Social evolution and the individual-as-maximising-agent analogy

Cédric Paternotte

Title : Social formal darwinism

Highlights:

- Maximising analogies are ubiquitous in evolutionary biology; nature or individuals can be seen as maximising agents.

- Global approaches (which specify maximands from general results) vs local approaches (which generalize maximands obtained in specific contexts).

- Some local approaches, which favour relatedness-based utility functions, lack robustness and cannot be fully generalised.

- It is difficult to identify prior constraints regarding what a 'realistic' utility function is – what agents can or cannot 'really' care about – and so to exclude global approaches as unrealistic.

Title:

Social evolution and the individual-as-maximising-agent analogy

Author:

Cédric Paternotte

Affiliation:

SND research team, UFR de Philosophie, Sorbonne Université

1 rue Victor Cousin, 75005 Paris

cedric.paternotte@sorbonne-universite.fr

Social evolution and the individual-as-maximising-agent analogy

Cédric Paternotte*

Abstract. Does natural selection tend to maximise something? Does it produce individuals that act as if they maximised something? These questions have long occupied evolutionary theorists, and have proven especially tricky in the case of social evolution, which is known for leading to apparently suboptimal states. This paper investigates recent results about maximising analogies – especially regarding whether individuals should be considered as if they maximised their inclusive fitness – and compares the fruitfulness of global and local approaches. I assess Okasha & Martens's recent local approach to the individual-as-maximising-agent analogy and its robustness with respect to interactive situations. I then defend the relative merits of a comparable global approach, arguing that it is conceptually on a par and heuristically advantageous.

1. Introduction

Does natural selection tend to maximise something? Does it produce individuals who act as if they maximised something? These two questions, or maximisation analogies, have long occupied evolutionary theorists. Early formal results suggested that natural selection acts so as to maximise mean fitness (Fisher 1930). In contemporary evolutionary theory, the success of kin selection has caused many to consider that natural selection makes evolutionary individuals act as if they maximised their inclusive fitness. However widely – and perhaps falsely – believed, such claims have been mitigated or not formally established. According to their current interpretation, Fisher's results do not show that we should expect natural selection to cumulatively increase mean fitness; kin selection has led to a number evolutionary analyses but not to a justification of inclusive fitness maximisation. General criticisms about adaptationism, whether due to genetic constraints or to social contexts, have also fueled doubts as to whether natural selection should really be expected to maximise something at all.

Such frustrating results have led to a renewed exploration of the maximisation claims, through a number of frameworks. First, there are two distinct claims or analogies to explore, depending on whether nature or evolutionary individuals may be seen as maximisers. Second,

* SND Research team, Faculté des Lettres, Sorbonne Université. cedric.paternotte@sorbonne-universite.fr

there are two broad ways in which to explore them. Some start from general evolutionary results, which they hope to apply to concrete situations. Others aim to work their way up from maximisation results in specific contexts to more encompassing ones. Is one method preferable? How could we hope to generally determine the domain of validity of maximisation analogies - what maximises what under which conditions and in which evolutionary contexts?

This paper has two related aims. First, it intends to describe some recent works on maximisation analogies and assess their limitations. Second, it focuses on a recently developed approach in order to critically discuss its fruitfulness when compared to a rival approach, derived from Hamilton's rule, which has recently undergone criticism. Most of the paper focuses on the specific case of social evolution and on the question about whether evolutionary individuals maximise their inclusive fitness – which is probably the most widely believed maximisation claim in evolutionary theory. More precisely, I assess Okasha & Martens recent local approach to the individual-as-maximising-agent analogy, its robustness with respect to interactive situations. I then defend the relative merits of a comparable global approach, arguing that it is conceptually on a par and heuristically advantageous.

The paper unfolds as follow. Section 2 discusses two possible understandings of maximisation and two possible approaches – global or local – for exploring maximisation analogies. Section 3 surveys various global approaches used to investigate the maximising analogies in the context of social evolution and underlines their current limitations. Section 4 then tackles local approaches – so-called individual-as-maximising-agent analogies – focusing in particular on recent work by Okasha and Martens, and uses their very method to suggest that the existence of a general function that individuals would maximise is unlikely. Section 5 discusses the issue of individual control, which leads to one way of re-establishing the relevance and heuristic advantage of the global maximisation approach. Section 6 concludes on the fruitfulness of global and local approaches for the maximisation research program in evolutionary biology and mentions one avenue for future work.

2. Kinds of maximisation

2.1. Nature and individuals as maximising agents

Natural selection may be said to lead to maximisation in different senses, which allow one to map existing approaches. A first useful distinction concerns the level at which the maximising analogy is located.

Nature itself may be seen as a maximiser; alternatively, evolutionary individuals (members of Darwinian populations) may be seen as maximisers. These two options constitute two possible analogies through which evolution may be understood.

The nature-as-maximising-agent (NMA) analogy is typically explored in the following way. From the analysis of the dynamics of an evolutionary system, one strives to find a variable that will always or typically increase when natural selection is the only or main evolutionary force. The idea is that natural selection would tend to bring a variable towards its optimum and that this tendency would be most manifest when other evolutionary forces are weak or idealised away. Formally, one looks for a variable based on characteristics of the population, which typically depends on the population members' phenotype, such that it would tend to increase under the action of natural selection.

By contrast, the individual-as-maximising-agent (IMA) analogy is explored by investigating whether biological agents or evolutionary individuals typically behave as if they were maximising a variable – as if they were somehow akin to the rational choice-makers who populate classical economics (and who are supposed to maximise their expected utility). This does not presuppose that evolutionary individual (are likely to) possess elaborate cognitive abilities, but only that natural selection may have shaped them so as to act as if they were striving to maximise something.

The distinction between seeing nature or individuals as maximisers is easily confused with that between the maximand – what is being maximised – being a population-level variable or an individual-level one. Although similar in spirit, these distinctions do not perfectly overlap. Individuals may behave as if they maximised a complex function, which may depend on population-level variables. In other words, that individuals maximise something does not entail that this has to be defined exclusively in terms of individual characteristics [1]

Conversely, natural selection may in principle seem to act so as to maximise a function of individual-level variables. For instance, for a population of altruistic agents in which altruistic acts always bear the same cost and provide the same benefit to others, natural selection may act so as to maximise a function of these cost and benefit, which are ultimately properties of individual traits. Still, it is difficult to imagine that such cases do not involve population-level variables as well – which would happen to be maximised just when all agents maximise some individual-level ones.

[1] One may object that individual may not be analogous to rational choice makers if they maximise variables that are not under or defined at the population-level – more on this in section 5.

This leads to an intuitive link between population and individual-level maximisation. The two analogies would be equivalent when population-level maximisation is equivalent to universal individual-level maximisation – when natural selection acts as a maximiser just if all members of a Darwinian population act as maximisers as well (of some related but possibly distinct variable). However, appearances are deceptive. For it is well-known in both the rational choice and evolutionary literatures that a group of rational or optimal agents may be collectively irrational or suboptimal. The Prisoner's Dilemma, in both its rational and evolutionary versions, teaches us that much. Evolutionarily speaking, that all agents maximise something may put a population in a collectively suboptimal state.

Similarly, nature may act as a maximiser even if most evolutionary agents appear not to maximise anything. Suppose that nature acts so as to maximise the mean fitness, or the fitness variance in a population. Then, it may be possible or necessary (respectively) that the agents' fitnesses are widely different, and that some of them appear to maximise their fitness while others significantly less so. If variation is the fuel of natural selection, nature may even act all the more optimally when individual phenotypes enjoy multiple degrees of adaptation.

Of course, the gap between the nature and individual-as-maximising-agent analogies depends on conceptual and empirical conditions. The aforementioned possible cases of incompatibility may be empirically rare, or conceptually impossible (for instance if all results in which nature maximises something happen to be cases in which individual are also maximisers). Still, as long as we are ignorant of the extension of the set and of the prevalence or rarity of such cases, the nature and individual-as-maximising agent analogies are best kept separate.

2.2. Global and local approaches

A second useful distinction concerns conceptual approaches to maximisation, which can be global or local. First, maximising results may be investigated formally in fully general conditions, for instance by using general equations for evolutionary dynamics from and investigating what maximand they allow, if any. A *global approach* starts from the most general perspective and then operates by adding idealisations or constraints in order to identify a maximand. For instance, one may start with the Price equation (Price 1972) and then investigate whether any variable may be maximised in the absence of drift, with perfect heritability of traits, in the absence of migration, etc.[2] This process of

[2] In its simplified version, the Price equation is this: $\overline{w}\Delta p = Cov(w_i, pi)$, where w_i and p_i are respectively the fitness and the value of a measurable trait for an

gradual restriction of scope is not necessary in principle; it is, however, often motivated by the hope for fully general maximising results.

A second, *local approach* consists in starting from specific evolutionary contexts in which a maximand can be identified, and to explore whether it can be generalised to other contexts or to a wider class of situations – in a word, whether it is *robust* with respect to changes of the evolutionary context. For instance, one may start with a basic evolutionary situation represented by a game-theoretic model in which evolutionary agents do behave as maximisers and then investigate whether they still do in games with different payoffs, more or less available actions and/or players.

The global and local approaches thus operate in opposite directions and could also be labelled top-down and bottom-up approaches. That they operate in opposite directions does not entail that they are exclusive; on the contrary, they are often complementary. As long as maximising results are not established, it makes as much sense to start from general negative results and go local in order to see positive ones appear, or to start from local positive results and go global in order to see when they disappear.

To summarise so far: nature or individuals can be seen as maximisers; one may search for a maximised variable by choosing global or local approaches. This leads to at least three different kinds of projects (rather than four). Using a global approach, one may wonder whether (and if yes what) nature maximises (global NMA), as well as whether individuals do (global IMA). In particular, the global approach can be used to find both population-level and individual-level maximands. By contrast, the local approach is typically not used to investigate whether nature acts as a maximiser. This is because local approaches start from specific evolutionary contexts in which the set of available phenotypes is severely constrained (individuals can typically be of two or three types, with straightforward genotype-phenotype links). But what nature may maximise is probably a function of all possible phenotypes – here the aim remains to understand why the members of Darwinian populations often seem so well adapted. In principle, we should expect any exploration of the nature-as-maximising-agent analogy to be based on a global approach. As a result, in practice the local approach is used only to investigate the individual-as-maximising-agent analogy (local IMA).

In the next section, I turn to the assessment of these three options. Before this, one last clarification is in order. The discussion of the links

---

individual i; $\overline{w}$ is the average fitness of the population. This version is obtained from the general one by assuming that there is no overall transmission bias of the trait in the population. See Okasha 2006: 18 ff.

between natural selection and maximisation may be perceived as a mere corollary of the classic debate about adaptationism. However, this perception would be mistaken. In its methodological guise (Godfrey-Smith 2001), adaptationism strives to explain biological traits by supposing that they once were evolutionarily beneficial for their bearers. A usual adaptationist strategy thus consists in elucidating the possible effects of a target trait on its bearer's fitness and in determining what value of the trait would maximise it. In social contexts, one may rather explore the benefits of a trait in terms of inclusive fitness. The point is that in an adaptationist strategy, the nature of the maximand is typically presupposed rather than justified. It is because evolutionary individuals are thought to be brought by natural selection to act as if they maximised their personal or inclusive fitness that such strategies can be employed. By contrast, the maximising approaches introduced so far aim to *justify* the appeal to a given maximand. Indeed, defenders of such approaches often claim that they want to provide a steady basis for assumptions that are routinely made in scientific practice. Note that this conceptual difference does not mean that the adaptationism and maximisation debates do not share common points – for instance, both have gradually shifted their emphasis from global to local approaches.

3. Global approaches

Over the history of evolutionary theory, the nature-as maximising-agent (NMA) analogy has been developed almost as soon as formal approaches emerged, starting with Fisher's (1930) 'fundamental theorem of natural selection'. According to a naive interpretation of this theorem, the mean fitness of a population always increases under the action of natural selection – that is, it is as if natural selection strived to maximise the average fitness of a population. This diagnosis proved puzzling because of the numerous cases in which average fitness appears to remain constant or even to decrease. One famous example of the former is the so-called 'heterozygote dominance', in which a population's fitness remains constant due to genetic constraints that prevent the fitter types from reaching fixation; one equally famous example of the latter is the evolution of altruism, in which natural selection favours 'selfish' behaviours over altruistic ones, which leads to a decrease of average fitness.

The split between this apparently undisputable formal result (Ewens 1989, Lessard 1997) and biological good sense was fixed by later interpretations. For according to the fundamental theorem, mean population fitness would increase from one generation to the next if the agents'

environment was fixed. However, the environment, which comprises the population's genetic composition, also changes under the action of natural selection. As a consequence, in real situations, mean fitness may increase, decrease or stay constant. In other words, nature acts as if it maximised a precisely defined variable only in ideal circumstances. Although encouraging, this result does not support the general use of a NMA analogy, which we would like to illuminate the actual effect of natural selection.

There exist other attempts to find a level maximand for natural selection. Still, in the remainder of the paper, I focus on the individual-as-maximising-agent (IMA) analogy, for several reasons. First, it has been explored by global as well as local approaches, and so provides a useful contrast between their fruitfulness. Second, it has long been claimed that evolutionary agents behave as if they were maximising their inclusive fitness.[3] This claim, which has enjoyed a renewed interest in recent years, can also be assessed by contrasting global and local approaches. Third, contrary to the NMA, the IMA is compatible with global as well as local approaches, which allows one to compare them.

Grafen's (2002, 2007, 2008, 2014a) 'Formal Darwinism' project provides a recent example of a global approach for the IMA analogy. Grafen intends to provide formal links between population genetics (changes in gene frequencies) and what he calls 'optimisation programs' – the existence of a variable that may maximise a certain function under given constraints. More precisely, Grafen establishes necessary and sufficient links between evolutionary agents solving optimisation programs (that is, having a phenotype that maximises a function) and the population being at equilibrium. Such links can be established without an explicit formulation of the function that is maximised; they can be seen as axioms that any maximand should fulfill (Okasha & Paternotte 2014). This clearly constitutes a global approach: it formulates fully general conditions for the IMA analogy to hold, which are susceptible to lead to the identification of actual maximands in specific contexts.

Applying this framework to the case of social evolution, Grafen (2006) is able to show that inclusive fitness approximates the individual's maximand, under the constraint of additivity (e.g. that "the effects of others on one individuals fitness combine by adding up" (Grafen 2008: 543)). While maintaining the spirit of Grafen's approach, Lehmann and Rousset (2014) counter his result by showing that many

[3] To recall, the inclusive fitness of an agent is the sum of its own fitness and of the fitnesses of its partners in interaction, weighted by their genetic relatedness. For instance, if an altruist agent provides a b benefit to one partner at a personal cost of c, the agent's personal fitness is $b + r.c$, where r is the relatedness coefficient. More on this below.

maximands are possible in a social context (even Hamilton's inclusive effect[4] always provides the direction of selection); again, this only holds under restrictive conditions (Grafen 2014b: 288). Those who do not want to take sides on the respective import of different results obtained in different models and under different constraints are left with a pessimistic diagnosis. Either inclusive fitness is but an approximation of what individuals appear to maximise, or the inclusive fitness effect indicates nothing more than the direction of natural selection. In both cases, nothing guarantees that the IMA analogy holds generally.

There is, however, another defense for the claim that individuals act as if they maximised their inclusive fitness, that is based on the famous Hamilton's rule. Supposing that an altruistic trait that provides a b fitness benefit to its bearer's partner in interaction, at a personal fitness cost of c, the rule states that the trait will spread in the population (that is, that the proportion of the trait's bearers in the population will strictly increase as generations succeed) if and only if $rb > c$,[5] where r is the relatedness coefficient.[6] This may appear to lead to a justification of the IMA, namely the claim that individuals maximise their inclusive fitness. For if an individual maximised its personal fitness, altruism could never evolve: regardless of the benefits one receives from its partners, behaving altruistically is always more costly than not. But if altruism evolves when $rb - c > 0$, individuals can be seen as maximising their inclusive fitness: altruism spreads precisely when an actor's inclusive fitness ($rb - c$) is strictly positive, and as fast as this value is high.

Unfortunately, Hamilton's rule only holds under the assumption that fitness benefits and costs are additive, which has led to criticisms regarding its scope (as well as to that of kin selection in general; see Allen et al. 2013). However, the rule can be generalised so that it holds without any restrictive assumption. That is, one can show that altruistic traits will evolve if and only if a condition of the form $rB > C$ holds.[7] In this version, $-C$ expresses the regression of an agent's fitness

[4] Which in simple cases would be proportional to $rb - c$, with the b, c and r coefficients defined as in fn. 2.

[5] For this result as well as for all evolutionary analyses of sections 4 onwards, we suppose infinite populations of haploid individuals that meet in pairwise interactions. See the Appendix for examples of such analyses.

[6] The relatedness coefficient is defined as the regression of the partner trait (altruistic or not) on the actor's trait, which can be understood as the extent to which an actor's trait allows one to predict that its partner will have the same trait.

[7] Gardner et al. 2011 derive this general version by starting from the expression of the action of natural selection given by the Price equation, and 'partitioning [it] into its direct and indirect components' (1024). See also Birch & Okasha 2015 for an appraisal of Hamilton's rule in general, and Birch 2014 for the consequences on the controversy about kin selection.

on its gene frequency (holding its partner's gene frequency fixed), that is, the extent to which an agent's gene predicts its own fitness; and B expresses the regression of an agent's fitness on the gene frequency of its partner (holding its own fixed), that is, the extent to which an agent's partner gene frequency predicts the agent's fitness. It is thus clear that B and C are purely statistical values that do not necessarily coincide with b and c (although they do in the additive case).

Now, does the generalised Hamilton's rule provide any support for the IMA analogy? Apparently not. As Okasha (2016) rightfully notes, the question of the conditions for the evolution of altruism and that of the apparent goal of individual behaviour are distinct. More precisely, the validity of the $rB > C$ does not entail that individuals maximise their inclusive fitness. Moreover, although the C and B components have been claimed to correspond to "the direct and indirect components of inclusive fitness; the quantity that organisms are designed to maximize" (Gardner et al. 2011), this correspondence is dubious because B and C are not under the individual's control: "the value of that quantity that an individual receives does not solely depend on its own behaviour." (Okasha 2016). B and C are regression coefficients between fitness and gene frequencies, which can be calculated even if the link is non-linear and even if other variables affect an agent fitness. By contrast, b and c express a fitness benefit and cost that directly depend on an individual's altruistic behaviour.

Overall, the global approach to the IMA analogy identifies maximands given certain restrictions; its fully general results do not seem to provide support for the claim that individual act as if they maximised something – in particular not inclusive fitness. In the face of such results, local approaches may seem worth exploring. I now turn to one such approach that relies on game theory – an explicit characterisation of individual rationality in interactive contexts – in order to determine when the IMA analogy holds.

4. Local approaches

4.1. Utility transformations

Following Martens (2016), Okasha & Martens (2016) put forward a new local approach to make the individual-as-maximising-agent analogy precise and to determine when it holds. Their method is as follows. When faced with a given context of interaction, we should follow three steps. First, we should work out its evolutionary characteristics (determine which traits will evolve under which conditions). Second, we

should represent it in a game-theoretic matrix and determine its set of rational strategies (its Nash equilibria). Third, we should compare the conditions for which a trait/strategy spreads by natural selection and for which it rational. If the conditions are identical, then the IMA analogy is justified: a trait will spread exactly in the conditions under which it is rational for an agent to choose it.

Interestingly, this method allows us to explore the form of what rational agents may maximise, if anything. This is because given fitness benefits and costs, we may decide that a rational agent's utility is any function of those, and then test whether this utility function makes it rational to choose the options that would spread in the evolutionary case. In short, one can easily test when an evolutionary situation makes things as if traits were chosen by rational agents that maximise this or that utility function. We become able to precisely identify for what maximand evolution makes agents appear to aim.

	A	S
A	$b-c$	$-c$
S	b	0

Figure 1. Additive Prisoner's Dilemma. The represented payoffs are that of the row player (the game is symmetric).

One example helps illustrate this point. Consider again the situation in which agents may be selfish (type S) or altruistic (type A), in which case they provide a fitness benefit b to their partner at a personal fitness cost of c. Supposing pairwise interactions, the situation can be represent by an additive Prisoner's Dilemma, as in fig. 1.[8] As seen before, the evolutionary analysis shows that the proportion of the bearers of the altruistic trait in the population will increase whenever Hamilton's rule holds ($rb > c$; see Hamilton 1964).

Consider the rational counterpart. If rational agents played the same game, then strategy A could never be chosen. This is because S strictly dominates A – an agent always obtains strictly more by doing S than by doing A, regardless of what her partner does.

	A	S
A	$(b-c)(1+r)$	$-c+rb$
S	$b-rc$	0

Figure 2. Additive Prisoner's Dilemma – inclusive fitness payoffs

Now imagine that rational agents have utility functions that differ from their personal fitness. Suppose for instance that agents maximise

[8] The game is additive because the benefit and cost depend only on the agent's trait or strategy, regardless of what her partner does.

their inclusive fitness, that is, the sum of their personal fitness and of their partner's fitness weighted by the relatedness coefficient. Agents would thus be playing the game shown in fig. 2. In this game, one can show that (A, A) is a strict Nash equilibrium if and only if $rb > c$, that is, precisely if and only if Hamilton's rule holds.[9] There is an equivalence between the conditions for evolutionary success[10] and for rational choice under inclusive fitness maximisation: so we can say that natural selection makes evolutionary agents behave as if they were maximising their inclusive fitness. The IMA analogy is justified.

	A	S
A	$b - c + d$	$-c$
S	b	0

Figure 3. Synergy game

However, different games may justify the IMA analogy but for a different maximand – which is Okasha and Martens's other interesting result. Suppose the interactive situation is now represented by a synergy game (in which agents receive an additional fitness benefit d when both are altruistic – see fig. 3). Here again, there exists a utility function for which rational choice and evolutionary success have similar conditions; but this function is not the inclusive fitness one.

	A	S
A	$b - c + d$	$-c + rb + rd$
S	$(1 - r)b$	0

Figure 4. Synergy game – Grafen payoffs

The authors borrow from Grafen (1979) the following utility function: an agent's utility is the weighted sum of her personal fitness and of what her fitness would have been had her partner acted like her (where the weight of the former is $1 - r$, and of the latter r). The game with Grafen payoffs is represented in fig. 4., and one can

[9] A Nash equilibrium is a set of strategies such that no agent could obtain a strictly higher payoff by playing a different strategy while all her partners keep playing their part of the set. In the any game represented in this paper, in order to check whether (A, A) is a strict Nash equilibrium, it suffices to compare the top row and bottom row payoffs of the game's first column. If the top row payoff is strictly higher, then (A, A) is a strict Nash equilibrium – the row agent would gain strictly less by switching from A to S, given that her partner plays A. In our example, (A, A) is a strict Nash equilibrium iff the row player's payoff when she and her partner both play A is strictly higher than her payoff when she plays S and her partner plays A. Formally, this is the case iff $(b - c)(1 + r) > b - rc$, that is, iff $rb > c$.

[10] That is, for an increase of the proportion of the concerned trait's bearers within the population as generations succeed.

show general links between the existence of various Nash equilibria and specific evolutionary outcomes; in particular, if (A, A) is the only strict Nash equilibrium of this game, then A will spread until it reaches fixation (Okasha & Martens 2016:478-9). Here again, the IMA analogy is justified, although with respect to a different maximand. Finally, Okasha & Martens further note that the Grafen utility function is also a maximand in the additive case, just as inclusive fitness.

This is a promising approach. In the remaining sections, I use it to explore the domain of validity of the IMA analogy, after which I compare it to what a global approach could teach us.

4.2. Robustness issues

What do Okasha & Martens's (2016) results entail for the individual maximisation analogy? First, that it holds for two classes of games: additive Prisoner's Dilemmas as well as non-additive ones (synergy games). Second, that there is a general maximand that is common to all such situations, namely Grafen's utility function. Evolutionary agents only act as if they maximised their inclusive fitness payoffs in the additive case, while Grafen's utility function corresponds to both additive and non-additive ones. Overall, the domain of validity of the individual-as-maximising-agent analogy is extended.

However, the analogy remains bounded because Prisoner's Dilemmas are but one kind of game. Cooperative interactions may be modelled by a variety of games, all of which are possibly relevant to the evolution of social behaviour. Should we expect individuals to act as if they maximize Grafen's utility function in all cooperative interactions, or inclusive fitness payoffs in situations other than the additive case? The following suggests that it depends on the type of games we consider. Evolutionary populations turn out to behave as if agents maximised their Grafen payoffs in any 2-action interaction. However, when interactions involve 3 actions, the analogy does not necessary hold anymore (at least not with respect to Grafen payoffs).

	H	L
H	2	0
L	0	1

Figure 5. Coordination game

As a simple example, let us first consider a pure coordination game (Fig. 5), and apply Okasha & Martens (2016) analysis. The classical evolutionary analysis of the game is well-known. There are two stable evolutionary equilibria, in which the population is composed uniquely of H-types or uniquely of L-types (both H and L are evolutionary stable

strategies). There is an additional unstable polymorphic equilibrium in which two-thirds of the population are Ls and one-third are Hs.[11]

However, we must consider related agents. I show in the Appendix that the change in p (the proportion of H-types) over one generation is positive if and only if $p > \frac{1-2r}{3(1-r)}$ (assuming only natural selection is at work). This condition always holds when $r > \frac{1}{2}$. When not, the value of p for which p will increase (and the size of the basin of attraction of an all-H population state) depends on the degree of relatedness in the population (the higher it is, the lower p needs to be to start increasing). As a consequence, H always is an evolutionary stable strategy, but L only when

$$r \leq \frac{1}{2}$$

	H	L
H	$2(r+1)$	0
L	0	$r+1$

Figure 6. Coordination game – inclusive fitness payoffs

Do agents act as if they maximised one of the previously discussed utility functions? Consider the game's inclusive payoffs, displayed in fig. 6. Here the rational analysis is straightforward, as the game is equivalent to the original one. Both (H, H) and (L, L) are pure, strict Nash equilibria. This corresponds to the evolutionary results that both pure H populations and pure L populations are stable. There is also a mixed Nash equilibrium consisting in playing H with probability $\frac{1}{3}$ and L with probability $\frac{2}{3}$. For the individual maximisation analogy to hold, these probabilities should coincide with the proportions of H and of L types for which an (evolutionary) polymorphic equilibrium exists. In particular, the probabilities of playing H for which H is the rationally preferable to L should correspond to the values of p (the proportion of H types in the population) for which p increases. However, the threshold value of p (for p to be increasing) only equals $\frac{1}{3}$ when $r = 0$. As soon as r is strictly positive, then the frequency value for which the (evolutionary) polymorphic equilibrium is evolutionarily stable starts moving away from the probability value of the (rational) mixed Nash equilibrium. In other words, for non-zero relatedness values, evolutionary agents do not act as if they maximise their inclusive fitness payoffs.

With the Grafen utility function, the game becomes different (see Fig. 7.). (H, H) is always a strict Nash equilibrium (as $2 > r$), and

[11] This entails that any departure from this one-third/two-thirds proportion will ultimately lead to an all-H or to an all-L population.

	H	L
H	2	$2r$
L	r	1

Figure 7. Coordination game – Grafen payoffs

(L, L) is only a strict Nash equilibrium if $r < \frac{1}{2}$, that is, for low values of r (which is also the condition under which L is evolutionarily stable). The mixed equilibrium condition is that H is played with probability $\frac{1-2r}{3(1-r)}$, which, this time, corresponds to the evolutionary analysis.

This should not come as a surprise, even if, surprisingly, Okasha & Martens somewhat undersell the generality of their result. The individual-as-maximising-agent analogy holds in all synergy games for the Grafen utility. But synergy games, as shown in fig. 3, contain 4 possible outcomes, the payoffs of which depend on the linear combination of 3 parameters (namely b, c and d). Once a payoff is arbitrarily fixed, the parameters can be adjusted so as to fit any 2-player 2-action game–that is, any 2x2 game can be seen as a 2x2 synergy game. As a consequence, the IMA analogy holds in general for such games (see the Appendix for a proof that makes this clear for a generic 2x2 game). In Okasha & Martens's words (then only about synergy games): 'This restores the rational actor heuristic' (478), although for all 2x2 games this time.

	A	B	S
A	4	2	0
B	3	2	1
S	2	2	2

Figure 8. Public good game

Let us now consider a more complex situation: a 2-player 3-action public good game (Fig. 8). On the evolutionary interpretation, suppose there are 3 possible types in the population: As cooperate much, while B slightly less. A mutually beneficial outcome may be obtained when two As, or a A and a B, interact but not when two B do. Moreover, suppose that the cooperative behaviour depends on the same specific cooperative allele. Agents A are homozygous for this allele (they possess two), while agents B are heterozygous (they possess only one). Agents S are homozygous for the non-cooperative allele. The allele increases the cooperative benefit as well as the loss from an interaction with a non-cooperator, and its effect is cumulative.

As we use this game as a counterexample, it is sufficient to find one discrepancy between the evolutionary and the rational analyses.[12] Let us check the conditions under which A is an evolutionary stable strategy and under which (A, A) is a Nash equilibrium.

Although more tedious than in the previous examples, the calculations remain simple. One can show that the fitness of A agents will be strictly higher than that of B agents when $p > \frac{1-3r}{2(1-r)}$, and than that of S agents when $p > \frac{1-2r}{2(1-r)}$ (see the Appendix). For instance, for a relatedness coefficient of $\frac{1}{4}$, As are strictly fitter than Bs for $p > \frac{1}{6}$, and than Ss for $p > \frac{1}{3}$.

What is the rational analysis with respect to Grafen payoffs? These payoffs are more difficult to compute with 3 types, because the A and B types are related.[13] However, note that (A, A) is a Nash equilibrium if its payoff (4) is greater than the (B, A) payoff. The latter will be a weighted sum of the possible payoffs for a player who would choose B, that is, a weighted sum of 3, 2 and 1 (the payoffs of the B row of the game matrix). Such a sum cannot exceed 3, let alone reach 4. It follows that (A, A) is always a Nash equilibrium of the modified game with Grafen payoffs. As the evolutionary stability of A depends on r but its rationality does not, there is a disanalogy between the evolutionary and rational analyses. The individual-as-maximising-agent analogy with respect to the Grafen payoffs does not hold anymore.

Two possible options remain. Either there is a utility function of the original payoffs for which the IMA analogy still holds, or there is not. In the latter case, the domain of validity of the IMA analogy is bounded. In the former, it can be extended, although further extensions may only proceed from a case-by-case basis: because the utility functions being maximised may vary depending on the context of interaction. The IMA analogy may be robust even if its specifics are not; but robustness has to be investigated stepwise.

5. Individual maximisation and acceptable preferences

What we have seen so far suggests a limitation for the local approaches to the individual-as-maximising-agent analogy. As bottom-up approaches,

[12] In particular, I do not need to follow the analysis provided in Okasha & Martens (2016) for all possible cases concerning the nature and number of equilibria.

[13] As before, the Grafen payoff of a consequence is equal to the sum of the agent's payoff if her partner had chosen the same action, weighted by r, and of the agent's payoff given her partner's actual choice, weighted by $1 - r$. However, r cannot be simply expressed as $P(A|A) - P(A|S)$ anymore, and in particular not in a way that does not involve p ; see the last part of the Appendix.

barring possible general results over a wide class of games, they need to investigate interactive contexts separately. Even if positive results are obtained for particular games, nothing guarantees that they will extend to other ones. Moreover, the approach depends on basic intuitions regarding plausible utility functions. For instance, while Okasha & Martens find that Grafen payoffs better preserve the IMA analogy than inclusive fitness payoffs, the former are just one that the authors happened to be curious about. No heuristics are offered that may guide us towards relevant utility functions in a given class of games. When previously identified utility functions fail to preserve the analogy, no alternatives naturally present themselves and we should resort to hunches.

This restores some of the appeal of the global approach for the IMA, such as stemming from the generalised Hamilton's rule. Why not favour methods that identify generally maximised functions and work out how they apply in particular cases? Because of the doubts, introduced in section 3, that such functions help identify what individuals maximise – namely because they involve parameters that are not under the individual's control. I now return to this discussion in order to assuage such worries.

To recall, in the additive Prisoner's dilemma, Hamilton's rule, which states that altruism evolves if and only if $rb > c$, may appear to warrant the conclusion that agents act as if they maximised their inclusive fitness. This is because the inclusive fitness of an altruistic agent is precisely the sum of $-c$, its personal fitness cost, and of rb, the relatedness-weighted benefit to its partner. Altruism evolves just when its bearer's inclusive fitness is positive.

However, the generalised version of Hamilton's rule, which states that in any context altruism evolves if and only if $rB > C$, does not allow one to identify a similar maximand. This is because whether B and C accrue to agents does not depend solely on an agent's actions, but also on a host of additional parameters.

	A	S
A	$b - c + d$	$-c + rb + rd$
S	$(1 - r)b$	0

Figure 9. Synergy game – Grafen payoffs

Let us go back to Okasha & Martens's framework and the case of the synergy game (fig. 9). Here, they show that evolutionary populations will behave as would agents trying to maximise their Grafen payoffs, as shown again in fig. 9. Note that here, payoffs explicitly involve the relatedness coefficient r, a statistical parameter defined as $\frac{P(A|A)-p}{P(A)-p}$

(where p is the general proportion of the A type in the population), or as $P(A|A) - P(A|S)$.

	A	S
A	$(B-C)(r+1)$	$-C+rB$
S	$B-rC$	0

Figure 10. Additive Prisoner's Dilemma game – generalised payoffs

	A	S
A	$(b-c)(r+1)+2dP(A\|A)$	$-c+rb+dP(A\|A)$
S	$b-rc+dP(A\|A)$	0

Figure 11. Additive Prisoner's Dilemma game – explicit generalised payoffs. This game is obtained when replacing B and C in the previous game by their expressions in function of b, c, d and $P(A|A)$ – namely $B = b + P(A|A)\frac{d}{1+r}$ and $C = c - P(A|A)\frac{d}{1+r}$.

However, there is a second way to obtain a game that allows for the IMA analogy. Recall that when the local version of Hamilton's rule ($rb > c$) holds, evolutionary populations behave as rational agents faced with the additive Prisoner's Dilemma with inclusive fitness payoffs (fig. 10). Similarly, when the generalized version of Hamilton's rule ($rB > C$) holds, populations will behave as rational agents facing the same game, where b and c are replaced by their generalised versions B and C respectively (fig. 11), where $B = b+P(A|A)\frac{d}{1+r}$ and $C = c-P(A|A)\frac{d}{1+r}$ (Martens 2016).

So there are two ways by which one may try to recover the IMA analogy–to understand evolutionary individuals as behaving as if they were rational, maximising agents. One may consider that the individuals act as if they maximised the Grafen payoffs in a synergy game; we have seen that all 2x2 games can be considered as synergy games, and that the Grafen payoffs warrant the IMA analogy in such cases. Alternatively, one may use another fully general result, that is, that altruism evolves to fixation only when the generalized Hamilton's rule holds ($rB > C$), which corresponds to agents maximising their inclusive fitness payoffs in an additive Prisoner's Dilemma (where b and c are replaced by B and C in fig. 1 and 2).

To repeat there are two different games that allow one to obtain the conditions for the evolution of altruism: either a synergy game with Grafen payoffs (fig. 9 – let us call it the Grafen SG), or an additive Prisoner's Dilemma with generalised payoffs (fig. 10 or 11 – let us call it the general PD). The question now is whether one, and if yes which one, best supports the IMA analogy. We have seen that formally speaking, both approaches can be used without any loss of generality. However,

the IMA analogy is not purely formal–it relies on the notion of a rational agent. Indeed, on this basis several arguments may be offered in favour of the Grafen SG approach; but I think these can be resisted.

In what follows, the discussion turns on the nature of preferences with which a rational agent may be endowed. One obvious difference between the two games is that the general PD features one parameter that does not appear in the Grafen SG, namely $P(A|A)$ – the probability that an altruist's partner is an altruist as well. It is thus crucial to determine whether the IMA analogy is compatible with utility functions that involve parameters such as $P(A|A)$; if we decide it is not, then the Grafen payoff approach will emerge as the only acceptable one.

The rational counterpart of $P(A|A)$ would be an agent's belief that her partner chooses action A given that she herself intends to choose A. But rational choice theory usually maintains a strict demarcation between agents' preferences (or payoffs) and beliefs. Given an agent's preferences (the origin of which is not discussed) as well as her beliefs (taken to be independent from the interactive situation), one is to compute expected utility of all the agent's options and straightforwardly deduce her choice. Beliefs are not supposed to factor in the payoffs, so by analogy parameters such as $P(A|A)$ should not appear in the game payoffs.

This invites at least two rejoinders. First, preferences may be allowed to depend on beliefs in alternative models of rational choice, for instance in those able to cover the sour grapes story, in which the fox's belief that it cannot reach the grapes weakens its desire to obtain them. However, such behaviours may always be labelled as irrational, so the corresponding models should not be deemed adequate counterparts that may guide the IMA analogy. On this point, let me just stress that the strength of this argument hinges on our definition of rationality, which often happens to hinge in turn on the formal models at our disposal.

The second rejoinder to the banishment of $P(A|A)$ from game payoffs is more decisive. For note that both the Grafen SG and the general PD involve payoffs in which features r, the relatedness coefficient. As seen above, r is a statistical value, which to recall is formally defined as follows: $r = \frac{P(A|A)-p}{P(A)-p}$ (where p is the frequency of the A type in the population). We see that $P(A|A)$ already features r, and so in the payoffs of all the transformed games considered so far – whether based on inclusive fitness payoffs or the Grafen utility function. Any criticism of $P(A|A)$ would seem to equally affect r, and as a consequence, either the IMA analogy never holds (not even for the additive Prisoner's

Dilemma, which goes against the consensus opinion in the literature), or it holds for the general PD just as well as for the Grafen SG.

But maybe r can be treated differently from $P(A|A)$. After all, r clearly is of biological significance and so may be a relevant parameter of choice for a rational agent. This is because r measures the average relatedness to one's partner, which relates to the increased odds that altruists receive cooperative benefits (as compared to non altruists). However, the significance of $P(A|A)$ may appear doubtful. Considering an agent of type S who would receive the $b - rc + dP(A|A)$ utility in the general PD (fig. 11 above), Martens expresses such a doubt:

> "But why should the agent care about the value of a synergistic benefit not received [d] (if we assume the agent is selfish), and why should the agent care about a probability, $P(A|A)$, that is in no way relevant? (Indeed, if any conditional probability should be relevant to the agent in this particular case, it would be of the form $P(A|S)$.)" (Martens 2016: 18)

Here, the analogy with rational choice theory is useful again. First, even in a rational setting, that a parameter is involved in a utility function does not entail that the agent should 'care' about this parameter. Suppose that my utility for a consequence is equal to the weighted sum of my material payoff and of my partner's. The adequacy of this utility function in describing my behaviour does not require that I care about the weight themselves. For instance, these weights may represent the strength of my empathetic feelings. The things I care about are the various agents' payoffs. And such utility functions are indeed common in rational choice theory. Whether it be Fehr & Schmidt's (1999) fairness utility function or Bicchieri's (2006) utility function triggered by social norms, various weights are involved, which may express the strength of certain feelings, whether it be one's aversion for inequity or one's tendency for conformism. Indeed, in our case, $P(A|A)$ (as well as r) are precisely weights of this form, that is, they only appear as multipliers of payoffs of the base game (for instance $P(A|A)$ only appears as multiplying the synergistic payoff d).

One may still object that the meaning of $P(A|A)$ is obscure. r can receive a natural 'rational' interpretation, such as the strength with which one cares about kin members. However, $P(A|A)$ is no more enigmatic: as seen above, it may straightforwardly be seen as equivalent to a belief, and beliefs may influence preferences.

This is where the second part of Martens's claim kicks in: even if a probability may in principle influence a payoff, *this* probability should not affect *this* payoff. The payoff of an agent doing S should depend

neither on what may have happened had he chosen A instead nor on the probability with which it may have happened.

Although intuitively appealing, this claim does not stem from characteristics of rational choice theory. There are many instances of utility transformations in which agents care about counterfactual propositions involving different choices they might have made. Take regret theory (Loomes & Sugden 1982), where an agent's utility for a consequence depends on what she may have obtained had she acted otherwise. Or consider again Bicchieri's (2006) social norm utility function, according to which the utility of a consequence is affected by whether alternative options are prescribed by a social norm or not. My point is not based on the validity of such accounts, but merely on the fact that they never triggered any objection from rational choice theorists due to the fact that utilities linked to actions may depend on what may have happened had other actions been chosen. Because this feature is not conceptually necessary for rational choice theory, it should not be used for adjudicating the adequacy of utility functions for the IMA analogy.

This argument also deals with one last possible objection to the use of the general PD. Grafen's utility has the following conceptual advantage: the utility depends only on the payoffs of the agent for the various possible consequences of her action. In particular, the utility for a particular box of the game only depends on the various agent's payoffs that appear on the same line. This may be seen as an advantage for the same reason as before, that is, because an agent should not have to care about what she may obtain had she acted otherwise. (Note that this stance would also exclude the inclusive fitness utility from the list of acceptable utility transformations.) The claim can also be countered just as before, by realising that rationality should be compatible with the agent's caring about counterfactual scenarios in which they may have chosen different courses of action.

Overall, I see no decisive argument left in favour of the banishment of the general PD in favour of the Grafen SG. It is not clear that the former presupposes preferences that may be those of a rational agent any less than the latter does. Here, both the local and the global approaches lead to payoff transformations that equally preserve the IMA analogy.

6. Conclusion

Does natural selection act as if it maximised something, or produces individuals that do? Such maximisation analogies can be explored through global and local approaches, depending on whether we proceed from

general results then applied to specific settings, or from local results then gradually extended to bigger classes of situations. Because some global results have only been obtained under restrictive conditions, or under general conditions that do not square well with maximisation analogies, local approaches have been on the rise. We have focused on one such approach, due to Okasha & Martens, which offers what I think is the conceptually clearest way to test the validity of maximisation analogies in a variety of interactive contexts. However, the approach is heuristically limited, and results obtained so far have not provided any utility function that individuals would appear to maximise in general. I have then argued that a global approach could be used in a similar spirit, so as to help derive context-specific utility functions from a general principle, namely the generalised form of Hamilton's rule.

None of this is meant to lead to a definitive verdict regarding global or local approaches to maximisation analogies. I have only tried to mitigate arguments that would motivate a clear preference for local approaches. I hope I have managed to show that adjudicating between them is not a simple task and depends on subtle considerations. In any case, just as adaptationist methods, all approaches to maximisation analogies can be seen as being part of a research program. Ultimately, identifying maximands or proving their non-existence in a variety of situations is what will allow a definite approach to succeed.

One last word on the limits of the formal approaches considered in this paper. All models involved the evolution of simple traits, depending on a single allele. However, the intuition that evolutionary individuals appear to maximise something often stems from their apparent design, that is, either on the complex nature of some traits or on the intricate layout of several traits. As noted by Birch (2014), a satisfying treatment of maximising analogies should aim to tackle such cases. In short, future models should be able to cover cases of epistatic or complex traits, in the hope that such cases will make it easier to identify a value that individuals appear to maximise.

Appendix

Coordination game

Here we provide the evolutionary analysis of the coordination game with related agents, by following Okasha & Martens's (2016) method.

Note $P(X|Y)$ the probability that an agent's partner is of type X given that the agent is of type Y. By definition, $r = \frac{P(H|H)-p}{P(H)-p}$, where p is the

proportion of the type H in the population and $P(H)$ the probability that a given agent is of type H.

Considering an agent of type H ($P(H) = 1$), from the definition of r we obtain :
$P(H|H) = r + (1 - r)p$.
Considering an agent of type L ($P(H) = 0$), from the definition of r we obtain :
$P(H|L) = (1 - r)p$.

Applying this to the coordination game (fig. 5), we obtain:
The fitness of the H type: $w_H = 2.P(H|H) = 2(r + (1 - r)p)$
The fitness of the L type: $w_L = 1.P(L|L) = 1 - P(H|L) = 1 - (1 - r)p$

From the Price equation, the variation in p from one generation to the next is: $\Delta p = \frac{p(1-p)}{\overline{w}}.(w_H - w_L)$ Δp is proportional to: $w_H - w_L = 2(r + (1 - r)p) - 1 + (1 - r)p = 3(1 - r)p + 2r - 1$

The H will be favoured by natural selection in the sense that $\Delta p > 0$ if and only if $p > \frac{1-2r}{3(1-r)}$. **QED**

2x2 games

	A	S
A	w	x
S	y	z

Figure 12. 2x2 game. The A and S are mere labels and do not correspond to altruistic or selfish types.

Let us consider a general 2x2 game (Fig. 12).

Evolutionary analysis.

The analysis is similar to that of the previous case. Here $r = \frac{P(A|A)-p}{P(A)-p}$. From this definition we obtain:

The fitness of the A type:
$w_A = w.P(A|A) + x.P(S|A) = w(p + r(1 - p)) + x(1 - r)(1 - p)$
The fitness of the S type:
$w_S = y.P(S|S) + z.P(A|S) = y(1 - r)p + z(1 - (1 - r)p)$

As before the conditions for the evolution of the A type would be given

by the value of $w_A - w_S$, which is given by $c + rb$ in the case of the Prisoners Dilemma, and $c+rb+d(r+(1r)p)$ in the case of the synergy game.

Rational analysis.

	A	S
A	w	$wr + x(1-r)$
S	$zr + y(1-r)$	z

Figure 13. 2x2 game – Grafen payoffs

We turn to the rational analysis of the general game for players with Grafen utility functions (fig. 13). Note p the probability that a player plays the strategy A.

The expected utility of a player choosing A is:
$U(A) = w.p + (wr + x(1-r)).(1-p) = w(p + r(1-p)) + x(1-r)(1-p)$
The expected utility of a player choosing S is:
$U(S) = (zr + y(1-r)).p + z.(1-p) = y(1-r)p + z(1-(1-r)p)$

A is the player's unique rational choice iff $U(A) > U(S)$. But we see that the expressions of $U(A)$ and $U(S)$ are identical to the expressions of w_A and w_S (respectively). Their difference will thus be given by the same expression as well. As a consequence, the conditions under which A evolves under natural selection are the same as the conditions for which A would be the rational choice for a player who maximises a Grafen utility function. **QED**

3-player public good game

We now consider the public good game represented in fig. 8. Under which conditions would the A type evolve? Individuals are now diploid: A types are homozygous for the altruistic allele and B types heterozygous. The formula for the relatedness coefficient thus becomes, in the case of an agent of a A type:
$r = \frac{P(A|A) + \frac{1}{2}P(B|A) - p}{1-p}$
Which gives: $P(A|A) + \frac{1}{2}P(B|A) = (1-r)p + r$

For a B type, we get:
$r = \frac{P(A|B) + \frac{1}{2}P(B|B) - p}{\frac{1}{2} - p}$
Which gives: $P(A|B) + \frac{1}{2}P(B|B) = (1-r)p + \frac{r}{2}$

The fitnesses of the various types are as follows.
Type A: $w_A = 4.P(A|A) + 2.P(B|A) + 0.P(S|A) = 4(r + (1-r)p)$
Type B: $w_B = 3.P(A|B) + 2.P(B|B) + P(S|B) = 2((1-r)p + \frac{r}{2}) + 1$
Type S: $w_S = 2$

As a result, we obtain:
$w_A > w_B$ iff $p > \frac{1-3r}{2(1-r)}$
$w_A > w_S$ iff $p > \frac{1-2r}{2(1-r)}$

QED

References

Allen, B., Nowak, M. A. and Wilson, E. O. Limitations of inclusive fitness. *Proc. Natl. Acad. Sci. USA*, 110:20135–39, 2013.

Bicchieri, C. *The Grammar of Society - The Nature and Dynamics of Social Norms*. Cambridge University Press, Cambridge, 2006.

Birch, J. Hamiltons rule and its discontents. *The British Journal for the Philosophy of Science*, 65:381–411, 2014.

Birch, J. and Okasha, S. Kin selection and its critics. *Bioscience*, 65:22–32, 2015.

Ewens, W. J. An interpretation and proof of the fundamental theorem of natural selection. *Theoretical Population Biology*, 36:167–180, 1989.

Fehr, E. and Schmidt, K. A theory of fairness, competition and cooperation. *Quarterly Journal of Economics*, 114:817–868, 1999.

Fisher, R. A. *The genetical theory of natural selection*. Oxford University Press, Cambridge, 1930.

Gardner, A., West, S. and Wild, G. The Genetical Theory of Kin Selection. *Journal of Evolutionary Biology*, 24:1020–1042, 2011.

Godfrey-Smith, P. Three kinds of adaptationism. In S. H. Orzack and E. Sober (Eds), *Adaptationism and Optimality, 335–357*. New York: Cambridge University Press, 2001.

Grafen, A. The hawk-dove game played between relatives. *Animal Behavior*, 27:905–907, 1979.

Grafen, A. A first formal link between the Price equation and an optimisation program. *Journal of Theoretical Biology*, 217:75-91, 2002.

Grafen, A. Optimisation of inclusive fitness. *Journal of Theoretical Biology*, 238:541-563, 2006.

Grafen, A. The formal Darwinism project: a mid-term report. *Journal of Evolutionary Biology*, 20:1243-1254, 2007.

Grafen, A. The simplest formal argument for fitness optimization. *Journal of Genetics*, 87:421-433, 2008.

Grafen, A. The formal darwinism project in outline. *Biology and Philosophy*, 29(2):155–174, 2014a.

Grafen, A. The formal darwinism project in outline: response to commentaries. *Biology and Philosophy*, 29(2):281–292, 2014b.

Hamilton, W. D. The genetical evolution of social behaviour. *Journal of Theoretical Biology*, 7:1–52, 1964.

Lehmann, L. and Rousset, F. Fitness, inclusive fitness and optimization. *Biology and Philosophy*, 29(2):181–195, 2014.

Lessard, S. Fisher's fundamental theorem of natural selection revisited. *Theoretical Population Biology*, 52:119-136, 1997.

Loomes, G. and Sugden, R. Regret Theory: An Alternative Theory of Rational Choice Under Uncertainty. *The Economic Journal*, 92:805–824, 1982.

Martens, J. Inclusive fitness and the maximizing agent analogy. *The British Journal for the Philosophy of Science*, 2016. doi: 10.193/bjps/axw003

Okasha, S. On Hamiltons Rule and Inclusive Fitness Theory with Nonadditive Payoffs. *Philosophy of Science*, 83(5):873–883, 2016.

Okasha, S. and Paternotte, C. Adaptation, fitness and the selection-optimality links. *Biology and Philosophy*, 29(2):225–232, 2014.

Okasha, S. and Martens, J. Hamiltons rule, inclusive fitness maximization, and the goal of individual behaviour in symmetric two-player games. *Journal of Evolutionary Biology*, 29:473–482, 2016.

Okasha, S. and Martens, J. The causal meaning of Hamiltons rule. *Royal Society Open Science*, 3: 160037, 2016.

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