Acta Biotheoretica (2020) 68:453–467
https://doi.org/10.1007/s10441-020-09380-1

REVIEW ESSAY

Essay Review: Exploring the Conceptual Foundations of Post-Hamiltonian Evolutionary Biology—Rationality and Evolution of Social Agents

Samir Okasha. *Agents and Goals in Evolution*. Oxford: Oxford University Press, 2018. 254p. $40. Jonathan Birch. *The Philosophy of Social Evolution*. Oxford: Oxford University Press, 2017. 266p. $19,74

Philippe Huneman[1]

Received: 18 March 2020 / Accepted: 24 March 2020 / Published online: 1 April 2020

Abstract
Evolutionary theorists often talk as if natural selection were choosing the most adapted traits, or if organisms were deciding to do the most adaptive strategy. Moreover, the payoff of those decisions often depend on what others are doing, and since Hamilton (1964), biologists possess conceptual tools such as kin selection and inclusive fitness to make sense of outcomes of evolution in these contexts, even when they seem unadaptive (such as sterility). The link between selection and adaptation through which selection or organisms can be seen as agents, as well as the scope and nature of Hamiltonian conceptions of social evolution, stimulated many formal elaborations (such as, initially, Fisher's "Fundamental theorem of natural selection"), but also raise major philosophical issues about causation and statistics, and about rationality and adaptation or selection. Two recent philosophy books, Okasha's *Agents and goals in evolution,* and Birch's *Philosophy of social evolution*, tackle those question. This essay reflects on them in order to think of those two issues. After having reviewed the books, I try to sketch some philosophical lessons onto which they concur.

Keywords Okasha · Birch · Social evolution · Natural selection · Rationality · Agency · Kin selection · Hamilton · Fitness

✉ Philippe Huneman
philippe.huneman@gmail.com

[1] Institut d'Histoire et de Philosophie des Sciences et des Techniques, CNRS/Université Paris I Panthéon Sorbonne), Paris, France

1 Introduction

Darwin established that natural selection is the major driver of adaptive evolution. From its inception, evolutionary biology thereby tied the notion of selection to the idea of what's good for the organism: it seems that natural selection intrinsically tends towards maximizing the fit between organisms and their environment. According to Darwin in the *Origin of Species*, "natural selection is daily and hourly scrutinising, throughout the world, every variation, even the slightest; rejecting that which is bad, preserving and adding up all that is good; silently and insensibly working, whenever and wherever opportunity offers, at the improvement of each organic being in relation to its organic and inorganic conditions of life" (Darwin 1859). Thus, the very idea of selection plausibly fits an analogy with the careful choice of the best options for traits in an environment. Yet, Darwin himself was unsure about this analogy, since in later editions he used the phrase "survival of the fittest" to avoid connotations of conscious choice and selection.

His view of selection as creating adaptation and design was purely conceptual; but when the modern theory of evolution, in the form of the Modern Synthesis, emerged, selection became understood in the context of a mathematical framework for modeling evolution: population genetics. The question arose how to mathematically make sense of the connection between selection and optimality, as attested by attempts such as Fisher's "fundamental theorem of natural selection" (FTNS).

This question is at the core of the research strategy known as "adaptationism", which assumes that adaptation is pervasive and uses it to issue predictions about organisms. Even though adaptationists typically acknowledge the presence of constraints on selection that prevent adaptation to be achieved overall, the tie between selection and optimality is still assumed, in the sense that selection without constraints should lead to adaptation. For instance, behavioral ecologists, in their quest for the adaptive meaning of traits, tend to view selection as "scrutinising" populations in search for the best adapted individuals and theorize about organisms as "maximizing agents" (Grafen 2014) .

Social evolution is another aspect of evolutionary theory that Darwin struggled with, but on which mathematics has helped us to make progress. In effect, while he himself insisted on selection favoring the good of the individual ("improvement of each organic being", says the quote above), some individuals like sterile workers in beehives seem to present features that benefit others individuals (and are costly, since they sacrifice their own reproduction to work for the queen). The evolution of social traits such as this altruism has famously been a major puzzle for evolutionary biology. William Hamilton's seminal papers in 1964, and his notions of "inclusive fitness" and "kin selection" allowed researchers to get a grip on a set of seemingly paradoxical phenomena from the viewpoint of individual-level selection.

While hugely successful empirical work has been done using the analogy with agents, and knowledge of subtle social traits such as hymenopterans' sex ratio or

alarm calls has been gained using Hamiltonian notions, major conceptual issues still remain unresolved. This has resulted in recurring controversies on those issues: witness the numerous papers coming out for or against inclusive fitness (Nowak et al. 2010), or the hot debates prompted by Grafen's attempt to seek a foundation for the analogy of agency in evolution (Okasha and Paternotte 2014). These are philosophical issues to the extent that they concern the meaning of the major concepts employed by evolutionary biologists.

Two recent monographs propose philosophical analyses of each of these two issues, intending to clarify the conceptual aspects of the ongoing research. Samir Okasha, in *Agents and Goals in Evolution* (Oxford University Press, 2019), exhaustively analyses agential thinking, its justifications and its limits in biology, focusing on the relations between selection and rationality, and utility and fitness. Jonathan Birch, in *The Philosophy of Social Evolution* (Oxford University Press, 2017) offers a complete account of the evolutionary biology of social traits that was initiated by Hamilton's seminal work. I will consider those two books in turn, and will emphasize the many convergences between them, before speculating on some morals for philosophers of evolution.

2 Samir Okasha: *Agents and Goals in Evolution*

Samir Okasha's book provides the most informed examination of "agential thinking" in evolutionary theory. This term does not only denote a manner of phrasing—we say that gorillas "look for" mates, or that a male lion "tries to kill: the offspring of his new mate,—but also refers to "models and explanatory strategies" (3). Choices concerning the latter can make a difference to the practice of science. Humans, for sure, can set goals and try to achieve them as rational agents, while organisms adapt to their environments, in which they apparently strive to survive and flourish; hence adaptation as well as rationality may both appear as a power of pursuing goals. At the most general the philosophical aim of the book is to reflect on these "two senses of purpose" (4), as instantiated by rationality and by adaptation. As Okasha says, the book asks why this agential thinking is so pervasive and considers the relation between agential or intentional talk and the proper evolution of intentionality in biology.

The angle of the book is both analytical and critical. It systematically investigates the justifications for an agential thinking and assesses the prospects of this talk in specific scientific settings. Such agential thinking can take various forms: it may take the shape of an "intentional stance" *à la* Dennett, which attributes goals and beliefs to agents, or it may take the form of rational choice and decision theory (including what Sober (1998) called "heuristics of personification"). Both are used in biological modeling. "Agent", in turn, accepts several meanings, ranging from a minimal notion of "doing something," to a more AI like notion of flexibility of behavior, to the philosophical notion of having purposes, and to the economist's notion of rationality as a behavior apparently maximizing a utility function (14). The philosophical notion—perhaps the richest—is less present in contemporary biology, but the other ones coexist in it.

In the end, Okasha defends a moderate position: he disagrees with philosophers like Godfrey-Smith (2009) and Sober (1998), who regard the use of agential thinking as a mistaken and positively misleading tradition, as well as with biologists such as Alan Grafen, whose program of "Formal Darwinism" (FD) aims to provide full legitimacy to the analogy with "maximizing agents" (the book indeed includes an extensive discussion of Grafen's research program). What is moderate here, is the constant care that is taken to distinguish what can be theoretically or a priori asserted about natural selection, and what should be empirically established. But this position is carved out through an invaluable discussion of almost all aspects of agential thinking and purposive language across evolutionary biology.

The book's structure is based on a key distinction between "type 1 agential thinking" that applies to organisms—in the sense that organisms are supposed to choose their phenotypes in a way parallel to what a rational agent would do—and "type 2 agential thinking" that applies to selection itself—in the sense that through selection "mother nature" chooses alternatives based on an evaluation and maximization of a fitness-like quantity, which exactly parallels rational choice. This distinction follows from the difference between selection, which is the process driving evolution, and adaptation, the expected product of the selection process.

Interestingly, Okasha starts by distancing himself from the common belief (in philosophy of biology, at least) that adaptation is the result of natural selection. Biologists like Grafen have also been keen to emphasize this point, and analyzing it will prove crucial to understand why agential thinking matters, may hold, and should be limited. Indeed, even though adaptations result from natural selection, it is not a priori true that natural selection always results in adaptation (be it the fixation of the most adapted genotype in the population, or the emergence of the best possible phenotype). Granted, philosophers of biology are familiar with the classical adaptationism debate, which focuses on constraints, and assumes that without constraints selection would produce adaptation (labeling 'adaptationism' the debate on the power of selection demonstrates such initial equation of selection and adaptation). But Okasha is interested in something else, namely the fact, established many times by population geneticists, that natural selection alone, without any constraints, may fail to reach adaptation, because of genetic make-up (e.g. heterozygote superiority), or frequency-dependence, or some other reason.

On this basis, the question of the goal-directed character of selection and the legitimacy of type 2 agential thinking boils down to determining the conditions under which selection can be expected to maximise something. A major conclusion of the book is that this type 2 agential thinking is less promising—or its validity less restricted—than the legitimacy of type 1 agential thinking.

The book is full of precious insights that any discussion of agency, rationality or fitness maximisation in evolution will have to integrate. I will emphasize some of them here, and wrap up the discussion with three themes that I see as a major contribution of the book.

Okasha first summarizes the various justifications for talking of agents in biology regarding selection and organisms. Most of them rely either on the idea of rationality, and hence rational choice (which supposes alternative options somewhere, be they in the nervous system of the organism or in the model), or on the idea of

having a goal (which doesn't per se includes alternatives and options). The latter type includes what started with Fisher's FTNS, namely attempts to show that fitness maximization is theoretically warranted by the structure of selection. It is explored in the second section of the book. His examination of the former type leads to an in-depth exploration of the connections between rational choice and natural selection (in Sect. 3).

Justifications for type 2 agential thinking are the weakest because of two crucial issues: environmental variation and frequency dependence. If, as Maynard Smith (1982) argued, natural selection chooses options in the same way as rationality selects alternatives through maximisation of utility, traits should be ranked; yet fitness ranking assumes constancy of environment, which often does not obtain (sometimes as an effect of selection itself). Hence the analogy often fails. If, on the other hand, selection is analogous to a goal-oriented process towards some maximization, the fact that the fitness of traits or alleles can change according to the frequency of those traits—itself due to selection—undermines this analogy.

While this means that the prospects for the overall justification of type 2 agential thinking are bleak, Okasha extensively discusses theoretical approaches to fitness maximization that are rooted in population genetics. The FTNS intends to analytically prove a trend intrinsic to natural selection towards fitness maximisation by equating the mean change in fitness of populations between generations to a positive quantity (additive genetic variance). Okasha surveys strategies used to save the empirical content of the theorem, which mostly depend on a distinction between change directly and indirectly due to selection (through change of environement, which includes genetic background). Finally, Okasha argues that the FTNS can't logically prove that natural selection leads to a fitness maximum but, at most, that "where high degrees of adaptation are found in nature, the twin hypotheses of natural selection plus environmental constancy constitute one possible explanation" (95).

Sewall Wright's notion of a fitness landscape constitute another attempt at grounding trends in the logics of selection. Here, maximisation exists as hill-climbing. And exactly like the FTNS, which conceives of it as a maximization intrinsic to natural selection, possibly counterbalanced by "deterioration of environment", hill-climbing is due to selection and possibly counteracted by "perturbing factors" (83). In both cases, defenders of maximisation see the *quantity* due to selection as essentially positive and regard the *apportioning* of fitness change due to selection and the other fitness changes as an empirical issue. What Okasha shows, is that it is hard to theoretically draw the line between these two things: in the fitness landscapes case, one can hardly say that "non-random mating," which may prevent hill-climbing, is an extraneous factor; in the case of the FTNS, indirect effects of selection on environments are so pervasive that it's hard to see them as logically distinct from direct effects.

The message here consists in deflating the high hopes invested by those early population geneticists into mathematical modeling as a ground for agential thinking type 2. While selection indeed may produce adaptation, the link between them cannot be presupposed because models only tell us what is possible. The actual effects should be empirically attested, which supports "a general message:

that adaptationism in biology must ultimately be justified on empirical rather than theoretical ground" (96). This lesson converges with an analysis of fitness maximisation by Birch (2016), and with what Birch's book deduces from a careful analysis of the Hamiltonian concepts in the case of social evolution, which I will discuss in the next section.

But first, let us return to type 1 agential thinking, about organisms. Okasha argues that it appears in a better shape than type 2 agential thinking, though it includes many more facets. Justifications for agential thinking about organisms rely on several approaches: the flexibility of behavior, which occurs in many animals; the goal-directed aspects of some behaviors, such as hunting prey, and the applicability of the rational choice formalism. Importantly, none of these require ascribing extensive cognitive abilities to organisms.

Philosophers are familiar with the etiological theories of function, which rephrase functional ascription to traits in terms of natural selection statements (Neander 1991); this type of finality is pervasive in biological discourse. Yet agential thinking is something else, since the whole organism is the agent (whereas only traits have functions). Agential thinking therefore demands another kind of Darwinian rationale. It requires something more demanding than mere biological functionality, what Okasha terms *unity of purpose*: "its different traits have evolved because of their contributions to a single overall goal: enhancing the organism's fitness" (29). The "because" here attests that this is a causal condition.

Okasha explores two major aspects of this agency: rational choice theory (extensively in the third section), and Grafen's Formal Darwinism (in chapter four), which intends to provide mathematical links between selection, as seen in population genetics, and adaptation, as studied in behavioral ecology. If these links are realised, then one is entitled to say what the FTNS misses, namely that agential thinking in the form of a maximizing agent analogy is legitimate. However once again Okasha shows that the prospects for an a priori link between selection and adaptation are weaker than claimed. A condition of additivity is assumed by Grafen but it is unlikely to be always realised. Frequency-dependence also threatens the connection. Okasha quickly examines "adaptive dynamics", a theoretical perspective which has been conceived explicitly to address frequency dependent selection, which is pervasive in evolution. Yet, against the expectations of many, adaptive dynamics isn't capable of justifying that organisms are fitness maximizers. Ultimately, all mathematical justifications for type 1 agential thinking should therefore be deflated, since they require some empirical work to be done.

Type 1 agential thinking at least requires unity of purpose, argued Okasha. Such unity is analogous to a condition on rationality in rational choice theory. The third section of the book deals with this pressing issue of the parallel between selection and rationality, and hence utility and fitness. This encompasses two related issues: how does rationality evolve, and why are rationality and selection likely to be understood in the same way? Okasha's analysis provides two answers. Adaptation is seen as a proto-rationality, since it's all about finding best choices in an environment, where utility is defined by fitness. He uses Kacelnik's (2006) distinction between economic, psychological and biological rationality, the first one being about utility and the latter about fitness. Next, he shows that selection should favor economic

or biological rationality over *arationality* (the absence of any reason); but the hard question concerns whether selection can select for *irrationality*, namely choices that contradict the "rational" option, the one that maximises fitness (taken as utility). Assuming that organisms are optimized by natural selection, one would indeed expect that they are rational in the sense of always making utility-maximising decisions when we take fitness as utility. But, Okasha argues, sometimes a "parting of the ways" between selection and rationality of decision-making occurs: "The conceptual link between what is adaptive and what is rational, and the formal link between maximization of fitness and utility, does not mean that one may be reduced to the other" (198). The analysis here is subtle and doesn't deliver a simple message; the distinction between aggregative and idiosyncratic risk, as well as the concavity of the utility function, and the inconsistencies of inter-temporal choice, are elements that allow rationality and selection to come apart, but can often be bypassed by redefining the options at stake. Okasha concludes that "the organism as rational agent heuristics must be treated with care, not regarded as a definitional truth, even on the assumption that the organism's behaviour has been optimized by natural selection" (199).

In exploring type 1 agential thinking, Okasha considers the legitimacy of seeing as agents other things than organisms, namely genes or groups. The empirical issue here is whether they satisfy the unity of purpose condition. Dawkins famously used the agent metaphor to talk of genes; and groups such as swarms or herds are often ascribed some agency. In Okasha's view, the legitimacy of groups seen as agents relies on their showing unity of purpose, which implies the absence of within-group selection. Contrary to Gardner and Grafen (2009), clonality is therefore not enough for agency, since the alignment between individuals and group interest is not always causal but can be merely correlational. In contrast, Okasha proposes an interesting concept to make sense of this unity of purpose, namely the "biological veil of ignorance", taken from Harsanyi in economics. If "individuals are deprived of information", then "the ensuing inability to discriminate between possibilities can restrict individuals to pursue goals" (65). In this perspective, recombination appears as a way to scramble this information, hence meiosis appears as a warrant for unity of purpose. Interestingly, in biology this veil concept works better than in economics, since the units of payoffs are the common currency of fitness value (70). Proving that sometimes evolution is even better fit than economics to implement concepts of rational choice theory is a precious insight of the book, in line with Maynard Smith (1982) intuition that applying rationality concepts to selection is often easier than with economic agents, since fitness is a more objective concept than utility.

Chapter five is devoted to social evolution and inclusive fitness. Okasha emphasized that in Hamilton's rule the costs and benefits are understood in causal terms (as causal effects on fitness), and that some additivity of social actors' contributions is assumed (120). But once this latter condition obtains, the relatedness coefficient can be interpreted as "a measure of how much one player values their partner's payoff", which means that one can use rational choice theory to interpret the rule. This heuristic also works with non-additive payoffs, provided one changes the notion of inclusive fitness as Grafen (1979) did, namely by making "the value an agent places on a [social] action "depend not only upon "the actual

payoff that the action brings", but also on "the personal payoff that would have ensued had their opponent reciprocated and chosen the sam action themselves" (127). It follows that the Nash equilibria of the game played with such payoffs correspond to fixation in an evolutionary process where the values represent fitness coefficients.

I'll highlight three results of this rich and deep analysis. First, Okasha provides us with a sort of internal critique of adaptationism. Gould and Lewontin's spandrels paper (Gould and Lewontin 1979) presented an external critique, based on the idea that selection faces limits that are set by constraints on variation. However, most readers of the spandrels paper took for granted that in absence of external factors, selection will always yield adaptation. Okasha shows that the very possibility of distinguishing what is external from what is internal to the working of selection is not always given and cannot be a priori assumed. Thus, there may be internal reasons, proper to the process of selection, to doubt adaptationism, and one should consider empirically, on a case by case basis, whether those reasons are manifest.

Second, the analyses given here have epistemological consequences. Regarding Formal Darwinism and more generally Fisher-style attempts to prove a priori that selection or organisms maximise, Okasha argues that these presuppose an epistemic stance according to which an explanation ought to show that the explanandum *must* obtain. As Okasha argues convincingly, one should rather favor a conception of explanation according to which explanations show how, assuming the explanans, the explanandum appears *possible*. In this sense, indeed, attempts like the FTNS can be successful. I take this to be a general claim about the relationship between population genetics and behavioral ecology. The workings of selection as modeled by population genetics can only be shown to make adaptation possible, both in terms of mean fitness of the population (FTNS, Fitness landscape) or organisms' strategy choice (FD). In order to move beyond an explanation of what is possible, empirical moves are always required. By taking this view, Okasha's appears to adopt Brandon's idea of "how possibly explanations" (Brandon 1990), as explanations that need to be complemented with other considerations in order to explain what actually occurs. To explain adaptation, 'how actually-explanations' require systematics, paleontology and ecology along with population genetics, to ground an explanatory statement.

The third question concerns the ontological status of 'agents'. There is an ongoing interest in *agency* as an irreducible explanatory property in biology. Denis Walsh, who is the most prominent advocate of this view, takes a stance that is very different from Okasha's (Walsh 2015). While Okasha conceives of agency on the basis of the analogy between selection and rationality, and thereby follows Maynard Smith among others, Walsh (2015) develops 'agency' on the basis of works like West-Eberhardt's on phenotypic plasticity and on the recent theory of niche-construction. Their theoretical and conceptual takes on agency are therefore very different. For Walsh and many others, agency is a genuine property of organisms, while for authors like Grafen or Dawkins, on which Okasha relies, agency is first and foremost a question of heuristics. It thus seems that there are two distinct concepts of agency. A next step in conceiving of agency in biology would consist in comparing and evaluating these two approaches. It would be a modern reappraisal

of the philosophical critique of notions of 'goal' and 'purpose' in biology, as first undertaken by Kant's *Critique of Judgement* before the rise of Darwinian biology.

On this matter, one question: isn't Okasha's notion of agency only an analogy? Everything depends upon the notion of 'rationality' one adopts. Economic rationality doesn't require cognitive abilities; thus one could argue that the same rationality is realised by humans as economic agents and by organisms when they are legitimately seen as maximizing agents. Reason here is not a feature proper to humans, thus, agency is ascribed in proper sense to animals or plants. But if one does not agree on this monism of economic rationality, and, along with Kacelnik, sees three or more kinds of reason, and finds them irreducible, then talking of "agents" in biology may be just a heuristic. In his examination of risk and "parting of the ways" between reason and selection, Okasha states that rational norms can't be naturalized. Thus, it seems that economic rationality, which may often be transcribed into adaptiveness seen as a kind of biological rationality, is not the full sense of reason. Hence "biological agency" names an analogy, and the difference between Okasha's "agency" and Walsh's "agency" relies at least on this dimension of heuristics.

3 Jonathan Birch: *The Philosophy of Social Evolution*

When Grafen elaborated the Formal Darwinism, which supposedly justifies the 'Maximizing agent analogy', he explicitly intended to conciliate Fisher's mathematical defense of type 2 agential thinking with the acknowledgement of social interactions and frequency-dependence as pervasive biological facts. Those facts force biologists to enlarge the measure of evolutionary success beyond one's proper offsprings, hence Grafen conceived of "inclusive fitness" as the proper maximand of selection (Grafen 2006).

Social evolution is the object of Birch's book, which offers an extensive philosophical clarification of a set of issues that were hotly debated since the 90s, and are still the focus of highly-mathematized, theoretical debates. For example, the very notion of "inclusive fitness"—the major tool to handle instances of biological altruism such as sterile ants—was attacked by a paper in 2010 (Nowak et al. 2010), which was met with a response signed by a large fraction of the community of experts on social evolution (Abbot et al. 2010). Other controversies include the competition between inclusive fitness and "neighbor-modulated fitness" (see below) as privileged tools to address cooperation, and kin selection vs. multilevel selection as competing processes to explain altruism.

Birch addresses these debates through a philosophical analysis of what "social behavior" should mean and what possible explanatory strategies they may require. Intended as "'one long argument' for the cogency and explanatory power of Hamilton's ideas" (10) the book investigates the three (often conflated) major Hamiltonian concepts, which ground work on social evolution: Hamilton's rule, kin selection and inclusive fitness. The differences between these concepts yield the distinctions between explanatory strategies in the domain (7). The analysis provided in the book successfully shows that the fierce debates often are due to researchers talking past to each other on those concepts. But Birch's ambitions are wider, and he includes two

more recent areas of research, highly concerned by social evolution: microbiology, and cultural evolution. In these two areas, besides his analysis of the major concepts of social evolution in biology, he offers some appealing empirical hypotheses that will hopefully inspire biological research. While some overlap with the books by Okasha (2006) and Bourke (2011) on Darwinian social evolution obviously exists, Birch's book rather completes those two works—Okasha's deals with multilevel selection, which does not directly enter the Hamiltonian framework, while Bourke focuses on the evolution of individuality, which is one aspect of social evolution.

Birch's analysis relies on one definition (that he substantially justifies) and one strong thesis. He defines cooperative behavior as behavior that: (a) is selected, (b) for the benefits of others,[1] and (c) in a "recent history" regime (23) (namely, it's maintained by selection, whatever its origins). Here he shifts attention from cooperation as a vernacular notion—roughly: what benefits others—characterizing a social behavior, to a theoretical notion likely to support an explanatory framework about social phenomena. "Function" and "adaptation" underwent the same process in evolutionary biology. In each case, becoming theoretically operational involved recognizing the role of selection at the core of the concept. More precisely, the cooperative action is always part of a *strategy* in a *task* (e.g., the action guiding others along the tracks of a bear is part of the strategy 'coordinate with others' in the task 'hunting'). A strategy-relative and task-relative quadripartition of cooperative action follows, depending on the nature of the payoff for the focal individual and for the others.

Hamilton's rule formalizes the conditions under which a behavior can evolve given certain payoffs, as captured in the well-known formula $c < br$. Birch's uses a formulation of the rule by Queller (1992), and modifies it, to propose a "generalized Hamilton rule (HRG)" (38). Here, b and c are regression coefficients of an individual partner's fitness onto her own, or her partner's fitness, and r is also defined in terms of population statistics: they are not "properties of token interactions" (45). Birch's strong thesis is that HRG is neither an axiom, nor a fact, but an "organizing principle" for social evolution theory (39) (exactly like Brandon's idea that "natural selection" is neither a fact nor an a priori law but an explanatory principle for a wide diversity of facts (Brandon 1996)). This means that it is not a law of nature, but it "organizes social evolution research by allowing us to locate specific modeling results in a space of explanations" (50). Birch construes such a "space of explanations" based on two axes, rb and c. Regions defined by the signs of rb and c denote different kinds of explanations. "Selective explanations" occur when $rb > c$. Among these, depending on whether rb and c are together lower or higher than 0, we get explanations based on direct fitness (namely fitness of the actor), or on indirect fitness—or a hybrid case in which rb is positive (indirect fitness) but c is negative (namely, there is a benefit for the actor). The HRG thus provides a typology of explanations, grounded on statistically construed values of coefficients. It allows us to discriminate between explanations, and at the same time entitles us to

[1] Which is exactly as West et al. (2007) define cooperation, hence defining altruism as a sub-case of it in which the actor's payoff is negative.

recognize that diverse causal processes—such as kin recognition, kinship or limited dispersal—may yield the same type of explanation (here, an indirect fitness explanation). The coefficients in HRG are not necessarily a measure of causal influence (of a behavior on someone's fitness), hence a pure causal reading of HRG is impossible, since in the cases of synergy (e.g. non-additivity of payoffs, also explored by Okasha's *Agents and goal in evolution*, see above) *c* and *b* fail to represent causes (73–76).

This framework constitutes a major clarification of two recurring debates—on kin vs. group (or multilevel) selection and on inclusive fitness vs. its alternatives,—which also concern the role of causation vs. statistics in explanations. In both cases, the philosophical issue is the interpretation of what *prima facie* stands as a formal equivalence between two concepts. "It is crucial to distinguish between the formal equivalence of two statistical descriptions of change and the identity (or otherwise) of two types of causal process responsible for change. The former does not imply the latter." (84).

First, regarding the group vs. kin controversy, Birch argues that we have here two causal processes, which rely on two distinct population structures responsible for indirect fitness effects: "kin selection occurs in populations that are structured such that relatives tend to interact differentially, while group selection occurs in populations in which there are stable, sharply bounded, and well-integrated social groups at the relevant grain of analysis." (101) Network analysis and its notion of clustering coefficients and relative density are interestingly used as a way to test where to locate actual populations in a gradient standing between groups made of "sharply bounded subgroups", and neighbor-structured networks where each individuals interacts with its own neighbors. Birch argues that K (kin selection) and G (group selection) are properties of populations (rather than organisms), and here too, he offers a conceptual space in which selection processes can be situated according to their degree of realizing kin selection and group selection. "*K* and *G* can be imagined as the axes of a two-dimensional space, and we can think of kin selection and group selection as large, overlapping regions of that space." (101). K-selection and G-selection, then, contribute specifically to two distinct evolutionary situations: "The significance of *K* lies in the fact that high-*K* populations may support the evolution of stable altruistic and spiteful behaviour—behaviour that is not suppressed by modifier alleles at other genomic loci. The significance of *G* lies in the fact that high-*G* populations meet a basic precondition for an evolutionary transition in individuality. Populations at any level of biological organization can be given a position in *K*-*G* space." (110).

Second, the controversy between neighbour-modulated (or "personal") fitness—where fitness benefits are an "unweighted sum of effects on [the focal actor's] own reproductive success"—and inclusive fitness—where benefits are computed as a weighted sum of contributions of the focal actor on the others' fitness—involves many subtle distinctions. Both concepts were conceived by Hamilton (1964). After being widely used by modellers for a long time, "inclusive fitness" has, in the 80s, gave way to "personal fitness", because some argued it was more mathematically tractable. Both are indeed equivalent in terms of most predictions they allow. However, Birch shows that only inclusive fitness is causally defined, while personal fitness registers phenotypic correlations. Yet, under some conditions about additivity

and weak selection, they are really equivalent. Birch's assessment is nuanced: for personal fitness to properly measure evolutionary success, less conditions are required, but inclusive fitness has the advantage of providing a criterion for "adaptive improvement", which is what matters in cumulative selection. "At all stages in this hypothetical process, the actor's inclusive fitness provides a consistent criterion for improvement: all and only those mutants which differentially promote the inclusive fitness of the actor are favoured." (136).

Yet, concurring with Okasha's analysis, Birch shows that there is no a priori expectation that inclusive fitness should be maximized; and his lesson is the same: "these formal results [can't] support a 'general expectation of something close to inclusive fitness maximization', even in a highly qualified sense. (…) we should not overstate the ability of purely theoretical arguments to support empirical generalizations, no matter how hedged, about natural populations." (138).

Beyond these clarifications, the book offers novel insights on two widely discussed recently areas in which social evolution is at issue. Among microbes, we know now that lateral gene transfer (LGT) is common. This means that bacteria that are involved in group behavior such as producing a "public good" (e.g. a substrate helping them to invade a host) can change their relatedness during their life, so indirect fitness changes may occur across time thanks to this process. Thus, even though social evolution and LGT are often seen as two areas of evolutionary research, a same phenomenon of indirect fitness changes is at work in both evolutionary settings. With LGT, the particularity is that "when organisms are horizontally exchanging genes for social phenotypes at a non-negligible rate, we can no longer even talk of an organism's genic value *simpliciter*. This property may be altered by a plasmid transfer event, and may therefore vary diachronically (i.e. over time) during the course of the organism's life cycle. Strictly speaking, we can only talk of an organism's genic value *at a particular time in the life cycle*." (154, my emphasis).

This leads Birch to rewrite Price equation and the HRG with a genic value that changes over time. This new rule, "HRM", has a coefficient of relatedness r_M such that, "in contrast to the standard concept of relatedness, r_M takes account of genetic correlations between actors and recipients created by horizontal transmission events. These events matter even if they occur (..) *after* the time at which public goods were produced." (156).[2] The natural extension of this HRM is a critical examination of the notion of a "society of cells", sometimes used to talk about multicellular organisms. "In taking a social perspective on the multicellular organism, we are making a methodological bet: we are betting that there are deep and illuminating (rather than superficial and misleading) parallels between multicellular organisms and other complex societies in the natural world, such as eusocial insect colonies, and we are betting that social evolution theory will provide us with the tools we need to explore these parallels." (170) Birch advances another empirical hypothesis here: evolution

[2] A consequence is an empirical hypothesis: "public-goods-producing plasmids may be able to spread by natural selection even if there is no genetic assortment at the moment of social interaction, if they are likely to have an opportunity to transfer horizontally at a later time point into individuals who, by virtue of having been free riders when the public good was produced, are fitter than average." (164).

of multicellularity requires a positive feedback on redundancy of tasks (which permits robustness of a group) upon the size of the groups, such increase in redundancy allowing both larger groups, and then larger possibilities of redundancy. This hypothesis adds on to the set of models we use to understand transitions to multicellular individuality (as explored by Michod (1999, 2005), Bourke (2011) and others).

The last chapter of the book is concerned with cultural evolution. Birch intends to define an analogon of HRG for cultural evolution, by extending inclusive fitness to a notion of "cultural fitness", defined as "the number of apprentices they are able to recruit." (217). Birch argues against the common notion of "cultural group selection": cultural fitness does not need group selection and its demanding requisites. Cultural selection means "selection on differences between individuals with respect to their *cultural variants*." It occurs when transmitted cultural variant impinge onto either the reproductive success of beneficiaries (CS1)—or on their cultural fitness (CS2). Here again Birch offers a "conjecture (…): the course of human social evolution in the Palaeolithic involved a gradual decoupling of cultural fitness from biological fitness, and that there was a gradual transition in the most important form of cultural selection from CS1 to CS2" (201). This conjecture echoes the decoupling fitness hypothesis, according to which transitions towards individuality rely on a shift from "multilevel selection 1" (where fitness of groups is measured in terms of offspring of individuals of the groups) to "multilevel selection 2" (where fitness measures the number of daughter-groups of a group) (Michod 2005).

Philosophically speaking, the affinity between HRC and HRM is that in each case there is a transmission that changes genic values, possibly occurring at any moment of the life cycle. Thus HRM and HRC are two facets of an extension of HRG towards a time-extended theory of fitness change and relatedness; and in both cases the rule has to be applied to an "ideal life cycle" to account for changes over time.

4 Some Reflections on Darwinism's Novel Conceptual Foundations

Those two wide-ranging philosophical investigations explore the conceptual foundations of evolutionary biology. They are complementary, and overlap on three messages:

- The weaknesses of a purely theoretical attempt to formulate laws and trends about selection in general and social contexts (e.g., inclusive fitness as a maximand). As a consequence, the philosophical moral to be drawn of those explorations is deflationary: the conceptual frameworks built by Wright, Fisher and Hamilton allow for an in-depth empirical understanding of evolution, but they don't yield a priori truths about what natural selection, let alone evolution, should produce, and where it should lead.
- Another major theme is the philosophical significance of formal equivalences—be they between forms of fitness, of selection (Birch), or of rationality and selection (Okasha). Sameness of processes and identity of concepts cannot immediately be predicated on the basis of such equivalences. Often, their validity is constrained by some assumed conditions. Equivalences of for-

mulas cannot prove that the causal processes they refer to are identical. And in a pragmatic sense, the choice of one rather than another may pertain to some explanatory interest; for instance, inclusive fitness is not better than personal fitness, except if one is interested in improvement criteria for cumulative selection.

– More generally, the relation between causation and statistics is at the core of those explorations, because many evolutionary *concepts* are causal while *modeling-tools* are statistical. For example, the generality of HRG is gained by considering variables as regression coefficients: the price of this high generality is that one cannot in principle causally interpret such coefficients.

Those two latter points converge towards a general position, which could be termed, if not pragmatism, at least explanatory pluralism. It implies giving up our hopes that the vivid theoretical controversies in biology will go away thanks to a powerful encompassing new theory.

Two final remarks: regression coefficients as used by modelers may receive an interpretation in terms of *information*. Birch sees that relatedness *r* may be seen as an information on the probability that the interactor is more likely than average to be cooperating. In turn, fitness itself could be seen as an information, as argued by Franck (2009). If this is right, the next task for philosophers interested in scrutinizing the conceptual foundations of Darwinism should be to assess those informationally framed formulations of the theory, and especially, under what conditions they can be translated into some of the perspectives discussed above—especially the agential perspective (agents being always information gatherers and emitters).

Finally, a case addressed by Okasha while handling the "parting of the ways" issue invites us to think of deeper parallels between the two books, in a speculative manner. "Inter-temporal choice" in economics names the issue of the constancy of our choices over time. Favoring X over X + dX at time t should, if someone is rational, lead to a specific valuing of X over X + dX at a later time t'. Such discounting of future units of time compared to a present unit should be exponential in theory, but the actual discounting curves are more like hyperbolic ones, attesting what's called a "preference for the present". Yet discounting may also concern "social distance": experiments have considered how much one would value other individuals, in proportion to some kind of emotional or familial distance. For instance, how much would I give (of a fixed received amount) to a brother as compared to the nephew of my cousin, or to my best friend as compared to a colleague (Jones and Rachlin 2006)? The discounting function here empirically matches the hyperbolic shape of the time discounting functions, at least in humans. This prompts a question about the evolution of those two discounting functions: should they be understood on a par? Is there an estimator of "social distance" embedded in the sense of temporal distance? In any case, if there is any non-accidental connection between such two discounting functions, their evolution should tell us something about the connection of irrationality (as a parting of the ways between evolution and rationality) and social evolution. This is only an a hint of the richness of the perspectives opened up by these two groundbreaking books, and of the way they echo each other.

Acknowledgements The author is very grateful to Joeri Witteveen for his insightful remarks, criticisms and suggestions, as well as a thorough language check.

References

Abbot P et al (2010) Inclusive fitness theory and eusociality. Nature 471(7339):E1–E2
Birch J (2016) Natural selection and the maximization of fitness. Biol Rev 91(3):712–727
Bourke AFG (2011) Principles of social evolution. Oxford University Press, Oxford
Brandon RN (1990) Adaptation and environment. Princeton University Press, Princeton
Brandon R (1996) Concepts and methods in evolutionary biology. Princeton University Press, Princeton
Darwin CR (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
Franck S (2009) Natural selection maximizes Fisher information. J Evol Biol 22:231–244
Gardner A, Grafen A (2009) Capturing the superorganism: a formal theory of group adaptation. J Evol Biol 22(4):659–671
Godfrey-Smith P (2009) Darwinian Populations and Natural Selection. Oxford University Press, Oxford
Gould SJ, Lewontin R (1979) The spandrels of San Marco and the Panglossian Paradigm: a critique of the adaptationist programme. Proc Royal Soc B 205:581–598
Grafen A (2006) Optimization of inclusive fitness. J Theor Biol 238(3):641–663
Grafen A (2014) The formal darwinism project in outline. Biol Philos 29(2):155–174
Hamilton WD (1964) The genetical evolution of social behaviour I and II. J Theor Biol 7(1):1–52
Jones B, Rachlin H (2006) Social Discounting. Psychol Sci 17(4):283–286
Kacelnik A (2006) Meanings of rationality. In: Hurley SL, Nudds M (eds) Rational animals. Oxford University Press, Oxford, pp 87–106
Maynard Smith J (1982) Evolution and the Theory of Games. Cambridge University Press, Cambridge
Michod RE (1999) Darwinian dynamics: evolutionary transitions in fitness and individuality. Princeton University Press, Princeton
Michod RE (2005) On the transfer of fitness from the cell to the multicellular organism. Biol Philos 20(5):967–987
Neander K (1991) The Teleological Notion of 'Function'. Australasian J Philos 69:454–468
Nowak MA, Tarnita CE, Wilson EO (2010) The evolution of eusociality. Nature 466(7310):1057–1062
Okasha S (2006) Evolution and the levels of selection. Oxford University Press, Oxford
Okasha S, Paternotte C (2014) The formal darwinism project [special issue]. Biol Philos 29(2):153–154
Queller DC (1992) A general model for kin selection. Evolution 46(2):376–380
Sober E (1998) Three differences between evolution and deliberation. Modelling rationality, morality and evolution. Oxford University Press, Oxford, pp 408–422
Walsh D (2015) Organisms, agency and evolution. Oxford University Press, New York
West SA, Griffin AS, Gardner A (2007) Social semantics: altruism, cooperation, mutualism, strong reciprocity and group selection. J Evol Biol 20(2):415–432