and Jennifer H. Fewell. 2004. Normalized mutual entropy in biology: quantifying division of labor. *American Naturalist* 164:677-682.

Gould, Stephen Jay. 2002. *The structure of evolutionary theory*. Cambridge, MA: Harvard University Press.

Grafen, Alan. 1979. The hawk-dove game played between relatives. *Animal Behaviour* 27:905-907.

———. 1982. How not to measure inclusive fitness. *Nature* 298:425-426.

———. 1984. Natural selection, kin selection and group selection. In J. R. Krebs & N. B. Davies (eds.), *Behavioural ecology* (2nd edition). Oxford: Blackwell, 62-84.

———. 1985a. A geometrical view of relatedness. *Oxford Surveys in Evolutionary Biology* 2:28-89.

———. 1985b. Hamilton's rule OK. *Nature* 318:310-311.

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

———. 2000. Developments of the Price equation and natural selection under uncertainty. *Proceedings of the Royal Society B: Biological Sciences* 267:1223-1227.

———. 2002. A first formal link between the Price equation and an optimization program. *Journal of Theoretical Biology* 217:75-91.

———. 2006a. Optimization of inclusive fitness. *Journal of Theoretical Biology* 238:541-63.

———. 2006b. A theory of Fisher's reproductive value. *Journal of Mathematical Biology* 53:15-60.

———. 2007a. The formal Darwinism project: a mid-term report. *Journal of Evolutionary Biology* 20:1243-1254.

———. 2007b. Detecting kin selection at work using inclusive fitness. *Proceedings of the Royal Society B: Biological Sciences* 274:713-719.

———. 2007c. An inclusive fitness analysis of altruism on a cyclical network. *Journal of evolutionary Biology* 20:2278-2283.

———. 2008. The simplest formal argument for fitness optimisation. *Journal of Genetics* 87:421-433.

———. 2009. Formalizing Darwinism and inclusive fitness theory. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364:3135-3141.

Grozinger, Christina M., Noura M. Sharabash, Charles W. Whitfield and Gene E. Robinson. 2003. Pheromone-mediated gene expression in the honey bee brain. *Proceedings of the National Academy of Sciences USA* 100: 14519-14525.

Haldane, J. B. S. 1955. Population genetics. In M. L. Johnson, M. Abercrombie and G. E. Fogg (eds.) *New Biology 18*. London: Penguin, 34-51.

Hamilton, W. D. 1963. The evolution of altruistic behaviour. *American Naturalist* 97:354-356.

———. 1964. The genetical evolution of social behaviour. *Journal of Theoretical Biology* 7:1-52.

———. 1970. Selfish and spiteful behaviour in an evolutionary model. *Nature* 228:1218-1220.

———. 1971. Selection of selfish and altruistic behaviour in some extreme models. In J. F. Eisenberg and W. S. Dillon (eds.), *Man and beast: comparative social behavior*. Washington, DC: Smithsonian press, 57-91.

———. 1972. Altruism and related phenomena, mainly in social insects. *Annual Review of Ecology and Systematics* 3:193-232.

———. 1975. Innate social aptitudes of man: an approach from evolutionary genetics. In R. Fox (ed.), *Biosocial anthropology*. New York: Wiley, 133-55.

Hastings, Michele D, David C. Queller, Frank Eischen and Joan E. Strassmann. 1998. Kin selection, relatedness, and worker control of reproduction in a large colony epiponine wasp, *Brachygastra mellifica*. *Behavioral Ecology* 9:573-581.

Heisler, I. Lorraine and John Damuth. 1987. A method for analyzing selection in hierarchically structured populations. *American Naturalist* 130:582–602.

Helanterä, Heikki and Tobias Uller. 2010. The Price equation and extended inheritance. *Philosophy and Theory in Biology* 2:e101.

Henrich, Joseph. 2004. Cultural group selection, coevolutionary processes and large-scale cooperation. *Journal of Economic Behavior and Organization* 53:3-35.

Henrich, Joseph and Robert Boyd. 2001. Why people punish defectors: conformist transmission stabilizes costly enforcement of norms in cooperative dilemmas. *Journal of Theoretical Biology* 208:79-89.

Herbers, Joan M. 1981. Reliability theory and foraging by ants. *Journal of Theoretical Biology* 89:175-190.

Hölldobler, Bert and E. O. Wilson. 1990. *The ants*. Cambridge, MA: Harvard University Press.

———. 2009. *The superorganism: the beauty, elegance and strangeness of insect societies*. New York: W. W. Norton & Company.

———. 2011. *The leafcutter ants: civilization by instinct*. New York: W. W. Norton & Company.

Holman, Luke, Charlotte G. Jørgensen, John Nielsen and Patrizia d'Ettorre. 2010. Identification of an ant queen pheromone regulating worker sterility. *Proceedings of the Royal Society B: Biological Sciences* 277:3793-3800.

Immler, Simone, Harry D. M. Moore, William G. Breed and Tim R. Birkhead. 2007. By hook or by crook? Morphometry, cooperation and competition in rodent sperm. *PLoS ONE* 2:e170.

Jablonka, Eva and Marion J. Lamb. 2005. *Evolution in four dimensions: genetic, epigenetic, behavioral, and symbolic variation in the history of life*. Cambridge, MA: MIT Press.

Kerr, Benjamin and Peter Godfrey-Smith. 2009. Generalization of the Price equation for evolutionary change. *Evolution* 63:531-536.

Kramer, Paul J. 1984. Misuse of the term 'strategy'. *Bioscience* 34:405.

Andreas Kupz, Greta Guarda, Thomas Gebhardt, Leif E. Sander, Kirsty R Short, Dimitri A Diavatopoulos, Odilia L. C. Wijburg, Hanwei Cao, Jason C. Waithman, Weisan Chen, Daniel Fernandez-Ruiz, Paul G. Whitney, William R. Heath, Roy Curtiss III, Jürg Tschopp, Richard A. Strugnell & Sammy Bedoui. 2012. NLRC4 inflammasomes in dendritic cells regulate noncognate effector function by memory $CD8^+$ T cells. *Nature Immunology* 13:262-269.

Lande, Russell and Stevan J. Arnold. 1983. The measurement of selection on correlated characters. *Evolution* 37:1210-1226.

Lehmann, Laurent, Laurent Keller, Stuart West and Denis Roze. 2007. Group selection and kin selection: two concepts but one process. *Proceedings of the National Academy of the Sciences USA* 104:6736-6739.

Levitt, Paul R. 1975. General kin selection models for genetic evolution of sib altruism in diploid and haplodiploid species. *Proceedings of the National Academy of the Sciences USA* 72:4531-4535.

Lewens, Tim. 2004. *Organisms and artifacts: design in nature and elsewhere*. Cambridge, MA: MIT Press.

———. 2007a. Functions. In Mohan Matthen and Christopher Stephens (eds.), *Handbook of the philosophy of science: philosophy of biology*. Amsterdam: North Holland, pp. 525-549.

———. 2007b. Adaptation. In David L. Hull and Michael Ruse (eds.), *The Cambridge companion to the philosophy of biology*. Cambridge: Cambridge University Press, pp. 1-21.

———. 2009. Seven types of adaptationism. *Biology and philosophy* 24:161-182.

Lewis, David K. 1973. Causation. *Journal of Philosophy* 70:556-567.

———. 2000. Causation as influence. *Journal of Philosophy* 97:182-197.

Lloyd, Elisabeth A. 1988. *The structure and confirmation of evolutionary theory*. Westport, CT: Greenwood Press.

———. 2012. Units and levels of selection. In Edward Zalta (ed.), *The Stanford Encyclopedia of Philosophy (Spring 2012 edition)*. URL (accessed 11/12/2012): http://plato.stanford.edu/archives/spr2012/entries/selection-units/>

McGlothlin , Joel W., Allen J. Moore, Jason B. Wolf and Edmund D. Brodie III. 2010. Interacting phenotypes and the evolutionary process. III. Social evolution. *Evolution* 64:2558-2574.

Machamer, Peter, Lindley Darden and Carl F. Craver. 2000. Thinking about Mechanisms. *Philosophy of Science* 67:1-25.

Marshall, James A. R. 2011a. Group selection and kin selection: formally equivalent approaches. *Trends in Ecology and Evolution* 26:325-332.

———. 2011b. Queller's rule OK: comment on van Veelen, 'When inclusive fitness is right and when it can be wrong'. *Journal of Theoretical Biology* 270:185-188.

Martens, Johannes. 2011. Social evolution and strategic thinking. *Biology and Philosophy* 26:697-715.

Mas-Colell, Andreu, Michael D. Whinston and Jerry R. Green. 1995. *Microeconomic theory*. New York: Oxford University Press.

Maynard Smith, John. 1964. Group selection and kin selection. *Nature* 200:1145-1147.

———. 1976a. Letter to New Scientist. *New Scientist* 71:247.

———. 1976b. Group selection. *Quarterly Review of Biology* 21:20-29.

———. 1980. Models of the evolution of altruism. *Theoretical Population Biology* 18:151-159.

———. 1982. *Evolution and the theory of games*. Cambridge, Cambridge University Press.

———. 1983. Models of evolution. *Proceedings of the Royal Society B: Biological Sciences* 219:315-325.

———. 1987. How to model evolution. In J. Dupré (ed.), *The latest on the best: essays on evolution and optimality*. Cambridge, MA: MIT Press, 119-131.

———. 2002. Commentary on Kerr and Godfrey-Smith. *Biology and Philosophy* 17:523-527.

Maynard Smith, John and Eörs Szathmáry. 1995. *The major transitions in evolution*. Oxford: Oxford University Press.

Michod, Richard E. 1982. The theory of kin selection. *Annual Review of Ecology and Systematics* 13:23-55.

———. 2007. Evolution of individuality during the transition from unicellular to multicellular life. *Proceedings of the National Academy of Sciences USA* 104:8613-8618.

Michod, Richard E. and W. D. Hamilton. 1980. Coefficients of relatedness in socio-biology. *Nature* 288:694-697.

Mitchell, Robert W., Nicholas S. Thompson and H. Lyn Miles (eds.). 1997. *Anthropomorphism, anecdotes, and animals*. Albany, NY: SUNY Press.

Moore, Allen J., Edmund D. Brodie III and Jason B. Wolf. 1997. Interacting phenotypes and the evolutionary process: I. Direct and indirect genetic effects on social interactions. *Evolution* 51:1352-1362.

Moore, Harry D. M. and D. A. Taggart. 1995. Sperm pairing in the opossum increases the efficiency of sperm movement in a viscous environment. *Biology of Reproduction* 52:947-953.

Moore, Harry D. M., Katerina Dvoráková, Nicholas Jenkins and William Breed. 2002. Exceptional sperm cooperation in the wood mouse. *Nature* 418:174-177.

Moore, Tom and Harry D. Moore. 2002. Marsupial sperm pairing: a case of 'sticky' green beards? *Trends in Ecology and Evolution* 17:112-113.

Nanay, Bence. 2010. Population thinking as trope nominalism. *Synthese* 177:91-109.

Neander, Karen. 1995. Explaining complex adaptations: a reply to Sober's reply to Neander. *British Journal for the Philosophy of Science* 46:583-587.

Northcott, Robert. 2005. Pearson's wrong turning: against statistical measures of causal efficacy. *Philosophy of Science* 72:900-912.

Nowak, Martin A. 2006a. Five rules for the evolution of cooperation. *Science* 314:1560-1563.

———. 2006b. *Evolutionary dynamics: exploring the equations of life*. Cambridge, MA: Harvard University Press.

Nowak, Martin A. and Roger Highfield. 2011. *Supercooperators: evolution, altruism, and why we need each other to succeed*. New York: Free Press.

Nowak, Martin A., Corina E. Tarnita and Edward O. Wilson. 2010. The evolution of eusociality. *Nature* 466:1057-1062.

Okasha, Samir. 2001. Why won't the group selection controversy go away? *British Journal for the Philosophy of Science* 52:25-50.

———. 2002. Genetic relatedness and the evolution of altruism. *Philosophy of Science* 69:139-149.

———. 2003. The concept of group heritability. *Biology and Philosophy* 18:445-461.

———. 2004a. The 'averaging fallacy' and the levels of selection. *Biology and Philosophy* 19:167-184.

———. 2004b. Multi-level selection and the partitioning of covariance: a comparison of three approaches. *Evolution* 58:486-494.

———. 2004c. Multi-level selection, covariance and contextual analysis. *British Journal for the Philosophy of Science* 55:484-504.

———. 2005a. Altruism, group selection, and correlated interaction. *British Journal for the Philosophy of Science* 56:703-724.

———. 2005b. Multi-level selection and the major transitions in evolution. *Philosophy of Science* 72:1013-1028.

———. 2006. *Evolution and the levels of selection*. Oxford: Oxford University Press.

———. 2008. Fisher's 'fundamental theorem' of natural selection: a philosophical analysis. *British Journal for the Philosophy of Science* 59:319-351.

———. 2009. Individuals, groups, fitness and utility: multi-level selection meets social choice theory. *Biology and Philosophy* 24:561-584.

———. 2010. Replies to my critics. *Biology and Philosophy* 25:425-431.

———. 2011. Reply to Sober and Waters. *Philosophy and Phenomenological Research* 82: 241-248.

Okasha, Samir and Cedric Paternotte. 2012. Group adaptation, formal Darwinism and contextual analysis. *Journal of Evolutionary Biology* 25:1127-1139.

Oli, Madan K. 2003. Hamilton goes empirical: estimation of inclusive fitness from life-history data. *Proceedings of the Royal Society B: Biological Sciences* 270:307-311.

Orlove, M. J. 1975. A model of kin selection not invoking coefficients of relationship. *Journal of Theoretical Biology* 49:289–310.

———. A reconciliation of inclusive fitness and personal fitness approaches: a proposed correcting term for the inclusive fitness formula. *Journal of Theoretical Biology* 81:577-586.

Orzack, Steven Hecht and Elliott Sober. 1994. Optimality models and the test of adaptationism. *American Naturalist* 143:361-380.

——— (eds.). 2001. *Adaptationism and optimality*. Cambridge: Cambridge University Press.

Oster, George F. and Edward O. Wilson. 1978. *Caste and ecology in the social insects*. Princeton, NJ: Princeton University Press.

Payne, Robert B. 2005. *The cuckoos*. New York: Oxford University Press.

Pepper , John W. 2000. Relatedness in trait-group models of social evolution. *Journal of Theoretical Biology* 206:355-368.

Pigliucci, Massimo. 2001. *Phenotypic plasticity.* Baltimore, MD: Maryland University Press.

———. forthcoming. On the different ways of 'doing theory' in biology. *Biological Theory*.

Pizzari, Tommaso and Kevin R. Foster. 2008. Sperm sociality: cooperation, altruism and spite. *PLoS Biology* 6:e130.

Price, George R. 1970. Selection and covariance. *Nature* 227:520-1.

———. 1971. Extension of the Hardy-Weinberg law to assortative mating. *Annals of Human Genetics* 34:455-458.

———. 1972a. Extension of covariance selection mathematics. *Annals of Human Genetics* 35:485-90.

———. 1972b. Fisher's 'fundamental theorem' made clear. *Annals of Human Genetics* 36:129-140.

Pust, Joel. 2001. Natural selection explanation and origin essentialism. *Canadian Journal of Philosophy* 31:201-220.

———. 2004. Natural selection and the traits of individual organisms. *Biology and Philosophy* 19:765-779.

Queller, David C. 1985. Kinship, reciprocity, and synergism in the evolution of social behaviour. *Nature* 318:366-7.

———. 1989. Inclusive fitness in a nutshell. *Oxford Surveys in Evolutionary Biology* 6:73-109.

———. 1992a. Quantitative genetics, inclusive fitness and group selection. *American Naturalist* 139:540-58.

———. 1992b. A general model for kin selection. *Evolution* 46:376-80.

———. 1992c. Does population viscosity promote kin selection? *Trends in Ecology and Evolution* 7: 322-324.

———. 1994. Genetic relatedness in viscous populations. *Evolutionary Ecology* 8:70-73.

———. 1996. The measurement and meaning of inclusive fitness. *Animal Behaviour* 51: 229-232.

———. 1997. Cooperators since life began. *Quarterly Review of Biology* 72:184-188.

———. 2000. Relatedness and the fraternal major transitions. *Philosophical Transactions of the Royal Society B: Biological Sciences* 355:1647-1655.

———. 2011. Expanded social fitness and Hamilton's rule for kin, kith and kind. *Proceedings of the National Academy of Sciences* USA 108:10792-10799.

Queller, David C, Francesca Zacchi, Rita Cervo, Stefano Turillazzi, Michael T. Henshaw, Lorenzo A. Santorelli and Joan E. Strassmann. Unrelated helpers in a social insect. *Nature* 405:784-787.

Queller, David C. and Joan E. Strassmann. 1998. Kin selection and social insects. *BioScience* 48:165-175.

———. 2009. Beyond society: the evolution of organismality. *Philosophical Transactions of the Royal Society B* 355:1647-1655.

Rankin, D. J., E. P. C. Rocha and S. P. Brown. 2011a. What traits are carried on mobile genetic elements, and why? *Heredity* 106:1-10.

Rankin, D. J., S. E. Mc Ginty, T. Nogueira, M. Touchon, F. Taddei, E. P. C. Rocha and S. P. Brown. 2011b. Bacterial cooperation controlled by mobile genetic elements: kin selection and infectivity are part of the same process. *Heredity* 107:279-281.

Ratnieks, Francis L. W. and Carl Anderson. 1999. Task partitioning in insect societies. *Insectes Sociaux* 46:95-108.

Ratnieks, Francis L. W. and Tom Wenseleers 2008. Altruism in insect societies: voluntary or enforced? *Trends in Ecology and Evolution* 23:45-52.

Ratnieks, Francis L. W., Kevin R. Foster and Tom Wenseleers. 2006. Conflict resolution in insect societies. *Annual Review of Entomology* 51:581-608.

Rice, Sean H. 2004. *Evolutionary theory: mathematical and conceptual foundations*. Sunderland, MA: Sinauer.

Richerson, Peter J. and Robert Boyd. 2005. *Not by genes alone: how culture transformed human evolution.* Chicago, IL: University of Chicago Press.

Ridley, Mark and Alan Grafen. 1981. Are green beard genes outlaws? *Animal Behaviour* 29:944-955.

Rose, Michael R. and George V. Lauder. 1996. *Adaptation*. San Diego, CA: Academic Press.

Rosas, Alejandro. 2010. Beyond inclusive fitness? On a simple and general explanation for the evolution of altruism. *Philosophy and Theory in Biology* 2:e104.

Ross-Gillespie, Adin, Andy Gardner, Angus Buckling, Stuart A. West and Ashleigh S. Griffin. 2009. Density dependence and cooperation: theory and a test with bacteria. *Evolution* 63:2315-2325.

Rousset, François. 2004. *Genetic structure and selection in subdivided populations*. Princeton, NJ: Princeton University Press.

Sarkar, Sahotra. 2005. *Molecular models of life: philosophical papers on molecular biology*. Cambridge, MA: MIT Press.

Schlicting, Carl and Massimo Pigliucci. 1998. *Phenotypic evolution*. Sunderland, MA: Sinauer.

Sharpe, F. A. and L. M. Dill. 1997. The behavior of Pacific herring schools in response to artificial humpback whale bubbles. *Canadian Journal of Zoology* 75:725-730.

Skyrms, Brian. 2010. *Signals: evolution, learning and information*. Oxford: Oxford University Press.

Sober, Elliott. 1984. *The nature of selection: evolutionary theory in philosophical focus*. Chicago, IL: University of Chicago Press.

———. 1995. Natural selection and distributive explanation: a reply to Neander. *British Journal for the Philosophy of Science* 46:384-387.

———. 2008. *Evidence and evolution: the logic behind the science*. Cambridge: Cambridge University Press.

Sober, Elliott and David Sloan Wilson. 1994. A critical review of philosophical work on the units of selection problem. *Philosophy of Science* 61:534-555.

———. 1998. *Unto others: the evolution and psychology of unselfish behaviour*. Cambridge, MA: Harvard University Press.

smith, jeff [*sic*], J. David van Dyken and Peter C Zee. 2010. A generalization of Hamilton's rule for the evolution of microbial cooperation. *Science* 328:1700-1703.

Spirtes, Peter, Clark Glymour and Richard Scheines. 2000. *Causation, prediction and search (2nd edition)*. Cambridge, MA: MIT Press.

Stegmann, Ulrich. forthcoming. Causal control and genetic causation. *Noûs.*

Sterelny, Kim. 1996. The return of the group. *Philosophy of Science* 63:562-584.

Sterelny, Kim and Paul E. Griffiths. 1999. *Sex and death: an introduction to philosophy of biology*. Chicago, IL: University of Chicago Press.

Sterelny, Kim and Philip Kitcher. 1988. The return of the gene. *Journal of Philosophy* 85:339-361.

Strassmann, Joan E., Colin R. Hughes, David C. Queller, Stefano Turillazzi, Rita Servo, Scott K. Davis and Keith F. Goodnight. 1989. Genetic relatedness in primitively eusocial wasps. *Nature* 342:268-270.

Strassmann, Joan E. and David C. Queller. 2007. Insect societies as divided organisms: the complexities of purpose and cross-purpose. *Proceedings of the National Academy of Sciences USA* 104:8619-8626.

———. 2010. The social organism: congresses, parties and committees. *Evolution* 64:605-616.

———. 2011. Evolution of cooperation and control of cheating in a social microbe. *Proceedings of the National Academy of Sciences* USA 108:10855-10862.

Strassmann, Joan E., Yong Zhu and David C. Queller. 2000. Altruism and social cheating in the social amoeba Dictyostelium discoideum. *Nature* 408:965-967.

Taylor, Peter D. 1990. Allele frequency change in a class-structured population. *American Naturalist* 135:95-106

———. 1992. Altruism in viscous populations—an inclusive fitness model. *Evolution and Ecology* 6:352-356.

Taylor, Peter D., Geoff Wild and Andy Gardner. 2007. Direct fitness or inclusive fitness: how shall we model kin selection? *Journal of Evolutionary Biology* 20:301-309.

Taylor, Christine and Martin A. Nowak. 2007. Transforming the dilemma. *Evolution* 61:2281-92.

Toro, M., R. Abugov, B. Charlesworth and R. E. Michod. 1982. Exact *versus* heuristic models of kin selection. *Journal of Theoretical Biology* 97:699-713.

Traulsen, Arne. 2010. Mathematics of kin- and group-selection: formally equivalent? *Evolution* 64:316-323.

Traulsen, Arne and Martin A. Nowak. 2006. Evolution of cooperation by multilevel selection. *Proceedings of the National Academy of Sciences USA* 103:10952-10955.

Trivers, Robert L. (1985). *Social evolution*. Menlo Park, CA: Benjamin/Cummings.

Uyenoyama, Marcy K. and Marcus W. Feldman. 1980. Theories of kin and group selection: a population genetics perspective. *Theoretical Population Biology* 17:380-414.

———. 1981. On relatedness and adaptive topography in kin selection. *Theoretical Population Biology* 19:87-123.

———. 1982. Population genetic theory of kin selection II: the multiplicative model. *American Naturalist* 120:614-27.

Uyenoyama, Marcy K., Marcus W. Feldman and Laurence D. Mueller. 1981. Population genetic theory of kin selection: multiple alleles at one locus. *Proceedings of the National Academy of Sciences USA* 78:5036-5040.

van Veelen. 2005. On the use of the Price equation. *Journal of Theoretical Biology* 237:412-26.

———. 2009. Group selection, kin selection, altruism, and cooperation: when inclusive fitness is right and when it can be wrong. *Journal of Theoretical Biology* 259:589-600.

van Veelen, Matthijs, Julián Garcia, Maurice W. Sabelin and Marthijn Egas. 2010. Call for a return to rigour in models. *Nature* 467:661.

———. 2012. Group selection and inclusive fitness are not equivalent; the Price equation vs. models and statistics. *Journal of Theoretical Biology* 299:64-80.

Velicer, Gregory J. 2003. Social strife in the microbial world. *Trends in microbiology* 11:330-337.

Velicer, Gregory J. and Michiel Vos. 2009. Sociobiology of the myxobacteria. *Annual Review of Microbiology* 63:599-623.

Voland, Eckart. 1998. Evolutionary ecology of human reproduction. *Annual Review of Anthropology* 27:347-374.

Wade, Michael J. 1985. Soft selection, hard selection, kin selection and group selection. *American Naturalist* 125:61-73.

Waibel, Markus, Dario Floreano and Laurent Keller. A quantitative test of Hamilton's rule for the evolution of altruism. *PLoS Biology* 9:e1000615.

Walsh, D. M. 1998. The scope of selection: Sober and Neander on what natural selection explains. *Australasian Journal of Philosophy* 76:250-264.

Waters, C. Kenneth. 2007. Causes that make a difference. *Journal of Philosophy* 104:551-579.

———. 2011. Okasha's unintended argument for toolbox theorizing. *Philosophy and Phenomenological Research* 82:232-240.

Waters, Christopher M. and Bonnie L. Bassler. 2005. Quorum sensing: cell-to-cell communication in bacteria. *Annual Review of Cell and Developmental Biology* 21:319-346.

Weber, Marcel. 2006. The central dogma as a thesis of causal specificity. *History and Philosophy of the Life Sciences* 28:595-610.

Wenseleers, T., H. Helanterä, A. Hart and F. L. W. Ratnieks. 2004. Worker reproduction and policing in insect societies: an ESS analysis. *Journal of Evolutionary Biology* 17:1035-1047.

Wenseleers, Tom, Andy Gardner and Kevin R. Foster. 2010. Social evolution theory: a review of methods and approaches. In T. Székely, A. J. Moore and J. Komdeur (eds), *Social behaviour: genes, ecology and evolution*. Cambridge: Cambridge University Press, 132-58.

West, Stuart A., Stephen P. Diggle, Angus Buckling, Andy Gardner and Ashleigh S. Griffin. 2007b. The social lives of microbes. *Annual Review of Ecology, Evolution and Systematics* 38:53-77.

West, Stuart A. and Andy Gardner. 2010. Altruism, spite and greenbeards. *Science* 327:1341-1344.

West, Stuart A., Ashleigh S. Griffin and Andy Gardner. 2007a. Social semantics: altruism, cooperation, mutualism, strong reciprocity and group selection. *Journal of Evolutionary Biology* 20:415-432.

———. 2008. Social semantics: how useful has group selection been? *Journal of Evolutionary Biology* 21:374-383.

West, Stuart A., Claire El Mouden and Andy Gardner. 2011. Sixteen common misconceptions about the evolution of cooperation in humans. *Evolution and Human Behavior* 32:231-262.

West, Stuart A., Ido Pen and Ashleigh S. Griffin. 2002. Cooperation and competition between relatives. *Science* 296:72-75.

West-Eberhard, Mary Jane. 2003. *Developmental plasticity and evolution*. New York: Oxford University Press.

Westneat, David and Charles Fox (eds.). 2010. *Evolutionary Behavioral Ecology*. New York: Oxford University Press.

Wild, Geoff and Peter D. Taylor. 2006. The economics of altruism and cooperation in class-structured populations: what's in a cost? What's in a benefit? *Journal of Evolutionary Biology* 19:1423-1425.

Williams, George C. 1966. *Adaptation and natural selection: a critique of some current evolutionary thought*. Princeton, NJ: Princeton University Press.

Williams, Paul, Klaus Winzer, Weng C. Chan and Miguel Camara. 2007. Look who's talking: communication and quorum sensing in the bacterial world. *Philosophical Transactions of the Royal Society B* 362:1119-1134.

Wilson, David Sloan. 1975. A theory of group selection. *Proceedings of the National Academy of Sciences USA* 72:143-146.

———. 2008. Social semantics: toward a genuine pluralism in the study of social behaviour. *Journal of Evolutionary Biology* 21:368-373.

Wilson, David Sloan and Lee A. Dugatkin. 1997. Group selection and assortative interactions. *American Naturalist* 149:336-351.

Wilson, Edward O. 1971. *The insect societies*. Cambridge, MA: Harvard University Press.

———. 2012. *The social conquest of Earth*. New York: W. W. Norton & Company.

Wilson, Robert A. 2003. Pluralism, entwinement, and the levels of selection. *Philosophy of Science* 70:531-552.

———. 2005. *Genes and the agents of life: the individual in the fragile sciences: biology*. Cambridge: Cambridge University Press.

Wimsatt , William. 1981. Units of selection and the structure of the multi-level genome. *Proceedings of the Philosophy of Science Association* 2:122–183.

Wolf, Jason B., Edward D. Brodie III, James M. Cheverud, Allen J. Moore and Michael J. Wade. 1998. Evolutionary consequences of indirect genetic effects. *Trends in Ecology and Evolution* 13:64-69.

Woodward, James. 2010. Causation in biology: stability, specificity, and the choice of levels of explanation. *Biology and Philosophy* 25:287-318.

Wynne-Edwards, V. C. 1962. *Animal dispersion in relation to social behaviour*. London: Oliver and Boyd.