JOURNAL OF Evolutionary Biology

doi: 10.1111/j.1420-9101.2011.02310.x

A formal theory of the selfish gene

A. GARDNER*† & J. J. WELCH‡§

*Department of Zoology, University of Oxford, Oxford, UK
†Balliol College, University of Oxford, Oxford, UK
‡Institut des Sciences de l'Evolution, Université Montpellier 2, Montpellier, France
§Department of Genetics, University of Cambridge, Cambridge, UK

Keywords:

adaptation;
epistasis;
formal Darwinism;
gene as maximizing agent;
inclusive fitness;
optimization program.

Abstract

Adaptation is conventionally regarded as occurring at the level of the individual organism. In contrast, the theory of the selfish gene proposes that it is more correct to view adaptation as occurring at the level of the gene. This view has received much popular attention, yet has enjoyed only limited uptake in the primary research literature. Indeed, the idea of ascribing goals and strategies to genes has been highly controversial. Here, we develop a formal theory of the selfish gene, using optimization theory to capture the analogy of 'gene as fitness-maximizing agent' in mathematical terms. We provide formal justification for this view of adaptation by deriving mathematical correspondences that translate the optimization formalism into dynamical population genetics. We show that in the context of social interactions between genes, it is the gene's inclusive fitness that provides the appropriate maximand. Hence, genic selection can drive the evolution of altruistic genes. Finally, we use the formalism to assess the various criticisms that have been levelled at the theory of the selfish gene, dispelling some and strengthening others.

> If we allow ourselves the license of talking about genes as if they had conscious aims, always reassuring ourselves that we could translate our sloppy language back into respectable terms if we wanted to, we can ask the question, what is a single selfish gene trying to do? Dawkins (1976, p. 88)

Introduction

The cardinal problem of biology is to explain the process and purpose of adaptation, i.e. the apparent design of the living world (Paley, 1802; Williams, 1966; Leigh, 1971; Gardner, 2009). The conventional view of adaptation is that this is a property of individual organisms, that owes to the action of natural selection, and that functions to maximize the organism's (inclusive) fitness (Darwin, 1859; Hamilton, 1963, 1964, 1970; Grafen, 2002, 2006a). This paradigm underlies an enormously successful body of research within the biological sciences, from functional anatomy to behavioural ecology.

Correspondence: Andy Gardner, Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK.
Tel.: +44 1865 271271; fax: +44 1865 271249;
e-mail: andy.gardner@zoo.ox.ac.uk

Alternative views of adaptation have also been advocated. Most prominent is the 'selfish gene' view of Dawkins (1976, 1978, 1982; see also Hamilton, 1972), which claims that adaptation is properly located at the level of the gene, and which regards the unity of the individual organism as an illusion, stemming from a transient alliance between otherwise-warring genes. This idea has received enormous popular attention, but it has enjoyed only limited uptake in the primary research literature. Although it is intended to apply to all genes, very often it has dislodged the organism-centred view only where the integrity of the organism has been breached by Mendelian outlawry or genomic imprinting (reviewed by Burt & Trivers, 2006). Moreover, in some quarters, the gene's eye view has suffered heavy and sustained criticism from its conception (Langley, 1977; Lewontin, 1977; Stent, 1977; Wade, 1978; Bateson, 1978, 2006; Midgley, 1979; Daly, 1980; Sober & Lewontin, 1982; Mayr, 1983; Hampe & Morgan, 1988; Williams, 1992; Dover, 2000; Hull, 2001; Lloyd, 2001, 2005; Gould, 2002; Charlesworth, 2006; Okasha, 2006 Ch. 5; Godfrey-Smith, 2009; Noble, 2011). Whereas some of these critics defend an organism-centred approach to adaptation,

others are not concerned with questions of adaptation at all. Table 1 provides a catalogue of the substantive criticisms, which are discussed and developed later in this article.

A major obstacle to the acceptance of the gene's eye view is the lack of formal theory on this topic. The gold standard for evolutionary theory is dynamical analysis of the genetics of populations. In contrast, the verbal theory of the selfish gene is framed in terms of the gene's apparent desires and conflicts of interest. Such intentional language is anathema to many population geneticists, and it has been suggested that formal theory must expunge the notion of purpose altogether (Daly, 1980; Charlesworth, 2006; Godfrey-Smith, 2009). However, to do this would be to reject the entire basis of the gene's eye view of adaptation.

In this article, we take an alternative route, by formalizing the theory of the selfish gene in the manner of Grafen's (2002, 2006a, 2007, 2009; see also Gardner & Grafen, 2009) 'formal Darwinism' project. This project draws upon the earlier use of optimization thinking in behavioural ecology (Maynard Smith, 1982 Ch. 1, 1987; Charlesworth, 1990; Parker & Maynard Smith, 1990), by expressing ideas of intention and purpose in the formal language of optimization theory. Here, we use optimization theory to develop a 'gene as fitness-maximizing agent' analogy. We relate this analogy to a dynamical model of genic selection, deriving mathematical correspondences that demonstrate how to translate between optimization (purpose) and dynamical (process) statements. We show that the selfish-gene analogy holds in models with no social interaction between genes. When social interaction is permitted (i.e. when genes impact on each others' fitness), we show that the appropriate maximand for the gene is its inclusive fitness, and hence, genic selection may favour altruistic genes. By forcing us to be explicit about assumptions, our formalism finally allows us to assess the various criticisms that have been levelled at the theory of selfish genes.

A formal theory of the selfish gene

Population genetics

The proper basis of evolutionary biology is population genetics, so any formal theory of selfish genes must make reference to the genetics of populations. Here, we develop the dynamical aspects of the theory using standard population genetic principles and assumptions. We assume a very large, but finite, population of gene positions. These are places where physical portions of genetic material – hereafter, 'genes' – occur. For example,

Table 1 A catalogue of substantive criticisms of the selfish-gene concept. Note that we neglect many easily rebutted criticisms of the 'gene's eye view', e.g. those concerning supposed 'genetic determinism' or denial of human agency, that were prominent in early reviews of *The Selfish Gene* (Dawkins, 1976), but are not relevant to assessing a gene-centred view of adaptation (Dawkins, 1976, 1982, Langley, 1977; Lewontin 1977; Midgley, 1979, 1983; Stent, 1977).

No.	Difficulty	References	Comments
1	Genes do not have intentions	Langley, 1977; Stent, 1977; Midgley, 1979, 1983; Daly, 1980; Charlesworth, 2006	The selfish gene is an analogy, which concerns only one component of the evolutionary dynamics. These correspondences explain the presence of adaptation, or apparent design in nature, but do not imply that optimality will obtain
2	Who is the gene?	Stent, 1977; Daly, 1980; Williams, 1992; Hull, 2001; Lloyd, 2001, 2005; Bateson, 2006	The scrap of nucleic acid residing at a gene position can be usefully attributed with agency. The allele (i.e. the type rather than the token), whilst a 'beneficiary' of the evolutionary process, cannot be considered an intentional agent
3	Genes do not always behave selfishly	Burt & Trivers, 2006; West *et al.*, in press	With social interactions, genes act to maximize their inclusive fitness. Selection can therefore lead to altruistic and spiteful genes. Genes play a fundamental role in theories of adaptation, but this does not imply that they will behave selfishly
4	Nonadditive gene interactions render the gene's eye view useless	Langley, 1977; Sober & Lewontin, 1982; Mayr, 1983; Dover, 2000 Gould, 2002; Okasha, 2006	The action of selection – at any level – is a function of additive effects only. Genes act to maximize inclusive fitness, defined as a function of additive effects; this does not require that nonadditive effects are absent or unimportant to the evolutionary outcome
5	Phenotypes may be determined by multiple genes interacting with the environment in a complex way	Langley, 1977; Lewontin, 1977; Bateson, 1978, 2006; Daly, 1980; Mayr, 1983; Dover, 2000; Gould, 2002; Noble, 2011	Adaptive phenotypes must be characters that are under the sole control of the putative agent. This places severe limits upon what can be considered a gene-level adaptation
6	Selfish-gene theory obscures the causes of evolution	Sober & Lewontin, 1982; Dover, 2000; Gould, 2002; Lloyd, 2005; Okasha, 2006	The theory of selfish genes swaps proximate (mechanism) causes for ultimate (purpose) causes

if we consider two individuals, each containing three diploid cells with four loci per haploid genome, we have $2 \times 3 \times 2 \times 4 = 48$ gene positions. For simplicity, we assume that all gene positions are equivalent (genes do not belong to separate 'classes'; Grafen, 2006b) and that each is occupied by a single gene. Put another way, there are no intrinsic differences between gene positions, aside from the alleles that occupy them. We assume that only a finite number of allelic variants are possible. We assume discrete (although potentially overlapping) generations and also that the number of gene positions remains fixed at N. We assign every gene position (and hence also its occupant gene) in a focal generation a unique index $i \in I$, and for the purpose of computing population statistics, we give every gene position an equal weighting $1/N$. Thus, for any quantity x that varies across genes positions, the arithmetic average of this quantity over the population is $\mathrm{E}_I(x) = \sum_I x_i/N$. Notation is summarized in Table 2.

The fitness of gene i in the current generation is defined as the number of gene positions in the subsequent generation that receive their genetic material from gene position i. This captures both the physical survival of genetic material between generations and also the synthesis of new genetic material (replication). We allow for stochastic fitness effects: owing to finite gene positions and finite allelic variants, there are a finite number of possible states in which the next generation can manifest. We assign each possible outcome a unique index $\omega \in \Omega$, and the probability of this outcome is q^ω. The fitness of gene i under outcome ω is w_i^ω, and the average of this quantity over uncertainty is $w_i = \sum_\Omega q^\omega w_i^\omega$. Owing to the assumption of fixed population size, the average fitness of all genes is $\mathrm{E}_I(w^\omega) = \mathrm{E}_I(w) = 1$.

We assign every gene an 'allele' $a_i \in A$, according to its nucleic acid sequence, where A is the set of all possible alleles. Next, we assign every gene a numerical 'genic value' according to its allele, i.e. $g_i = \mathscr{G}(a_i)$, where $\mathscr{G}$ is the

Table 2 A summary of notation used in the evolutionary models and optimization programs.

Evolutionary model			Optimization programme
With and without social interactions:			
Number of genes/gene positions	N	N	Number of agents
Gene/gene-position index	i	i	Agent index
Set of gene/gene-position indices	I	I	Set of agent indices
Reproductive outcome	ω	–	–
Set of all possible outcomes	Ω	–	–
Probability of outcome ω	q^ω	–	–
Fitness of gene i under outcome ω	w_i^ω	–	–
Expected fitness of gene i	$w_i = \sum_\Omega q^\omega w_i^\omega$	–	–
Genic value of gene i	g_i	–	–
Average genic value of gene i's descendants under outcome ω	$g_i'^\omega = g_i + \Delta g_i^\omega$	–	–
Allelic state of gene i	a_i	–	–
Set of alleles	A	–	–
Genotype function	$\mathscr{G}(a)$	–	–
Phenotype of gene i	π_i	s_i	Strategy of agent i
Set of all phenotypes	P	S	Strategy set
Phenotype function	$\mathscr{P}(a)$	–	–
Without social interactions:			
Fitness function	$\mathscr{W}(\pi)$	$\mathscr{U}(s)$	Objective function
With social interactions:			
Unordered list of all phenotypes in the population	Π	$\wp$	Context parameter
Role index	j	–	–
Set of all role indices	J	–	–
Phenotype of role-j social partner of gene i	π_{ij}	–	–
Ordered list of phenotypes belonging to gene i's social set	$\tilde{\pi}_i$	–	–
Fitness function in context Π	$\mathscr{W}(\tilde{\pi};\Pi) = \mathscr{W}_\mathrm{B}(\Pi) + \sum_J \mathscr{W}_\mathrm{A}(\pi_j,j;\Pi) + \mathscr{W}_\mathrm{N}(\tilde{\pi};\Pi)$	–	–
Baseline fitness in context Π	$\mathscr{W}_\mathrm{B}(\Pi) = 1$	–	–
Additive effect of role-j social partner phenotype π upon personal fitness in context Π	$\mathscr{W}_\mathrm{A}(\pi_j,j;\Pi)$	–	–
Nonadditive effect of social partner phenotypes $\tilde{\pi}$ upon personal fitness in context Π	$\mathscr{W}_\mathrm{N}(\tilde{\pi};\Pi)$	–	–
Coefficient of relatedness between focal gene and its role-j social partner	$r_j = \mathrm{cov}_I(g_j,g)/\mathrm{cov}_I(g,g)$	–	–
Inclusive fitness in context Π	$\mathscr{H}(\pi;\Pi) = \mathscr{W}_\mathrm{B}(\Pi) + \sum_J \mathscr{W}_\mathrm{A}(\pi,j;\Pi) r_j$	$\mathscr{U}(s;\wp)$	Objective function in context $\wp$

'genotype function'. The assignment of genic values to alleles is arbitrary. For example, we may choose to assign a focal allele a genic value of 1 and all other alleles a genic value of 0, such that the average genic value corresponds to the population frequency of the focal allele. Alternatively, the assignment of genic values to alleles might be with reference to their phenotypic effects (i.e. average effects; Fisher, 1918). We denote the average genic value of the descendants of gene i in the next generation (under outcome ω) as $g_i'^{\omega} = g_i + \Delta g_i^{\omega}$. Change in genic value between parent and offspring genes ($\Delta g_i^{\omega} \neq 0$) may occur, owing to factors such as spontaneous mutation (i.e. change in a) or generational differences in the average effect of an allele (i.e. changes to $\mathscr{G}$; Fisher, 1941).

Genic selection

Selection is the part of the evolutionary process that gives rise to adaptation. Here, we give a formal account of selection in the model developed above, using Price's (1970, 1972) equation to describe the population genetic change that is driven by the differential fitness of genes. We express the change in average genic value between consecutive generations of the population as:

$$\Delta \mathrm{E}_I(g)^{\omega} = \mathrm{cov}_I(w^{\omega}, g) + \mathrm{E}_I(w^{\omega}\Delta g^{\omega}) \qquad (1)$$

The first term on the RHS of eqn 1 is the covariance, taken across all gene positions in the population, of a gene's fitness and its genic value, and this defines the action of 'selection' at the gene level. The second term on the RHS of eqn 1 is the expectation, taken across all gene positions in the population, of the product of a gene's fitness and the difference between its own genic value and that of its offspring, and this defines the 'transmission' effect. Transmission includes factors such as mutation (i.e. change in allelic variant) and change in the average effect of a gene between generations, and there is no reason to suspect that it is negligible (to be discussed later in this article).

Our aim is to determine whether selection (not evolution as a whole) leads to the appearance of agency at the level of the gene. Hence, in what follows, we focus exclusively upon the first term on the RHS of eqn 1, i.e. the change that is ascribed to selection. This is calculated as

$$\Delta_{\mathrm{S}} \mathrm{E}_I(g)^{\omega} = \mathrm{cov}_I(w^{\omega}, g) \qquad (2)$$

This expression still includes the impact of stochastic fitness effects, i.e. random drift. However, averaging over uncertainty eliminates such effects, giving an expression for the systematic selection of genes, hereafter 'genic selection' (cf. Grafen, 2000 and Gardner & Grafen, 2009):

$$\mathrm{E}_{\Omega}(\Delta_{\mathrm{S}} \mathrm{E}_I(g)) = \mathrm{cov}_I(w, g) \qquad (3)$$

This formalism for genic selection, mediated by fitness differences between genes, is analogous to expressions for natural selection, mediated by fitness differences between individual organisms (Price, 1970; see also Robertson, 1966, 1968). However, gene fitness may be literally different from organism fitness: it represents the gene's success in gaining gene positions in the subsequent generation, which may or may not involve the improvement of the carrier organism's reproductive success (or, indeed, fitness effects at the group level). For example, the formalism allows for gene conversion, transposition and meiotic drive: mechanisms that potentially incur fitness costs for the organism (Burt & Trivers, 2006) and that are traditionally regarded as accruing to the 'transmission' component of evolutionary change (Price, 1970).

Adaptation and the optimization program

A dynamical, population-genetic analysis provides the proper basis for evolutionary theory. However, at the core of the concept of adaptation is the idea that phenotypes manifest apparent purpose, function and design: concepts that are alien to population genetics (Gardner, 2009). A formal language of purpose is provided by optimization theory, where such ideas are captured in the form of an optimization program, e.g.

$$\begin{array}{cc} s & \max \quad \mathscr{U}(s) \\ & s \in S \end{array} \qquad (4)$$

The optimization program (4) describes an (implicit) agent with an agenda and an instrument to be employed in the pursuit of its agenda. Specifically, the agent has a set of strategies S available to it (i.e. ways in which it may wield the instrument), and each strategy $s \in S$ assigned a corresponding real value by the objective function $\mathscr{U}(s)$, according to how well the strategy performs in the pursuit of the agenda (the larger the value, the closer the agent is to having realized its objective).

Thus, the notion of purpose is expressed as a maximization problem: the agent seeks the strategy that will maximize its objective function (von Neumann & Morgenstern, 1944). This leads to a formal definition of an optimum. An optimal strategy is one that (i) belongs to the strategy set and (ii) when given as input to the objective function, returns an output that is equal to or greater than that of all other strategies in the set. Formally, an optimal strategy is s^* where $\mathscr{U}(s^*) \geq \mathscr{U}(s)$ $\forall\, s \in S$. Conversely, a suboptimal strategy is one that (i) belongs to the strategy set and (ii) returns a lower output from the objective function than at least one other member of the set. Formally, a suboptimal strategy is s° where $\exists\, s \in S{:}\mathscr{U}(s^\circ) < \mathscr{U}(s)$. Importantly, although the optimization program provides a formal definition for optimality, it does not imply that optimality obtains – i.e. it sets a maximization problem, which may or may not be solved by the agent. This formalization therefore captures two important ideas about adaptation that were emphasized by Paley (1802): biological adaptations show contrivance with respect to some end, but this does not

imply perfection, or even optimality within constraints (Parker & Maynard Smith, 1990; Gardner, 2009).

Gene as fitness-maximizing agent

The traditional view of adaptation at the level of the individual (Darwin, 1859; Hamilton, 1964, 1970, 1996 Ch. 2) has been formalized by identifying the agent described in an optimization program with an individual organism (Grafen, 2002, 2006a, 2007, 2009). Here, we are interested in forming a 'gene as maximizing agent' analogy, i.e. the idea that the gene is a purposeful agent with an agenda. Hence, the first step in forming the analogy is to identify the gene as the agent, which we do by assigning every agent the index i of the corresponding gene.

Second, we choose a phenotype, associated with the gene, to be the agent's instrument. Note that an instrument must be under the sole control of the corresponding agent. Hence, in order for the phenotype to fulfil this role, we must consider only phenotypes that are fully determined by the allele encoded by the corresponding gene. We denote the phenotype associated with this gene position by $\pi_i = \mathscr{P}(a_i)$, where $\mathscr{P}$ is the 'phenotype function' that relates allele to phenotype. Because there are a finite number of possible alleles, there are a finite set of possible phenotypes, which we denote P. With this notation in place, we may identify the phenotype as the gene's strategy, i.e. $\pi_i \leftrightarrow s_i$, and the set of possible phenotypes as the strategy set, i.e. $P \leftrightarrow S$.

To complete the analogy, we must to identify a candidate for the objective function $\mathscr{U}$, which the agent aims to maximize. We choose the gene's expected reproductive success, and in this section of the article, we assume that it is given by a very simple function of the gene's phenotype, i.e.

$$w_i = \mathscr{W}(\pi_i)/\mathrm{E}_I(\mathscr{W}(\pi)) \tag{5}$$

where $\mathscr{W}$ is the 'fitness function'. Equation 5 assumes that the phenotypes associated with other genes do not impact upon the focal gene's expected fitness, except for a density-dependent scaling effect that is a necessary consequence of the assumption of fixed population size. This strong and unrealistic assumption of no social interaction between genes is relaxed in the next section of this article. With this notation in place, we may identify the fitness function with the objective function, i.e. $\mathscr{W} \leftrightarrow \mathscr{U}$. Thus, the idea that the gene wields its phenotype as a means of achieving personal fitness is formally captured as an optimization program:

$$\begin{array}{ll} \pi & \max \quad \mathscr{W}(\pi) \\ & \pi \in P \end{array} \tag{6}$$

This provides a formal statement of what we mean when we say that a gene is responsible for a phenotype and that the function of the phenotype is to maximize the gene's fitness. It provides a rigorous basis for forming statements about the optimality of phenotypes. In particular, an optimal phenotype is one that (i) belongs to the set of possible phenotypes and (ii) achieves an expected fitness that is equal to or greater than that of any other phenotype in the set. Formally, an optimal phenotype is π^* where $\mathscr{W}(\pi^*) \geq \mathscr{W}(\pi) \ \forall \ \pi \in P$. Conversely, a suboptimal phenotype is one that (i) belongs to the set of possible phenotypes and (ii) achieves lower expected fitness than at least one other member of the set. Formally, a suboptimal phenotype is π° where $\exists \ \pi \in P: \mathscr{W}(\pi^\circ) < \mathscr{W}(\pi)$. Importantly, although the gene as maximizing agent analogy (6) provides a formal definition for phenotype optimality, it does not imply that optimality obtains – i.e. it defines the function of the phenotype, without stating that the gene achieves maximum fitness.

Formal justification for the selfish gene

We have developed an evolutionary model that describes how selection acting upon genes drives genetic change of populations – summarized by dynamical equation (3). We have also developed a formal account of what it means to say that a gene is striving to maximize its fitness – summarized by optimization program (6). Here, we seek formal justification for the selfish-gene view, by establishing mathematical correspondences between the equations of motion (genetic change) and the calculus of purpose (genes maximizing their fitness). The translation of purpose into dynamics reveals that the optimization view recovers the results of a population-genetic analysis. The translation of dynamics into purpose captures the way in which genic selection gives rise to the emergence of apparently selfish genes.

We derive six correspondences between our dynamical and purposeful accounts of genic adaptation (Table 3; derivations given in Appendix). Correspondences I–V translate gene-optimization scenarios into population-genetic results. Specifically, if all agents described in the optimization view are behaving optimally, then this corresponds to a scenario in the population-genetic model where there is no selection with respect to any genic value (no 'scope for selection'; I), and where no allele in the permitted set can be introduced to the population, which will increase in frequency from rarity under the action of genic selection (no 'potential for positive selection'; II); if all agents are suboptimal, but equally so, then there is no scope for selection (III) but there is potential for positive selection, i.e. there exists an allele that, if introduced into the population, is expected to increase in frequency from rarity, under the action of genic selection (IV); and if agents vary in their optimality, then there is scope for selection, and the selective change in the average of every genic value and in every allele frequency is given by its covariance, across all gene positions, with relative attained maximand (i.e. relative fitness; V). Arising from these five correspondences is a further correspondence (VI), translating in the opposite

Table 3 Correspondences between dynamical and purposeful accounts of gene-level adaptation. Formal justification for the 'gene as fitness maximizing agent' view rests upon the ability to translate this way of thinking into formal population genetics, and vice versa.

Numeral	Correspondence
I	If all agents are optimal, there is no scope for selection (no expected change in the average of any genic value)
II	If all agents are optimal, there is no potential for positive selection (no introduced allele can increase from rarity due to selection)
III	If all agents are suboptimal, but equally so, there is no scope for selection (no expected change in the average of any genic value)
IV	If all agents are suboptimal, but equally so, there is potential for positive selection (there exists an allele that, if introduced, can increase from rarity due to selection)
V	If agents vary in their optimality, there is scope for selection, and change in the average of all genic values, and in all gene frequencies, is given by their covariance with relative attained maximand
VI	If there is neither scope for selection nor potential for positive selection, all agents are optimal

direction, which states that if there is neither scope for selection nor potential for positive selection, then all agents must be behaving optimally.

These correspondences are analogous to those derived by Grafen (2002, 2006a), for the 'individual as maximizing agent' analogy, that justify seeing organisms as economic agents, and by Gardner & Grafen (2009), for the 'clonal group as maximizing agent' analogy, that do the same for colonies of clonal organisms. We have shown that genes also formally qualify as adaptive agents, insofar as our simplified biological model can be regarded as painting an accurate picture of the real world. One major limitation of the model is that it assumes that genes do not impact upon each other's fitness, aside from a density-dependent scaling effect. The next section relaxes this assumption.

A formal theory of the social gene

Phenotype, personal fitness and inclusive fitness

In this section, we allow for a more realistic link between phenotypes and fitness. In particular, we permit social interactions between different genes. We assume that the fitness of a focal gene is dependent upon the phenotypes expressed by a given 'social set' of genes, potentially including the focal gene itself. We assume that this network of socially interacting gene positions is established prior to the assignment of alleles to these gene positions. (However, once populated with genes, the genic and allelic values of socially interacting gene positions may be correlated.) In addition, we assume that the social network appears identical from the perspective of every gene position, such that there are no intrinsic differences between gene positions, aside from the alleles that they – and their social partners – carry.

The impact upon fitness of a social partner's phenotype may depend upon the 'role' within which this social partner is acting (Grafen, 2006a). For example, the fitness effect of the phenotype of a social partner in the role of 'homologous gene in the same diploid cell' may be different from the fitness effect of the same phenotype expressed by a social partner in the role of 'homologous gene in a neighbouring cell' or, indeed, in the role of 'self'. We define a set of roles J, and we assume that for every focal recipient gene, there is one and only one gene in each of the actor roles $j \in J$. Moreover, we assume that every individual occurs once and only once in each role $j \in J$. Thus, we identify the role-j actor gene that mediates the fitness of recipient gene i with a subscripted ij. For example, the phenotype of this actor is π_{ij}. The ordered list of phenotypes expressed by genes in all the actor roles with respect to a focal recipient gene i is $\tilde{\pi}_i$. Note that roles need not be reciprocal: if a gene acts in a given role so as to mediate the fitness of another gene, then the latter need not act in the same role to mediate the fitness of the former. Indeed that the former is a member of the latter's social set does not imply that the latter is a member of the former's social set.

We also allow the overall phenotypic composition of the population to mediate a gene's fitness. This allows for 'playing the field' interactions, *sensu* Maynard Smith (1982, p. 23), in addition to the social interaction with explicit partners considered above. We compile an unordered list of all the phenotypes expressed in the population (one entry for every gene) and denote this Π. Thus, we may define a new fitness function, such that

$$w_i = \mathscr{W}(\tilde{\pi}_i; \Pi) \tag{7}$$

Without the loss of generality, we can decompose personal fitness into a sum of separate effects: $\mathscr{W}(\tilde{\pi}_i;\Pi) = \mathscr{W}_{\mathrm{B}}(\Pi) + \sum_J \mathscr{W}_{\mathrm{A}}(\pi_{ij},j;\Pi) + \mathscr{W}_{\mathrm{N}}(\tilde{\pi}_i;\Pi)$, where subscripts B, A and N denote baseline fitness (a function of population composition only), the additive effect of the phenotype of social partner of given role upon fitness in the context of the given population composition and the residual (nonadditive) effect of social partner phenotypes on fitness in the context of the given population composition, respectively (see Appendix for details).

In addition to personal fitness, we may also compute a focal gene's inclusive fitness by (i) re-assigning all additive fitness effects to the actor (rather than to the recipient, as was done above); (ii) weighting each additive fitness component by the appropriate coefficient of relatedness, defined as $r_j = \mathrm{cov}_I(g_j,g)/\mathrm{cov}_I(g, g)$, which is equal to the slope of the least-squares linear regression of role-j social partner genic value against a gene's own genic value (Orlove & Wood, 1978; Queller, 1992; Frank,

1998) and (iii) neglecting all nonadditive social effects. This yields

$$\mathscr{H}(\pi_i;\Pi) = \mathscr{W}_{\mathrm{B}}(\Pi) + \Sigma_J \mathscr{W}_{\mathrm{A}}(\pi_i, j;\Pi) r_j \qquad (8)$$

Equation 8 is analogous to the definition of inclusive fitness for individual organisms (Hamilton, 1964, 1970; Grafen, 2006a; Gardner *et al.*, 2011). Here, we have made clear that the neglecting of nonadditive effects takes place in the computation of inclusive fitness, not in the construction of the underlying evolutionary model. That is, we do allow for arbitrary nonadditive interactions between genes in our model. Note that nonadditive effects were assumed absent in Grafen's (2006a) model of inclusive fitness–maximizing organisms.

Classification of the social behaviours of individual organisms is made according to the sign of the additive fitness impact on actor and recipient: +/+ behaviours are mutually beneficial, +/− behaviours are selfish, −/+ behaviours are altruistic and −/− behaviours are spiteful (Hamilton, 1964, 1970; West *et al.*, 2007). This scheme readily extends to the social behaviours of genes. Assuming two roles, 'self' ($j = 1$) and 'social partner' ($j = 2$), the additive impact upon actor fitness is $\mathscr{W}_{\mathrm{A}}(\pi_i,1;\Pi)$ and the additive impact upon recipient fitness is $\mathscr{W}_{\mathrm{A}}(\pi_i,2;\Pi)$. Because each of these quantities can take positive or negative values, a gene's behaviour can be classified as either mutually beneficial (+/+), selfish (+/−), altruistic (−/+) or spiteful (−/−). Behaviours that have fitness consequences for further recipients are more difficult to classify in this simple pairwise scheme. However, this ambiguity also exists in the classification of organismal social behaviours (West & Gardner, 2010).

Gene as inclusive fitness–maximizing agent

In the previous section, which excluded social interaction, a formal analogy of gene as maximizing agent was developed in which the (personal) fitness function $\mathscr{W}$ was identified as the objective function $\mathscr{U}$. The input of the fitness function was the gene's phenotype π, and this was identified as the strategy s that provides the input for the objective function.

The incorporation of social interactions into the model raises some difficulties for this approach. First, the personal fitness function takes as one of its arguments the phenotypic composition of the entire population, Π. This cannot be considered analogous to an agent's strategy in the gene's eye view, as it cannot be regarded as under the control of a single gene. Our solution is to make the objective function explicitly dependent upon a global context parameter – i.e. $\mathscr{U}(s;\wp)$ – and to identify the phenotypic composition of the population with this context parameter: $\Pi \leftrightarrow \wp$. Thus, the optimization problem is now considered to be conditional upon the (shared) context within which the agents find themselves. Future work to tackle game-theoretic aspects of the gene's eye view will need to make the dependence of the context parameter upon the strategies of other maximizing agents more explicit within the optimization program itself. However, this is beyond the scope of the present article.

Second, the personal fitness function takes as its other argument the full list of phenotypes $\tilde{\pi}_i$ expressed by those genes that mediate the focal gene's fitness (the social set). This cannot be considered analogous to the strategy of a single agent in the gene's eye view and nor can it be considered a global context parameter, as it is expected to vary from gene to gene. Our solution is to abandon the idea that the gene strives to maximize its personal fitness $\mathscr{W}$, and instead, we consider that it is inclusive fitness $\mathscr{H}$ that the gene is interested in maximizing. This takes as its arguments the phenotype π_i of the focal gene and the population context Π in which this gene finds itself. Hence, we write $\mathscr{H} \leftrightarrow \mathscr{U}$.

If all other aspects of the analogy remain as specified in the previous section, we may express the gene as inclusive fitness–maximizing agent analogy in the form of an optimization program:

$$\begin{array}{cc} \pi & \max \quad \mathscr{H}(\pi;\Pi) \\ & \pi \in P \end{array} \qquad (9)$$

This provides a formal statement of what we mean when we say that a gene wields its phenotype as an instrument in the pursuit of its objective to maximize its inclusive fitness. It provides a rigorous basis for making statements about the conditional optimality of phenotypes, in this respect. In particular, an optimal phenotype is one that (i) belongs to the set of possible phenotypes and (ii) achieves an inclusive fitness that is equal to or greater than that of any other phenotype in the set, conditional upon population composition. Formally, an optimal phenotype is π^* where $\mathscr{H}(\pi^*;\Pi) \geq \mathscr{H}(\pi;\Pi)\ \forall\ \pi \in P$. Conversely, a suboptimal phenotype is one that (i) belongs to the set of possible phenotypes and (ii) achieves lower inclusive fitness than at least one other member of the set, conditional upon population composition. Formally, a suboptimal phenotype is π° where $\exists\ \pi \in P{:}\mathscr{H}(\pi^\circ;\Pi) < \mathscr{H}(\pi;\Pi)$. For the purpose of evaluating the optimality of a phenotype that is absent from the given population, we assume the corresponding allele can be introduced at vanishingly low frequency with negligible disturbance to the population composition or genetic relatedness of social partners. Hence, we strictly interpret $\mathscr{H}$ as evaluating inclusive fitness of a phenotype π in the context of a population in the vicinity of Π.

Formal justification for the social gene

Here, we derive correspondences between the action of genic selection – summarized by dynamical equation (3) – and the idea of the gene as an inclusive fitness–maximizing

agent – summarized by optimization program (9). This generalizes the link between our dynamical and purposeful accounts of genic adaptation (Table 3; derivations given in Appendix).

Our six correspondences are the same as were derived for the model lacking social interactions: if all genes are behaving optimally, then there is neither scope for selection (I) nor potential for positive selection (II); if all genes are behaving suboptimally, but equally so, then there is no scope for selection (III) but there is potential for positive selection (IV); if genes vary in their optimality, then the change in the average of all genic values and in all allele frequencies is equal to the covariance, taken over all gene positions, of attained maximand (inclusive fitness; V); and if there is neither scope for selection nor potential for positive selection, then all genes are behaving optimally (VI). These formal correspondences reveal that the view of genes as inclusive-fitness maximizers can be translated into rigorous, population-genetic statements. They also formally capture the sense in which the dynamics of genic selection is governed by the design principle of inclusive-fitness maximization, leading to the emergence of genic adaptations that appear to function for this purpose.

Difficulties of the theory

Laborious as it is, a benefit of the formal approach taken in this article is to force us to be clear about what we mean by the theory of selfish genes: a subject upon which there has been little agreement (Daly, 1980; Maynard Smith, 1987; Lloyd, 2001, 2005). Working with a clear definition makes it much easier to identify and distinguish between the many criticisms that have been levelled at this approach. A list of distinct criticisms is catalogued in Table 1, and each is addressed in this section. All of these criticisms have been raised by multiple authors, but they have not always been separated in this way.

A first objection to taking the gene's eye view is simply that 'genes do not have intentions' (Stent, 1977; Langley, 1977; Midgley, 1979, 1983; Daly, 1980; Table 1, row 1). However, the point of formal Darwinism (Grafen, 2002, 2003, 2006a, 2007, 2009; Gardner, 2009; Gardner & Grafen, 2009) is to show that the metaphor of agency (as embodied in an optimization program) can have predictive power over a dynamical system (i.e. a population-genetic model) that contains no intentionality of any kind, let alone 'true' intention (whatever that means; Dennett, 1989 Ch. 3). We have shown that the metaphor can be applied to genes as well as to organisms.

But why invoke imaginary agents when a dynamical description of evolution would be less misleading and more complete (Stent, 1977; Haig, 1997; Grafen, 2003; Charlesworth, 2006; Godfrey-Smith, 2009; Noble, 2011)? A full description of evolutionary change includes transmission effects, genetic drift and nonadditive gene interaction, and these factors show no correspondence to the idea of gene-level intentionality (eqns 1–3 and 8), but may have substantial effects on evolutionary outcomes. A pragmatic response is to note the overwhelming empirical success of optimization-based research in the study of organism-level adaptation (Maynard Smith, 1982 Ch. 1, 1987; Parker & Maynard Smith, 1990; Grafen, 2007; West, 2009 Ch. 11). *A priori*, there is no reason to suspect that this approach must fail when applied at the gene level. Its scientific utility will ultimately be judged by empiricists and not by theoreticians or philosophers. More fundamentally, the language of intentionality is unavoidable if we are to fully explain the apparent design of the living world (Paley, 1802; Darwin, 1859; Leigh, 1971; Dennett, 1989; Gardner, 2009; Grafen, 2009). For this purpose, the correspondence of the optimization program to just one component of the total evolutionary change highlights two important facts about adaptation: that only selection (and not transmission effects or drift) has a tendency to generate apparent design (Grafen, 2003; Gardner, 2009; Lenormand *et al.*, 2009; Ewens, 2011) and that adaptation does not imply perfection of design, or even optimality within specified constraints (Paley, 1802; Langley, 1977; Lewontin, 1977; Dawkins, 1982; Maynard Smith, 1987; Parker & Maynard Smith, 1990; Gould, 2002; Gardner, 2009).

A second major difficulty for the theory of the selfish gene is to decide exactly what the 'gene' is (Stent, 1977; Bateson, 2006; Daly, 1980; Williams, 1966, 1992; Haig, 1997; Hull, 2001; Lloyd, 2005; Table 1, row 2). When being most explicit, Dawkins (1976) defines the gene as a 'distributed agent' comprising every copy of a particular allele. However, most applications of selfish-gene theory locate agency at the level of the physical scrap of nucleic acid – e.g. the transposon (Burt & Trivers, 2006) – and this is the approach we have taken here (see also Hamilton, 1972). There are conceptual problems with the idea of the allele as a maximizing agent: the allele is associated with a particular phenotype and hence is better seen as an encoded strategy, rather than an agent in its own right. In contrast, we can, for our purposes, readily imagine an intentional scrap of nucleic acid deciding which protein it is to encode, and this permits a formal separation of the notions of agent and strategy. There may be viable alternatives that recover the distributed-agent view, such as seeing the gene as a cloud of identical-by-descent scraps of nucleic acid. However, such agents are statistical rather than concrete objects, and it is difficult to assign diffuse probabilistic clouds a causal role in evolutionary biology.

A third criticism concerns the word 'selfish' (Table 1, row 3). In the primary literature, the term 'selfish gene' commonly denotes a genetic element that spreads despite decreasing the inclusive fitness of its 'host' organism (Burt & Trivers, 2006). The usefulness of taking the 'gene's eye view' in such cases is clear, especially

considering the conceptual tangles of early attempts to explain the high prevalence of selfish genetic elements in natural populations (e.g. Gershenson, 1928, pp. 506–507). However, the theory of the 'selfish gene' was originally intended to apply more widely, i.e. to all genes, not just to intragenomic outlaws. Specifically, selfish-gene theory represents an attempt to reduce organismal altruism to the self-interested behaviour of a more fundamental agent (Dawkins, 1976, 1978, 1982). Our formalism suggests that this attempt does not succeed. When a gene's fitness is mediated by the action of other genes, genes are expected to behave as if striving to maximize their inclusive fitness, rather than their personal fitness (eqns 8 and 9). Such social effects may therefore drive the evolution of unselfish behaviour among genes, such as altruism and spite (Burt & Trivers, 2006; West *et al.*, in press). Hence, the theory of the selfish gene appears misnamed. Certain elements of the theory can be rigorously justified – in particular, the notion of the 'striving' or 'intentional' gene – but only if we change the way we think about the agenda of this agent. The theory works as an application of Hamilton's (1964, 1970, 1996 Ch. 2) theory of inclusive fitness to the gene level, but not as a recasting of that theory as originally applied at the organism level (Dawkins, 1976, 1978, 1982).

The claim that genes must always be selfish is closely connected to the attribution of agency to the allele (Table 1, row 2). Because the allele can be seen as the ultimate beneficiary of the evolutionary process (Dawkins, 1976, 1978, 1982; Hampe & Morgan, 1988; Williams, 1992; Lloyd, 2001; but see Godfrey-Smith 2011), it is taken to follow that its actions are, by definition, selfish. But this argument fails if, as we argue, the allele cannot function as an intentional agent. More generally, the formalism that allows us to capture the notion of adaptation in mathematical terms does not require that we should seek 'fundamental' agents who are always selfish. Genes do play a fundamental role, on the dynamical side of adaptation theory: it is via the medium of gene-frequency change that adaptations – at any level of biological organization – are moulded by the action of selection (Fisher 1930; Grafen, 2002, 2003; Gardner, 2009; Gardner & Grafen, 2009; Ewens, 2011). However, this role should not be confused with the notions of intentionality and selfishness, which belong to the optimization side of adaptation theory.

A fourth criticism of the gene's eye view is that, because a gene's fitness is prone to vary as a function of its genetic background, considering each gene in isolation can be misleading (Sober & Lewontin, 1982; Dover, 2000; Gould, 2002; Okasha, 2006 Ch. 5; Table 1, row 4). A gene might, for example, be selected for in one genetic background and selected against in another. However, we have shown that genes can be considered as inclusive fitness–maximizing agents irrespective of the degree of such epistatic effects, by defining inclusive fitness (eqn 8) in such a way as to exclude nonadditive effects. Although such effects may mediate genotypic change, they do not contribute to the genetic change that defines the action of selection (Grafen, 2003; Crow, 2008; Ewens, 2011; Gardner *et al.*, 2011) and show no correspondence with the optimization program. So, although neglect of these epistatic effects will lead to a less complete (and potentially quite inaccurate) account of the evolutionary process (Charlesworth, 1990; Nagylaki, 1991, 1993; Gould, 2002; Okasha, 2006, p. 160; Table 1, row 1), the gene's eye view does function adequately as a theory of adaptation. Furthermore, this argument does not just apply to genes. Only additive genetic effects contribute to the action of selection acting upon organisms (Crow, 2008; Ewens, 2011; Gardner *et al.*, 2011), and only these show correspondences with the analogy of the organism as maximizing agent (Grafen, 2002, 2003, 2006a, 2007).

A fifth criticism of the selfish-gene perspective is that most phenotypes of interest are multigenic in their underlying architecture and/or shaped by environmental effects and hence cannot be interpreted as expressing the intentions of a single gene (Langley, 1977; Lewontin, 1977; Bateson, 1978, 2006; Daly, 1980; Dawkins, 1982; Mayr, 1983; Dover, 2000; Gould, 2002; Noble, 2011; Table 1, row 5). Our formalism supports this criticism: a strategy is, by definition, under the full control of a single agent. This puts strong constraints upon the kinds of phenotype that can be addressed by the selfish-gene perspective: the phenotype cannot be mediated by social or environmental effects. Gene-level adaptations will usually have to be sought at a relatively proximate level – such as at the level of proteins – or else be defined in a convoluted and nonintuitive manner.

However, it is not clear that this is really a limitation: any complex phenotype – such as the eye, which is traditionally viewed as an organism-level adaptation – can, at least in principle, be dissected into a large number of gene-level adaptations, each involving realized phenotypes that are conditional upon various 'environmental' cues. Are there any grounds for preferring one view of adaptation to the other (Dawkins, 1982 Ch. 1; Sterelny & Kitcher, 1988; Kitcher *et al.*, 1990; Lloyd, 2005)? Following Williams (1966), Occam's Razor can be employed to argue for the gene's eye view, on the basis that it accounts for phenomena as diverse as meiotic drive and (bits of) the eye, without positing unnecessary organism-level agents (Dawkins, 1982; Kitcher *et al.*, 1990). On the other hand, as most phenotypes are not impacted by Mendelian outlawry, the Razor can also be made to argue for the reverse position: why assume a very large number of gene-level agents, when a single organism-level agent will suffice? Moreover, the hallmark of design is the contrivance of multiple parts as if for a common purpose (Paley, 1802; Gardner, 2009), and organisms enjoy a much greater potential for the elaboration of complex design than do mere molecular agents (Hurst *et al.*, 1996; Hammerstein & Hagen, 2006).

Indeed, treating the organism as a maximizing agent formally captures this idea of common purpose, showing how the evolutionary dynamics of all genes, with the same pattern of inheritance, correspond to the same optimization program (Grafen, 2002, 2006a, 2007; Gardner, 2009; Gardner & Grafen, 2009).

A sixth criticism of the theory is that the gene's eye view misrepresents the 'causal structure' of adaptive evolution (Table 1, row 6). The same selective event can often be described in multiple, mathematically equivalent ways, corresponding to different partitions of the total evolutionary change (Sterelny & Kitcher, 1988; Kitcher *et al.*, 1990; Okasha, 2006). For example, heterozygote disadvantage can be modelled by assigning an organismal fitness to each diploid genotype or else by assigning context-dependent fitnesses to each allele, and the predictions will be exactly the same. Some critics have argued that, mathematical equivalence notwithstanding, the gene's eye view misrepresents the causal structure of selection in such scenarios (Wade, 1978; Sober & Lewontin, 1982; Dover, 2000; Gould, 2002; Lloyd, 2005; Okasha, 2006).

Insofar as they concern adaptation, these disputes relate to its proximate cause, i.e. the manner in which biological entities interact with their environment in a way that makes a difference to replicator success (Hull, 2001; Lloyd, 2001, 2005; Okasha, 2006). The approach taken here asks a different set of questions, concerning ultimate causation: why do we observe adaptations, at what level of biological organization are they manifest, and what do they appear designed to achieve (Gardner, 2009)? Accordingly, our notion of agency implies something other than 'important actor in the evolutionary process' or 'target of selection' (Bateson, 1978, 2006; Dover, 2000; Grafen, 2000; Gould, 2002). Rather, it concerns the notion of the 'maximizing agent', which is a crucial component of any formal theory of function, purpose, intention or design. In our formalism, a phenotype might be considered a gene-level adaptation, even if no within-organism selection has taken place. The grounds for choosing between different descriptions must be sought elsewhere (e.g. Table 1, row 5). Furthermore, by refocusing attention on the control of the adaptive phenotype – i.e. the identification of the maximizing agent who wields the instrument – the present framework stresses an issue with genuine empirical purchase (Dawkins, 1982; Hurst *et al.*, 1996; Mehdiabadi *et al.*, 2002; Grafen, 2008).

Conclusions

We have developed the first formal theory of the selfish gene, by constructing a 'gene as maximizing agent' analogy. We have linked notions of gene-level intentionality (captured by an optimization program) to evolutionary-genetic dynamics (captured by Price's equation). This has provided formal justification for viewing genes as purposeful, adaptive agents and has allowed us to address several criticisms that have been levelled at the selfish gene over the decades since its conception. We have aimed for conceptual clarity rather than full generality, so our model necessarily neglects certain biological complexities, such as class structure. We have assumed that all loci are equivalent and that the population is homogeneous in its social networks. Relaxing such assumptions is an important goal for future research. Nevertheless, we have allowed for arbitrary social interaction with an arbitrary number of social partners and arbitrary additive and nonadditive fitness effects. We have shown that, in general, the gene should be regarded as striving to maximize its inclusive fitness and not its personal fitness. Hence, genes may be altruistic and spiteful as well as selfish. We have laid the formal foundations for a research programme that views genes as inclusive fitness–maximizing agents. As with any research programme, the gene's eye view cannot be considered a simple hypothesis that can be falsified with a single experiment. Rather, it must ultimately be judged according to how well it facilitates hypothesis generation and empirical testing and advancing scientific understanding of the natural world.

Acknowledgments

We thank João Alpedrinha, Brian Charlesworth, Troy Day, Warren Ewens, Craig MacLean, Allen Moore, Denis Noble, Francisco Úbeda, Stuart West and especially Alan Grafen for helpful discussion. AG is supported by research fellowships from Balliol College and the Royal Society.

References

Bateson, P.P.G. 1978. Book review: *The selfish gene* by Richard Dawkins. *Anim. Behav.* **26**: 316–318.

Bateson, P.P.G. 2006. The nest's tale. A reply to Richard Dawkins. *Biol. Philos.* **21**: 553–558.

Burt, A. & Trivers, R. 2006. *Genes in Conflict*. Belknap Press, Cambridge, MA.

Charlesworth, B. 1990. Optimization models, quantitative genetics, and mutation. *Evolution* **44**: 520–538.

Charlesworth, B. 2006. Conflicts of interest. *Curr. Biol.* **16**: R1009–R1011.

Crow, J.F. 2008. Maintaining evolvability. *J. Genet.* **87**: 349–353.

Daly, M. 1980. Contentious genes: a commentary on *The selfish gene* by Richard Dawkins. *J. Soc. Biol. Struct.* **3**: 77–81.

Darwin, C.R. 1859. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. John Murray, London, UK.

Dawkins, R. 1976. *The Selfish Gene*. Oxford University Press, Oxford, UK.

Dawkins, R. 1978. Replicator selection and the extended phenotype. *Z. Tierpsychol.* **47**: 61–76.

Dawkins, R. 1982. *The Extended Phenotype*. Oxford University Press, Oxford, UK.

Dennett, D.C. 1989. *The Intentional Stance*. MIT Press, Cambridge, MA.
Dover, G. 2000. Anti-Dawkins. In: *Alas Poor Darwin: Arguments Against Evolutionary Psychology* (H. Rose & S. Rose, eds), pp. 55–77. Harmony Books, New York, NY.
Ewens, W.J. 2011. What is the gene trying to do? *Br. J. Philos. Sci.* **62**: 155–176.
Fisher, R.A. 1918. The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edinb.* **52**: 399–433.
Fisher, R.A. 1930. *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford.
Fisher, R.A. 1941. Average excess and average effect of a gene substitution. *Ann. Eugen.* **11**: 53–63.
Frank, S.A. 1998. *Foundations of Social Evolution*. Princeton University Press, Princeton, NJ.
Gardner, A. 2009. Adaptation as organism design. *Biol. Lett.* **5**: 861–864.
Gardner, A. & Grafen, A. 2009. Capturing the superorganism: a formal theory of group adaptation. *J. Evol. Biol.* **22**: 659–671.
Gardner, A., West, S.A. & Wild, G. 2011. The genetical theory of kin selection. *J. Evol. Biol.* **24**: 1020–1043.
Gershenson, S. 1928. A new sex-ratio abnormality in *Drosophila obscura*. *Genetics* **13**: 488–507.
Godfrey-Smith, P. 2009. *Darwinian Populations and Natural Selection*. Oxford University Press, Oxford.
Gould, S.J. 2002. *The Structure of Evolutionary Theory*. Harvard University Press, Cambridge, MA.
Grafen, A. 2000. Developments of the Price equation and natural selection under uncertainty. *Proc. R. Soc. B* **267**: 1223–1227.
Grafen, A. 2002. A first formal link between the Price equation and an optimization program. *J. Theor. Biol.* **217**: 75–91.
Grafen, A. 2003. Fisher the evolutionary biologist. *J. R. Stat. Soc. D* **52**: 319–329.
Grafen, A. 2006a. Optimization of inclusive fitness. *J. Theor. Biol.* **238**: 541–563.
Grafen, A. 2006b. A theory of Fisher's reproductive value. *J. Math. Biol.* **53**: 15–60.
Grafen, A. 2007. The formal Darwinism project: a mid-term report. *J. Evol. Biol.* **20**: 1243–1254.
Grafen, A. 2008. The simplest formal argument for fitness optimisation. *J. Genet.* **87**: 421–433.
Grafen, A. 2009. Formalizing Darwinism and inclusive fitness theory. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **364**: 3135–3141.
Haig, D. 1997. The social gene. In: *Behavioural Ecology*, 4th edn (J.R. Krebs & N.B. Davies, eds), pp. 284–304. Blackwell Science Ltd, Oxford, UK.
Hamilton, W.D. 1963. The evolution of altruistic behavior. *Am. Nat.* **97**: 354–356.
Hamilton, W.D. 1964. The genetical evolution of social behaviour. *J. Theor. Biol.* **7**: 1–52.
Hamilton, W.D. 1970. Selfish and spiteful behaviour in an evolutionary model. *Nature* **228**: 1218–1220.
Hamilton, W.D. 1972. Altruism and related phenomena, mainly in social insects. *Ann. Rev. Ecol. Syst.* **3**: 193–232.
Hamilton, W.D. 1996. *Narrow Roads of Gene land I: Evolution of Social Behaviour*. W.H. Freeman, Oxford, UK.
Hammerstein, P. & Hagen, E.H. 2006. Broken cogs or strategic agents? *Science* **312**: 530.
Hampe, M. & Morgan, S.R. 1988. Two consequences of Richard Dawkins' view of genes and organisms. *Stud. Hist. Philos. Sci.* **19**: 119–138.
Hull, D.L. 2001. *Science and Selection: Essays on Biological Evolution and the Philosophy of Science*. Cambridge University Press, Cambridge, UK.
Hurst, L.D., Atlan, A. & Bengtsson, B.O. 1996. Genetic conflicts. *Q. Rev. Biol.* **71**: 317–364.
Israels, A.Z. 1987. Path analysis for mixed qualitative and quantitative variables. *Qual. Quant.* **21**: 91–102.
Kitcher, P., Sterelny, K. & Waters, C.K. 1990. The illusory riches of Sober's monism. *J. Philos.* **87**: 158–161.
Langley, C.H. 1977. A little Darwinism. *Bioscience* **27**: 692.
Leigh, E. 1971. *Adaptation and Diversity*. Cooper, San Francisco, CA.
Lenormand, T., Roze, D. & Rousset, F. 2009. Stochasticity in evolution. *Trends Ecol. Evol.* **24**: 157–165.
Lewontin, R.C. 1977. Caricature of Darwinism. *Nature* **266**: 283–284.
Lloyd, E.A. 2001. Different questions: levels and units of selection. In: *Thinking About Evolution: Historical, Philosophical, and Political Perspectives* (R.S. Singh, C.B. Krimbas, D.B. Paul & J. Beatty, eds), pp. 267–291. Cambridge University Press, New York, NY.
Lloyd, E.A. 2005. Why the gene will not return. *Philos. Sci.* **72**: 287–310.
Maynard Smith, J. 1982. *Evolution and the Theory of Games*. Cambridge University Press, Cambridge, UK.
Maynard Smith, J. 1987. How to model evolution. In: *The Latest on the Best: Essays on Evolution and Optimality* (J. Dupré, ed.), pp. 119–132. MIT Press, Cambridge, MA.
Mayr, E. 1983. How to carry out the adaptationist program? *Am. Nat.* **121**: 324–334.
Mehdiabadi, N.J., Reeve, H.K. & Mueller, U.G. 2002. Queens versus workers: sex-ratio conflict in eusocial Hymenoptera. *Trends Ecol. Evol.* **18**: 88–92.
Midgley, M. 1979. Gene juggling. *Philosophy* **54**: 439–458.
Midgley, M. 1983. Selfish genes and social Darwinism. *Philosophy* **58**: 365–377.
Nagylaki, T. 1991. Error bounds for the fundamental and secondary theorems of natural selection. *Proc. Natl Acad. Sci. USA* **88**: 2402–2406.
Nagylaki, T. 1993. The evolution of multilocus systems under weak selection. *Genetics* **134**: 627–647.
von Neumann, J. & Morgenstern, O. 1944. *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ.
Noble, D. 2011. Neo-Darwinism, the Modern Synthesis, and Selfish Genes: are they of use in physiology? *J. Physiol.* **589**: 1007–1015.
Okasha, S. 2006. *Evolution and the Levels of Selection*. Oxford University Press, Oxford, UK.
Orlove, M.J. & Wood, C.L. 1978. Coefficients of relationship and coefficients of relatedness in kin selection: a covariance form for the RHO formula. *J. Theor. Biol.* **73**: 679–686.
Paley, W. 1802. *Natural Theology*. Wilks & Taylor, London, UK.
Parker, G.A. & Maynard Smith, J. 1990. Optimality theory in evolutionary biology. *Nature* **348**: 27–33.
Price, G.R. 1970. Selection and covariance. *Nature* **227**: 520–521.
Price, G.R. 1972. Extension of covariance selection mathematics. *Ann. Hum. Genet.* **35**: 485–490.
Queller, D.C. 1992. A general model for kin selection. *Evolution* **46**: 376–380.
Robertson, A. 1966. A mathematical model of the culling process in dairy cattle. *Anim. Product.* **8**: 95–108.
Robertson, A. 1968. The spectrum of genetic variation. In: *Population Biology and Evolution* (R.C. Lewontin, ed.), pp. 5–16. Syracuse University Press, Princeton, NJ.

Sober, E. & Lewontin, R.C. 1982. Artifact, cause and genic selection. *Philos. Sci.* **47**: 157–180.
Stent, G.S. 1977. You can take the ethics out of altruism but you can't take the altruism out of ethics. *Hastings Cent. Rep.* **7**: 33–36.
Sterelny, K. & Kitcher, P. 1988. The return of the gene. *J. Philos.* **85**: 339–361.
Wade, M.J. 1978. The selfish gene. *Evolution* **32**: 220–221.
West, S.A. 2009. *Sex Allocation*. Princeton University Press, Princeton, NJ.
West, S.A. & Gardner, A. 2010. Altruism, spite and greenbeards. *Science* **327**: 1341–1344.
West, S.A., Griffin, A.S. & Gardner, A. 2007. Social semantics: altruism, cooperation, mutualism, strong reciprocity and group selection. *J. Evol. Biol.* **20**: 415–432.
West, S.A., El Mouden, C. & Gardner, A. In press. 16 common misconceptions about the evolution of cooperation in humans. *Evol. Hum. Behav.*, doi: 10.1016/j.evolhumbehav.2010.08.001.
Williams, G.C. 1966. *Adaptation and Natural Selection*. Princeton University Press, Princeton, NJ.
Williams, G.C. 1992. *Natural Selection: Domains, Levels & Challenges*. Princeton University Press, Princeton, NJ.
Wright, S. 1934. The method of path coefficients. *Ann. Math. Stat.* **5**: 161–215.

Appendix

Social interactions absent

I. *If all agents are optimal, there is no scope for selection* – If all agents behave optimally, then $w_i = \mathscr{W}(\pi^*)/\mathscr{W}(\pi^*) = 1$ for all $i \in I$. Hence, from eqn 3, $E_\Omega(\Delta_S E_I(g)) = \mathrm{cov}_I(w,g) = \mathrm{cov}_I(1,g) = 0$.

II. *If all agents are optimal, there is no potential for positive selection* – Introduce a new allele $a' \in A$ into the population at vanishingly low frequency. Carriers of allele a' have phenotype $\pi' = \mathscr{P}(a')$, and all other genes, being optimal, have phenotype π^*. Assigning carriers of allele a' a genic value $g = 1$, and carriers of the other alleles a genic value $g = 0$, the response to selection is $E_\Omega(\Delta_S E_I(g)) = \mathrm{cov}_I(\mathscr{W}(\pi),g) = E_I(g)(1 - E_I(g))(\mathscr{W}(\pi') - \mathscr{W}(\pi^*))$. Since $\mathscr{W}(\pi') \leq \mathscr{W}(\pi^*)\ \forall\ \pi' \in P$, then $E_\Omega(\Delta_S E_I(g)) \leq 0$.

III. *If all agents are suboptimal, but equally so, there is no scope for selection* – If all agents behave suboptimally, but equally so, then $w_i = \mathscr{W}(\pi^\circ)/\mathscr{W}(\pi^\circ) = 1$ for all $i \in I$. Hence, from eqn 3, $E_\Omega(\Delta_S E_I(g)) = \mathrm{cov}_I(w,g) = \mathrm{cov}_I(1,g) = 0$.

IV. *If all agents are suboptimal, but equally so, there is potential for positive selection* – Introduce the allele a^* corresponding to the optimal phenotype (i.e. satisfying $\pi^* = \mathscr{P}(a^*)$) into the population at vanishingly low frequency. All other genes, being suboptimal, have phenotype π°. Assigning carriers of allele a^* a genic value $g = 1$, and carriers of the other alleles a genic value $g = 0$, the response to selection is $E_\Omega(\Delta_S E_I(g)) = \mathrm{cov}_I(\mathscr{W}(\pi),g) = E_I(g)\ (1 - E_I(g))\ (W(\pi^*) - \mathscr{W}(\pi^\circ))$. Since $\mathscr{W}(\pi^*) > \mathscr{W}(\pi^\circ)$, then $E_\Omega(\Delta_S E_I(g)) > 0$.

V. *If agents vary in their optimality, there is scope for selection, and change in the average of all genic values, and in all gene frequencies, is given by their covariance with relative attained maximand* – From eqn 3, $E_\Omega(\Delta_S E_I(g)) = \mathrm{cov}_I(w,g) = \mathrm{cov}_I(\mathscr{W}(\pi)/E_I(\mathscr{W}(\pi)),g)$.

VI. *If there is neither scope for selection nor potential for positive selection, all agents are optimal* – If all agents are optimal, there is neither scope for selection (I) nor potential for positive selection (II); if all agents are suboptimal, but equally so, then there is no scope for selection (III) but potential for positive selection (IV); if agents vary in their optimality there is scope for selection (V). These exhaust the possibilities in the optimization view; hence, if there is neither scope for selection nor potential for positive selection, all agents must be optimal.

Social interactions permitted

Partitioning fitness – In the main text, we partitioned a gene's personal fitness into a baseline, plus additive effects of phenotypes in the social set, plus the non-additive effect of phenotypes in the social set. Here, we make this partition explicit, showing how each component may be computed using least-squares regression, analogous to how fitness effects are computed for social interactions between individual organisms (reviewed by Gardner *et al.*, 2011). A complication for regression analysis is posed by the qualitative nature of the allelic and phenotypic variables a and π: regression analysis requires that variables represent numerical quantities. We resolve this problem by representing the qualitative variables by a set of binary dummy variables (Israels, 1987). In particular, we write a number of dummies equal to the number of qualities, such that the nth dummy represents the presence (1) or absence (0) of the nth quality, and the each quality is uniquely determined by a string of n dummies, one of which takes value 1 and the rest take value 0. We index allelic dummies by $x \in X$ and phenotypic dummies by $y \in Y$.

Figure A1 provides a path diagram of the causal link between genes and fitness (Wright, 1934). The LHS of Fig. A1 shows that the allele a_j carried by the role-j social partner causally determines the genic value g_j of the role-j social partner (solid, single-headed arrow), and that the genic value g_j of the role-j social partner is correlated (broken, double-headed arrow) with the genic value g of the focal gene. The RHS of Fig. A1 shows that the allele a_j carried by the role-j social partner causally determines the phenotype π_j of the role-j social partner and that the phenotype π_j of the role-j social partner causally determines the fitness w of the focal gene. True quantitative variables are shown in circles, and dummy variables representing qualitative variables are shown in boxes. For simplicity, Fig. A1 shows only two roles ($j = 1, 2$), two alleles ($x = 1, 2$) and two possible phenotypes ($y = 1, 2$).

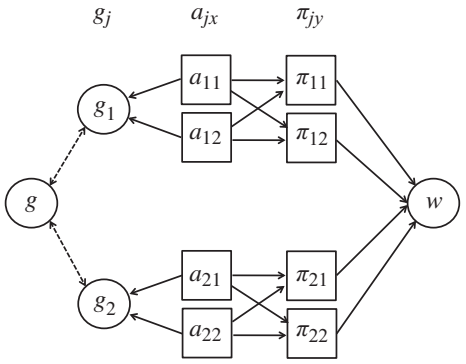

Fig. A1 Path diagram illustrating the causal link between genes and fitness. The allele a_j of a gene's role-j social partner causally determines (solid, single-headed arrow) the genic value g_j of the gene's role-j social partner, and the genic value g_j of the gene's role-j social partner is correlated (broken, double-headed arrow) with the genic value g of the focal gene. The allele a_j of a gene's role-j social partner also causally determines the phenotype π_j of the gene's role-j social partner, and the phenotype π_j of the gene's role-j social partner has a causal impact upon the personal fitness w of the focal gene. True quantitative variables are shown in circles, and dummy variables representing qualitative variables are shown in boxes.

However, the number of roles, alleles and possible phenotypes is arbitrary.

We write the following linear-regression model of fitness as a function of the phenotypes of the social set:

$$w = 1 + \sum_J \sum_Y \tilde{\beta}_{w,\pi_{jy}} \left(\pi_{jy} - \mathrm{E}_I\left(\pi_{jy}\right)\right) + \varepsilon \qquad \text{(A1)}$$

where the coefficients β are chosen so as to minimize $\mathrm{E}_I(\epsilon^2)$, i.e. they are least-squares regression coefficients. We will use the notation $\beta_{\kappa,\lambda}$ to denote the regression of dependent variable κ against predictor variable λ, and we will use the notation $\tilde{\beta}_{\kappa,\lambda}$ to denote the partial regression of κ against λ, i.e. holding fixed all other predictors in the same column as λ in Fig. A1.

Equation A1 defines the partition of personal fitness, with

$$\mathscr{W}_\mathrm{B}(\Pi) = 1 \qquad \text{(A2)}$$

$$\mathscr{W}_\mathrm{A}\left(\pi_j, j; \Pi\right) = \sum_Y \tilde{\beta}_{w,\pi_{jy}} \left(\pi_{jy} - \mathrm{E}_I\left(\pi_{jy}\right)\right) \qquad \text{(A3)}$$

$$\mathscr{W}_\mathrm{N}(\tilde{\pi}; \Pi) = \varepsilon \qquad \text{(A4)}$$

Genic selection – From eqn 3, the response to genic selection is $\mathrm{E}_\Omega(\Delta_\mathrm{S}\mathrm{E}_I(g)) = \mathrm{cov}_I(w,g) = \beta_{w,g}\mathrm{cov}_I(g,g) = \sum_J \sum_X \sum_Y \tilde{\beta}_{w,\pi_{jy}} \tilde{\beta}_{\pi_{jy},a_{jx}} \beta_{a_{jx},g_j} \beta_{g_j,g} \mathrm{cov}_I(g,g)$. From the main text, we have $r_j = \mathrm{cov}_I(g_j,g)/\mathrm{cov}_I(g,g) = \beta_{g_j,g}$, and from eqn A3, we have $\beta_{\mathscr{W}_\mathrm{A}\left(\pi_j,j;\Pi\right),g_j} = \sum_X \sum_Y \beta_{w,\pi_{jy}} \tilde{\beta}_{\pi_{jy},a_{jx}} \beta_{a_{jx},g_j}$. Hence, we may write $\mathrm{E}_\Omega(\Delta_\mathrm{S}\mathrm{E}_I(g)) = \sum_J \beta_{\mathscr{W}_\mathrm{A}\left(\pi_j,j;\Pi\right),g_j} r_j \mathrm{cov}_I(g,g) = \sum_j \beta_{\mathscr{W}_\mathrm{A}(\pi,j;\Pi),g} r_j \mathrm{cov}_I(g,g) = \mathrm{cov}_I(\sum_J \mathscr{W}_\mathrm{A}(\pi,j;\Pi) r_j, g) = \mathrm{cov}_I(\mathscr{W}_\mathrm{B}(\Pi) + \sum_J \mathscr{W}_\mathrm{A}(\pi,j;\Pi) r_j, g)$. The response to genic selection may therefore be expressed as:

$$\mathrm{E}_\Omega(\Delta_\mathrm{S}\mathrm{E}_I(g)) = \mathrm{cov}_I(\mathscr{H}(\pi;\Pi), g) \qquad \text{(A5)}$$

I. *If all agents are optimal, there is no scope for selection* – If all agents behave optimally, then $\mathscr{H}(\pi_i;\Pi) = \mathscr{H}(\pi^*;\Pi)$ for all $i \in I$. Hence, from eqn A5, $\mathrm{E}_\Omega(\Delta_\mathrm{S}\mathrm{E}_I(g)) = \mathrm{cov}_I(\mathscr{H}(\pi;\Pi),g) = \mathrm{cov}_I(\mathscr{H}(\pi^*;\Pi),g) = 0$.

II. *If all agents are optimal, there is no potential for positive selection* – We choose a population composition in the vicinity of Π such that a new allele $a' \in A$ is present at vanishingly low frequency and with negligible impact upon the genetic relatedness of social partners. Carriers of allele a' have phenotype $\pi' = \mathscr{P}(a')$ and all other genes, being optimal, have phenotype π^*. Assigning carriers of allele a' a genic value $g = 1$, and carriers of the other alleles a genic value $g = 0$, the response to selection is $\mathrm{E}_\Omega(\Delta_\mathrm{S}\mathrm{E}_I(g)) = \mathrm{cov}_I(\mathscr{H}(\pi;\Pi), g) = \mathrm{E}_I(g)(1 - \mathrm{E}_I(g))(\mathscr{H}(\pi';\Pi) - \mathscr{H}(\pi^*;\Pi))$. Since $\mathscr{H}(\pi';\Pi) \leq \mathscr{H}(\pi^*;\Pi)\ \forall\ \pi' \in P$, then $\mathrm{E}_\Omega(\Delta_\mathrm{S}\mathrm{E}_I(g)) \leq 0$.

III. *If all agents are suboptimal, but equally so, there is no scope for selection* – If all agents are equally suboptimal, then $\mathscr{H}(\pi_i;\Pi) = \mathscr{H}(\pi^\circ;\Pi)$ for all $i \in I$. Hence, from eqn A5, $\mathrm{E}_\Omega(\Delta_\mathrm{S}\mathrm{E}_I(g)) = \mathrm{cov}_I(\mathscr{H}(\pi;\Pi),g) = \mathrm{cov}_I(\mathscr{H}(\pi^\circ;\Pi),g) = 0$.

IV. *If all agents are suboptimal, but equally so, there is potential for positive selection* – We choose a population composition in the vicinity of Π such that the allele a^* corresponding to the optimal phenotype (i.e. satisfying $\pi^* = \mathscr{P}(a^*)$) is present at vanishingly low frequency and with negligible impact upon the genetic relatedness of social partners. All other genes, being suboptimal, have phenotype π°. Assigning carriers of allele a^* a genic value $g = 1$, and carriers of the other alleles a genic value $g = 0$, the response to selection is $\mathrm{E}_\Omega(\Delta_\mathrm{S}\mathrm{E}_I(g)) = \mathrm{cov}_I(\mathscr{H}(\pi;\Pi),g) = \mathrm{E}_I(g)(1 - \mathrm{E}_I(g))(\mathscr{H}(\pi^*;\Pi) - \mathscr{H}(\pi^\circ;\Pi))$. Since $\mathscr{H}(\pi^*;\Pi) > \mathscr{H}(\pi^\circ;\Pi)$, then $\mathrm{E}_\Omega(\Delta_\mathrm{S}\mathrm{E}_I(g)) > 0$.

V. *If agents vary in their optimality, there is scope for selection, and change in the average of all genic values, and in all gene frequencies, is given by their covariance with attained maximand* – From eqn A5, $\mathrm{E}_\Omega(\Delta_\mathrm{S}\mathrm{E}_I(g)) = \mathrm{cov}_I(\mathscr{H}(\pi;\Pi),g)$.

VI. *If there is neither scope for selection nor potential for positive selection, all agents are optimal* – If all agents are optimal, there is neither scope for selection (I) nor potential for positive selection (II); if all agents are suboptimal, but equally so, then there is no scope for selection (III) but potential for positive selection (IV); if agents vary in their optimality, there is scope for selection (V). These exhaust the possibilities in the optimization view; hence, if there is neither scope for selection nor potential for positive selection, all agents must be optimal.

Received 4 April 2011; revised 18 April 2011; accepted 23 April 2011