What, If Anything, is Biological Altruism?

Topaz Halperin and Arnon Levy

Department of Philosophy, The Hebrew University of Jerusalem

Forthcoming in *British Journal for philosophy of Science*

1. Introduction

The study of altruism is a cornerstone of modern evolutionary biology. Associated with foundational questions about natural selection, it is often supposed that understanding biological altruism is key to understanding social behavior more generally. Moreover, philosophers and biologists alike have suggested that explaining biological altruism can provide clues—or more than that—about the nature of sociality and norms in humans (FitzPatrick, 2016). For these reasons, traits that are regarded as altruistic, such as sterility in worker ants, warning calls in macaques and blood sharing in vampire bats, are among the best-known and most studied behavioral phenomena.

Typically, the central question about altruism in biology is put in something like the following way: Given that natural selection is a process in which fitter traits become more prevalent, what explains the continued prevalence of organisms that "donate" fitness to others? How could natural selection favor organisms that appear to regularly sacrifice their own survival and reproduction for the sake of others? When first introduced to this problem a common initial reaction—often voiced, e.g., by students in introductory courses—is to object to, or at any rate to wonder about, its anthropomorphic ring, embodied in terms such as 'donation', 'sacrifice' and, ultimately, 'altruism' itself. The ready answer, as anyone who has studied the topic even at an elementary level will know, is that biological altruism is defined in *non-psychological terms*. As Samir Okasha (2013) puts it: "In evolutionary biology, an organism is said to behave altruistically when its behaviour benefits other organisms, at a cost to itself. The costs and benefits are measured in terms of reproductive fitness, or expected number of offspring", Okasha goes on to clarify: "For the biologist, it is the consequences of an action for reproductive fitness that determine whether the action counts as altruistic, not the intentions, if any, with which the action is performed."

This idea, that *biological* altruism is purely an effects-based, behavioral notion, devoid of psychological or intentional content, is widely accepted among biologists and philosophers of biology. Our main goal in this paper is to challenge it: We argue that seemingly "dry", behavioral definitions of altruism carry a vestige of the psychological concept familiar to us from the human domain. Statements like Okasha's notwithstanding, whether an action is seen as altruistic is not simply a matter of its consequences for survival and reproduction alone; the identification of a behavior as altruistic, we will show, relies on an implicit attribution of intentionality, in the form of assumptions about which of the interacting organisms is the "initiator" or "author", of the interaction. In effect, designating a behavior as altruistic assumes that organisms (in general) can be seen as agents, in a fairly rich sense of the term. We aim, first, to show that this is the case. And, second, to raise doubts about whether the relevant notion of agency can be cashed out in an acceptable way. If both parts of this argument are successful, it calls for a substantial rethink of the notion of biological altruism.

The issue we have in mind is perhaps best approached in the following manner. In any two-way biological interaction with fitness consequences, there are three possible generic outcomes: Either the fitness of both parties increases ([+,+]), or both suffer a decrease in fitness ([-,-]) or, thirdly, one interactant's fitness increases while the other's decreases ([+,-]). Cases of altruism fall into the latter category, of course. But altruism is not alone in this. Crucially, behaviors of the "converse" sort, i.e. apparently antagonistic interactions such as predation and parasitism, are also plus/minus interactions. To distinguish these from altruism one must go beyond sheer fitness consequences. One must find a rationale for thinking of one of the interactors as "contributing" fitness to its partner—a case of altruism—or as selfishly "extracting" fitness from the partner. Put differently, to designate an interaction as altruistic (or selfish) is to implicitly assume an active/passive distinction. If the fitness benefits accrue to the active side, the behavior is selfish. If the active side loses fitness, it is altruistic. Our question, then, is what makes it the case that an organism should be seen as active, with respect to a certain gain or loss of fitness.

We start by looking at relatively simple definitions of altruism, and relatively simple exceptions to them. These present substantial, though not overwhelming, challenges to the concept of altruism. We discuss some potential solutions, noting their benefits and drawbacks. But the primary significance of this discussion is that it clarifies and clears the ground for harder cases – ones in which we think a subtler notion of organismal agency is at play.

These harder cases are also very important ones, pertaining to social insects. This is no doubt a key taxon for the study of social evolution. In particular, the phenomenon of worker sterility has been perhaps the most widely studied altruistic trait, and has been seen by many as a make-or-break case for theories of the evolution of altruism.[1] But we show that it is not straightforward to justify thinking of sterility as altruistic on the part of a worker ant (or bee), as opposed to an instance of selfishness on the part of the queen. More specifically, to make such a designation one must appeal to a more fundamental distinction between actors and recipients – one must first decide who the actor is, and only then can one ask whether the actor acted altruistically. We examine the relevant notion of actorhood and suggest that it is hard to make good sense of it.

A couple of preliminary remarks will help clarify our aims and assumptions. First, we assume that whether or not the organisms in the examples discussed below have the capacity to form intentions and execute deliberate action is irrelevant to our analysis. For, as noted, *biological* altruism is assumed to be independent of actual intentions. Even if some non-human organisms are capable of intentional action, models of biological altruism are supposed to apply far more broadly, potentially to *any* organism – including insects, bacteria and plants. 'Biological altruism' is meant to name the very same phenomenon across all these contexts (Bourke, 2011; Dawkins, 1976; Lewens, 2015; Okasha, 2013).

Second, our discussion, as should be clear, is centered on behavioral traits. General questions arise in this context – importantly, one can ask after an account of the notion of a trait: Which effects or dispositions of an organism count as its traits (as opposed to mere by-products)? How does one individuate a trait? Must traits be modular? Must they be products of evolution by natural selection? (Wagner, 2001) While we think these are important philosophical questions, we will not address them in a general manner here, as they go well beyond our main focus. We think, at any rate, that any reasonable stance on traits and their individuation will leave our core argument intact.

Third, when discussing behavior and its evolution, it is common to speak in terms of *strategies*—rather than particular actions—as the objects of selection. "[A] 'strategy'" says John Maynard Smith "is a behavioural phenotype; i.e. it is a specification of what an individual will do in any situation in which it may find itself" (Maynard Smith, 1982, 10). In this sense, a strategy is

[1] This is not to say that the idea that worker ants are altruists is universally accepted. A well-known paper by Nowak, Tarnita, & Wilson (2010) questions whether social insects display altruism (as well as the kin selectionist framework within which this phenomenon is often studied). But this is an exception; by and large social insects are seen as a major case of altruism. There are very few, if any, biologists who take altruism seriously as a biological phenomenon and do not regard social insects as one of its key instances.

a function from situations to actions describing what the organism is expected to do in each of them. We are not critical of the concept of a strategy per se. We do not view *this* concept as intentional or as otherwise problematic (Birch 2017, §1.5). But our focus is on actions, i.e. the constituents of strategies. Our claim, seen from this perspective, concerns how actions are assigned to organisms: What makes a given action the action of this organism rather than that one? This is a more basic question than specifying strategies, since strategies are made up of actions. In other words, if actions cannot be assigned to organisms in a satisfying manner, then neither can strategies. Consequently, and for the sake of simplicity, our discussion is conducted in terms of actions, avoiding the notion of a strategy.

Fourth and finally, it is possible to restrict altruism, by definition, to interactions among members of the same species (Bertram, 1982; Hamilton, 1963). Some of the problems we discuss, especially in the earlier parts of the paper, will be obviated by such a move. But not all. Be that as it may, we think a definitional restriction to conspecifics is unreasonable. For one thing, it has possible exceptions (Pitman et al., 2017). For another, it is unclear why altruism should be so restricted while selfish interactions, such as predation and parasitism, are not. Lastly, it is often claimed that a prediction of kin selection theory, perhaps the central framework for theorizing about social evolution, is that inter-specific altruism should not exist. In accordance, the empirical finding that altruism towards members of different species is rare can be taken as evidence for kin selection (Bourke, 2011, 71, 76-77; Foster, Wenseleers, & Ratnieks, 2006). If altruism is restricted to conspecifics *by fiat*, then the fact that it occurs exclusively among conspecifics cannot play this evidential role. Thus, we proceed by assuming that even if (as seems likely) altruism *in fact occurs* primarily between conspecifics, that is not a fact that we should rely on in spelling out *what altruism is*.

2. Effect-Based Definitions: (Relatively) Simple Issues

We begin, as noted, by discussing problems that pose relatively mild challenges to the standard definition of altruism. The goal, to be clear, is to problematize the concept of altruism – specifically, to show that it is not a straightforwardly “austere” concept – it is not merely a matter of the loss and gain of fitness, but rather embodies intentional intuitions and descriptive habits.

Usually, biological altruism is defined simply as an interaction in which one organism increases the fitness of others at its own expense. For example, Bourke (2011, 28) says:

> Altruism is defined as the social action in which the actor (or altruist in this case) loses offspring and the recipient (or beneficiary) gains offspring.

Likewise, Lewens (2015, 146) writes:

> A *biologically altruistic behavior* is usually understood by evolutionists to be one that augments the ability of others—call them "recipients"—to survive and reproduce, while damaging the survival and reproduction of the organism producing the behavior—call it the "actor". In other words, altruistic behaviors increase the reproductive fitness of recipients while reducing the reproductive fitness of actors. [Italics and quotes in original]

Very similar definitions are given by other prominent philosophers and biologists, including Hamilton (1963), Trivers (1971), E.O. Wilson (1975), Dawkins (1976), Maynard Smith (1980), Hoffman (1981), Kitcher (1998), Sober & D. S. Wilson (1998), Godfrey-Smith (2013) and Okasha (2018).

It should be noted that, as the quotes from Bourke and Lewens demonstrate, even simple definitions of biological altruism are typically couched in terms of an 'actor' and 'recipients'. In later sections we will discuss this distinction in detail. For now, however, we proceed as if it is clear – the actor is the focal organism which "performs" or "exhibits" the behavior in question.

A first, relatively minor issue with these austere definitions is that they count accidental fitness contributions as altruistic. If a deer slips into a pit, a wolf may obtain an easy meal. In falling, the deer increases the fitness of the wolf at its own expense. The fallen deer may also increase the fitness of other deer which thereby avoid becoming the meal themselves. Nonetheless, the fallen deer would not be classified as altruistic toward the wolf or its fellow deer. Conversely too: Suppose that while worker ants construct a nest, a rainstorm erupts and the ensuing flood destroys the fruits of their labor. Here, the ants' efforts don't actually contribute to the fitness of the queen, and hence cannot be considered altruistic according to the simple definition. However, the building of a nest by worker ants is seen as altruism par excellence. Thus, a definition that focuses on an interaction's sheer fitness profile does not align with behaviors regularly thought of as altruistic (J. Wilson, 2002).[2]

A possible amendment is to require that altruistic acts contribute to the recipient's fitness in the majority of cases, on average, in greater probability, etc. But this restriction helps only slightly.

[2] It may be said that in these cases it is intuitively clear that the deer's falling is not an action or a behavior, but merely an extrinsic accident that befell it. Likewise, it seems eminently intuitive that the ants' action—building a nest—was disrupted by bad luck, an event outside of their control. Such intuitive descriptions rely on a notion of *actor's control* – to be discussed at length later in the paper.

Systematic, highly probable non-altruistic fitness donations are ubiquitous. Most sea turtle hatchlings making their way to the ocean are devoured by predators, such as seagulls and herons, while still on the beach. But it hardly makes sense to consider that a display of altruism on behalf of the hatchlings. Similarly, slower deer that fail to escape predators increase the fitness of predators (as well as that of their fellow deer) at their own expense, but it would be inapposite to regard this as altruism by the deer. Finally, it is perfectly possible that in some environments most ant nests are swept away by rainstorms; yet that would not lead biologists to classify worker ants as non-altruists.[3]

3. A More Teleological Approach?

A purely effect-based definition faces serious counter-examples. This suggests that altruism involves something beyond strictly facts about fitness effects. It is not enough that one side gains and the other loses: altruism involves a *directed donation of fitness*. In other words, there is some sense in which altruistic acts are geared towards, or aimed at, increasing the fitness of their partners. Intuitively, in cases of deer falling into pits or escaping too slowly, the beneficiary gains a fitness advantage "for the wrong reasons": The deer do not "aim" to aid the predator. If anything, the opposite holds – there appears to be a sense in which such non-altruistic sacrifices for the sake of others are unintended or misdirected, and it is this sense, we think, that drives the judgement that these cases aren't *bona fide* altruism.

In view of problems like this, J. Wilson (2002, 87) suggests the following definition:

> An organism's behavior is biologically altruistic if and only if it is directed towards another organism with the *goal* of providing a benefit for that organism and where that benefit would have a propensity to cost the acting organism. [emphasis added]

Unfortunately, Wilson does not precisely define 'goal'. The most common way of understanding goal-directedness in this context is as signifying that the organism tends to reach the same end result (the goal) under different conditions, typically due to internal mechanisms that

[3] A more general concern is that these points reflect problems in the concept of fitness, rather than altruism. Indeed, fitness, and especially its connection to accidents and other chance events, raises vexing issues (Griffiths, 2008). But these are orthogonal to the claim we are making: if accidents affect deer survival and reproduction systematically, then surely they count toward deer fitness, and our point stands. If they don't, then such accidents are irrelevant for evolution, and *a fortiori* the interactions at issue are neither selfish nor altruistic. Either way, the question we are raising is separate from general worries about chance and fitness.

adjust its behavior given prevailing conditions (Boorse, 1976; Enç & Adams, 1992; Nagel, 1977). So understood, Wilson would say that an organism is altruistic when it adjusts a self-sacrificing behavior according to the context, so that overall its recipient benefits regardless of changing conditions.

This successfully resolves the issue of accidents. A fallen deer does not have the goal of conferring benefits on the wolf – it does not adjust its behavior to achieve that outcome. Moreover, given Wilson's definition, a prey's inability to avoid predation isn't an act of altruism towards predators. If anything, a slow-fleeing deer has the goal of escaping; it simply fails to meet that goal in this instance. Conversely, according to Wilson, behaviors can be seen as altruistic even if they do not benefit the target. Workers constructing a nest have the goal of providing shelter to their nest-mates, whether or not they are struck by a storm or get consumed by an anteater in the process.

However, the goal-directedness definition too faces problems. Most seriously, the definition countenances behaviors of hosts towards their parasites. For example, by taking care of the cuckoo chick, the warbler's goal is to benefit the chick: the warbler reaches the same end result – an adult cuckoo – despite changes in conditions. Moreover, the behavior of the warbler tends to reduce its own fitness, since attending to the cuckoo's chick comes at the expense of its own offspring. Wilson acknowledges that his definition of biological altruism includes parasitism and that restricting the definition to intraspecific instances doesn't resolve the problem; he is willing to bite this bullet, since, in his view, the behaviors of hosts involve a recognition error (e.g., the warbler misrecognizes the cuckoo as its own chick).

We think this particular bullet may perhaps be bitten, but it would then be hard to swallow. For there are many cases of parasitism which do not involve a recognition error. We illustrate using a case that will recur later in the paper, involving the terrestrial drumming katydid (or the oak bush-cricket; *Meconema thalassinum*). Being unable to swim, the katydid typically hops away from ponds. But upon digesting parasitic hairworms, nematomorphs of the *Paragordius tricuspidatus* species, the katydid jumps straight into ponds and immediately drowns. The parasitic hairworms then burst out of its body and into an aquatic environment required for *their* reproduction (Biron et al., 2005). Such "suicidal" behaviors on the part of hosts are quite common. If the host's behavior has a goal, it is to allow for parasite reproduction, by immersing in water. Yet such interactions are standardly seen as parasitic, and so as selfish, rather than involving altruistic self-sacrifice. This seems to us fatal to Wilson's suggestion: operating with a definition of altruism that applies to such a paradigmatic case of selfishness all but erases altruism as a distinctive behavioral category.

To be sure, more can be said about goal attribution in biology. A direct response to our last analysis is that we've attributed goals incorrectly. Perhaps hosts like the katydid do not have the goal of benefitting their parasites but are "manipulated" into doing so. Such descriptions are fairly common in the parasitology literature. Their underlying rationale seems to concern behavioral *control*: To say that an organism has been manipulated is to say that its behavior has somehow been subverted; control of its actions has been wrested from it by means of deceit. In section 5 we discuss the notion of control in some detail, identifying significant problems with it. But before we get there, we will consider the possibility of cashing out the teleological component of biological altruism using a selected-functional definition.

4. Appealing to Selected Functions?

Jonathan Birch (2017) has recently suggested that altruism should be defined, in part, by refernce to a trait's adaptive history (see also West, Griffin, & Gardner, 2007). Birch's full-blown definition of altruism is relatively complex. For our purposes we analyze this simpler formulation he provides (Birch 2017, 23):

> A behavior is altruistic if and only if it has, in recent history, been maintained by selection because of its positive effect on the reproductive success of other organisms, and despite its negative effect on the reproductive success of the actor.

As with a definition that appeals to goal-directedness, characterizing biological altruism in terms of adaptation removes the difficulties associated with accidental fitness benefits and predation. However, even if altruism is defined as an adaptation, this does not block the inclusion of parasitism. Generally speaking, parasite-host interactions evolve through natural selection because they benefit the parasite and despite decreasing the fitness of the host. Thus, assuming a Birch style definition, it should be seen as altruistic.

Let us spell this out a little further, since the point is relatively subtle. There is no doubt, we take it, that parasites like the hairworm benefit from the relevant interactions, and that they have been selected to so benefit. One way to look at the situation—the common way, to be sure—is to treat this benefit as deriving from the active, "manipulatory" behavior of the parasite. This results in classifying the behavior as selfishness on behalf of the parasite. But we can also focus on the host, and ask about its causal contribution to the evolution of this interaction. Here, we can break down the question in two, relating to origin and maintenance. As regards origin, it may seem that the parasite is justifiably treated as in the driver's seat: Isn't it exploiting a vulnerability on the part of the host, e.g. a vulnerability of the katydid's nervous system, that allows the parasite to cause it to jump into ponds? Perhaps, albeit absent detailed information about the mechanisms underlying the interaction, it is hard to be certain.

That said, our main point concerns the *maintenance* of parasitic interactions. For once we attend to maintenance, we must ask after the selection pressures on the host side: Why have hosts not evolved mechanisms to block the parasite? Why do they "allow" the parasite to "exploit" them? Indeed, this is a salient question for parasitologists. Moreover, Birch's definition directs us to these questions, for it explicitly requires that selection operated to maintain the altruistic behavior in recent evolutionary history.[4]

Now, very often, the answer to such questions is given in terms of the costs to the host – evolving the appropriate defenses is not worth it, in fitness terms (Poulin, Brodeur, & Moore, 1994). In other words, the maintenance of the host's vulnerability to the parasite *is fitness enhancing on average, relative to evolving the necessary defense mechanisms*. Thus, in many cases, selection has indeed maintained the host's behavior "because of its positive effect on the reproductive success of other organisms [i.e. the parasite] and despite its negative effect on the reproductive success [of the host]". Therefore, it can readily be fitted into Birch's definition, with the result that (at least some) hosts qualify as altruistically benefitting their parasites. We think this is an unacceptable result, one that threatens to empty the notion of altruism of explanatory content.

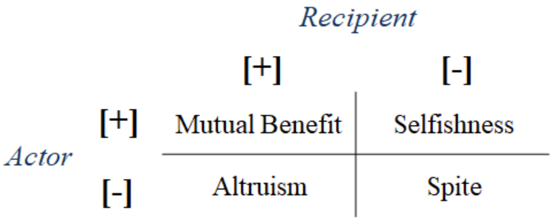
It should be noted that Birch does not consider parasites, in this context. West, El Mouden, & Gardner (2011, 236), who offer a very similar definition of altruism, dismiss the idea that hosts can be regarded as altruistically helping parasites. They suggest that the behaviors in question, i.e. those that hosts exhibit towards their parasite, were selected for in contexts where they are adaptive to hosts. Thus, in effect, parasites "exploit" a pre-existing behavioral adaptation of the host. Such a claim arguably holds for the warbler-cuckoo system, assuming the cuckoo "exploits" a behavior that warblers have been selected for exhibiting towards their own offspring. But this does not work in general: The katydids' jumping into water has not, as far as is known, evolved in a context where it was adaptive to katydids. Indeed, katydids have stayed away from water unless parasitized. We conclude that appealing to selection history will not solve the problem.

[4] Birch adds this requirement for good reasons: both general considerations having to do with function ascription (Godfrey-Smith, 1994) and because, typically, questions about an altruistic trait's maintenance are the most pressing ones. Furthermore, essentially the same argument applies to similar definitions of altruism that take adaptation into account, e.g. West et al.'s (2007) definition in terms of *current* selection pressures.

Here is where we have arrived in our overall argument: Some plus-minus interactions are described as selfish while others are taken to be altruistic. We have asked what distinguishes the two sorts of cases. Is there some identifiable feature of the situation that merits our attribution of intentional-like behavior, such that the minus side should be seen as “contributing” fitness to the plus side, as opposed to the plus side “extracting” fitness from the minus side? We have looked at a number of suggestions for such a feature—systematicity of effects; goal-directedness; selection history—and found them all wanting. Our next step is to examine the actor/recipient distinction. We suggest that while this distinction may appear clear, it is in fact blurry and anthropomorphic.

5. Altruism and the Actor/Recipient Distinction

As we’ve noted—and as quotes brought above demonstrate—definitions of altruism are often explicitly couched in terms of an actor/recipient distinction. With such a distinction in place, it is straightforward to define altruism: An organism behaves altruistically if it *acts* in a manner that results in a loss of fitness and thereby increases the fitness of a *recipient*. This can be seen by examining the following table (*Table 1*), due originally to Hamilton (1964), but now very commonly used to classify social interactions:



		Recipient	
		[+]	[-]
Actor	[+]	Mutual Benefit	Selfishness
	[-]	Altruism	Spite

Table 1. Hamilton's four-way scheme. [+] and [-] signs indicate whether the organism gained or lost fitness in the interaction.

As can be seen, the table embodies an asymmetry: A social interaction’s fitness profile alone cannot tell us whether it qualifies as “nice”—altruism or mutual benefit—or as “nasty”—selfishness and spite. One must also place it in the right column, i.e. decide which party to the interaction is the actor and which is the recipient. In particular, both altruism and selfishness have the same fitness profile (+/-). What distinguishes them is that in altruism the benefits accrue to the recipient whereas selfish actors benefit themselves.

Our discussion so far has effectively assumed that the “actor” is simply the organism that appears to be performing the behavior in question (see Section 1; West et al., 2007). But this leads to trouble, as appearances can readily mislead: Why not view the katydid as “performing the behavior” (in which case its behavior would be altruistic)? We need a better way of understanding the idea that in an interaction among organisms one side is an actor while the other is a passive recipient. Specifically, we need a criterion that allows us to classify "bad" cases, like the worm-katydid, as instances of selfishness, and "good" cases, like social insects (discussed below), as involving altruism.

5.1. Actors and causal control

An intuitive sounding suggestion is that being an actor is being responsible for an outcome. We can put the resultant criterion roughly as follows. An individual is an actor, relative to a focal action, in virtue of exerting *causal control* over said action. This idea has echoes in philosophical discussions of agency in general (G. Wilson & Shpall, 2016), namely, that a behavior *stems from* the individual, and/or its parts, and in this sense is under its control. Samir Okasha, for example, considers the contrast between "an insect colony's moving when it swarms with its moving when displaced by a hurricane. In the latter case, external factors wholly account for the movement; in the former, external factors, for example, ambient temperature suitable for swarming, are at most background conditions. Thus swarming is something that the colony does, not something that happens to it" (2018, 12). Okasha regards such cases as illustrative of "a minimal notion of agency" which is "no doubt hard to make precise" (ibid).

Such a minimal notion may well suffice for some purposes, including those Okasha puts it to use for. But we do not think it will do in the context of distinguishing altruism from selfishness. That is because in many social interactions both sides of a behavioral interaction can and do exert causal control. Thus, both will be deemed actors, and we will not be able to place them in Hamilton's table.

Consider, for instance, a worker ant's sterility. There are several sorts of mechanisms responsible for sterility, but none of them appear to match the idea that control, understood causally, resides with the worker. For example, in some species, workers have fully functional ovaries, but sterility is maintained by primer pheromones released by the queen or by other workers (Conte & Hefetz, 2008). In other species, outside temperature or nutrition at the embryonic, pupae and larval stages prevent offspring from developing functional ovaries (Abouheif & Wray, 2002; Khila & Abouheif, 2010). In such cases, arguably, whatever determines the ambient conditions (perhaps Mother Climate) is in control. It is not the worker-to-be. Other cases exhibit an even greater mismatch with the idea of worker control. For instance, in some ant species the queen fertilizes worker eggs with sperm from other species, while the eggs of winged queens are fertilized with sperm from conspecifics (Cahan & Vinson, 2003). If inter-species hybridization reduces workers' fecundity, then sterility appears partly under the control of the queen, and partly, perhaps, of males of a different species (!). Finally, in some ant species, workers may manage to lay male eggs, but these are eliminated by nest-mates (Ratnieks & Wenseleers, 2008). Clearly, in such cases workers do not control their fertility—they do not "actively forgo" reproduction—rather, their offspring are eaten. Hence, according to the empirical evidence, workers generally do not causally control their behavior.[5] The upshot is that sterile workers, a paradigm case of altruism in the biological world, cannot be regarded as altruistic.

Perhaps it could be argued, in response, that even if sterility is not entirely up to workers, they still "actively acquiesce" in the external factors that cause it (see Bourke, 2011; Ratnieks & Wenseleers, 2008). On this view, the only thing that the above discussion shows is that sometimes all parties to a given social interaction should be seen as actors. This way, even if the queen selfishly induces workers' sterility, they are still altruistic. But the move looks less appealing as soon as we recall the need to maintain parity between parasitism and altruism. Why not treat parasitic interactions in the same way? In "allowing" themselves to be "exploited", cannot hosts be equally seen as actively acquiescing in a parasite's manipulation – and hence as altruists? In other words, if the requirement of parity is heeded, this route too leads to a collapse of the altruism/selfishness distinction.

Perhaps, then, we should opt for a subtler causal criterion. Birch (2013) makes such a suggestion, relying on the concept of systematic counterfactual dependence (Lewis, 1973). [6] The idea, in essence, is that control amounts to fine-tuned causal dependence: if both the cause and the effect have multiple possible states, and if particular states of the effects are counterfactually dependent on particular states of the cause, then the causal relationship in question is fine-tuned (or *specific*). The claim, then, is that if there exists a specific causal relation then we can speak of

[5] We could have extended the discussion to the [+] side of workers' behavior, e.g. constructing the nest or rearing the queen's brood. To the best of our knowledge, workers do not have primary causal control of these behavioral traits as well. That said, since there is no biological altruism without costs to fitness, it suffices for our argument to show that workers do not have causal control over the [-] side of the interaction.

[6] Birch in fact aims to account for a somewhat narrower notion: *genetic* control over an organism's behavioral *strategy*. He does so by restricting the cause side of the relevant causal relationship to genes, and notes that these do not directly control an interaction's outcomes, but rather the ways in which an organism responds to behavioral stimuli (i.e. it is a strategy, in roughly the game theoretic sense). Our discussion, however, aims for greater generality, so we assess Birch's proposal as it applies to control *tout court*.

control. Thus, a worker ant has control over her sterility to the extent that sterility exhibits systematic counterfactual dependence on her bodily and/or behavioral states, i.e. her influence on these outcomes exhibits substantial specificity.

Will this subtler notion of control do the trick? We think not. For, as with the simpler causal-control criterion, both sides of an interaction may have substantial and, in particular, fine-grained influence on it. Indeed, we think this isn't a superficial or coincidental aspect of the phenomenon. Since social interactions often result from, and are maintained by, co-evolutionary processes, we should expect each side to have relatively fine-grained influence. We should expect that, say, both the queen and workers-to-be will have evolved in incremental, interdependent ways; to the extent that one of them can exert fine-grained control over sterility, so will the other. More concretely, while we are not aware of any direct empirical findings as to Birch-style control, the mechanisms discussed above—pheromones, nutrition, "policing" of eggs etc.—are such that it is hard to see how there would be far greater control on one side of the interaction. For instance, in a biochemical reaction, typically both reagents (say a pheromone and receptor) can be tweaked to affect the rate, efficiency, side-products and other properties of the reaction. And in the course of a co-evolutionary process, such changes are to be expected. A situation in which only one side's influence evolves to be highly specific represents a fairly remote possibility.

In sum, the simple as well as the more sophisticated notions of causal control are both ways of fleshing out the idea that the actor is the individual who exerts primary influence over an interaction. While this sounds intuitive, further scrutiny reveals that such a criterion cannot distinguish biological altruism in the right way.

5.2. Actors, unity-of-purpose and inclusive fitness

The next and final suggestion we'll consider has a more abstract character – operating in terms of the overall characteristics of the organism in question. In particular, the idea is to understand actors as a kind of biological agent, where an agent is understood to be a system that *maximizes inclusive fitness (IF) in a unified manner*. We will explain this suggestion and its rationale. But we will ultimately argue that it too fails: The appeal to IF generates a regress, since IF itself rests on a prior notion of actorhood.

The appeal to unity-of-purpose, like the notion of internal control discussed above, has been central to philosophical discussions of action and agency. Indeed, on some philosophical accounts, unity or integration is the defining feature of being an agent (Bratman, 1987; Korsgaard, 1989; Hyman, 2015). The intuition, echoing ideas that may be familiar from the human context, is that for something to be an agent it needs to display internal coherence; its parts must "work together" towards a common end. At a minimum, there should not be a substantial degree of conflict among a system's activities – their ends must be compatible with one another. This is why we have difficulty assigning agency in cases of split personality, or even in cases where someone appears to be "pulled in different directions" (Sinnott-Armstrong & Behnke, 2000).

So the idea would be that an individual is an actor with respect to a given behavior if and only if they display agency with respect to that behavior. In turn, they would display agency if and only if the behavior coheres with the individual's unified overall purpose. The question that immediately arises is this: If such a criterion is to be applied in a *biological* context, how should we understand purpose, so as to judge whether there is unity-of-purpose? Okasha, following Hamilton and the kin selection tradition, asserts that in social interactions the answer is to be given in terms of IF: Organisms are unified-in-purpose to the extent that their various traits and behaviors contribute to an increase in IF.

Now, an organism's IF is comprised of the sum of two components: direct and indirect. The direct component refers to the organism's contribution to its own offspring; the indirect component relates to its contribution to the production of offspring by other organisms, weighted by their degree of genetic relatedness. Crucially, IF is regarded as the quantity that organisms are selected to increase, in any type of social interaction. Specifically, selfish behaviors increase IF via the *direct* component, while altruistic behaviors increase IF via the *indirect* component. Put differently, for the IF of selfish organisms, the direct component is positive and the indirect component is negative; for the IF of altruistic organisms, the reverse holds.

The basic idea behind this explanation for altruism is that while an organism may suffer in terms of its own survival and reproduction (i.e. its direct fitness) by performing an "altruistic" act, its IF will rise if the act bestows sufficient benefits on closely related individuals (i.e. on its indirect fitness). Thus, organisms' traits can be seen many times, and perhaps especially in cases of "altruism", as "working together" towards the end of greater IF.

Taking the example of worker ants, here is how the notion of unity-of-purpose might be invoked to save the concept of altruism. Sterility, along with many of its other physiological and morphological traits, raises the worker's IF by increasing its contribution to the reproductive output of its genetic relatives (despite the decrease to its own reproductive output). So the ant's traits can be said to "work together" to increase its IF. According to Okasha's view, since sterility belongs to a set of traits which together conduce to the IF of the workers, it supplies a 'rationale' for the behavior of the worker, grounding its status as an agent. In contrast, when an organism's traits are misaligned in terms of their contribution to IF, we have disunity-of-purpose, and so cannot ascribe agency to the organism in question. Such is the case, for instance, in parasitic manipulation: The katydid's jumping into the water *decreases* its IF and therefore it is *dis*unified-in-purpose with the katydid's other traits. Consequently, the jumping cannot be seen as a means to increasing the katydid's IF. If anything, the jumping coheres with the *parasite's* goal, inasmuch as it increases *its* IF. Thus, we cannot treat the katydid as an agent with a unified purposeful activity, but we may treat the hairworm as-if it were a purposeful manipulator.

As these examples illustrate, behaviors classified as altruistic are associated with unity-of-purpose, hence agency, since they lead to an overall increase in IF. In the worker ant, a host of adaptive features—sterility, life history, various behavioral traits etc. —are aligned, jointly contributing to its IF. In contrast, in parasitic interactions unity—and therefore agency—break down: Promoting the parasite's survival and reproduction runs counter to many of the host's other traits. Thus, it appears that viewing control/actorhood in terms of unity-of-purpose may well allow us to distinguish altruism from antagonistic interactions like parasitism.

But this appearance is misleading – unity-of-purpose does not supply a sound basis for distinguishing altruistic actors from passive recipients, because unity-of-purpose, it turns out, *relies on that very distinction*, generating a regress. To see this, let us look more closely at how IF is defined. Okasha quotes Hamilton's original definition: "[T]he personal fitness which an individual actually expresses … once it is stripped of all components which can be considered as due to the individual's social environment … then augmented by certain fractions of the quantities of harm and benefit which the individual himself causes to the fitnesses of his neighbours … The fractions in question are simply the coefficients of relationship." He then adds: "This definition sounds complicated but the underlying idea is simple, namely to re-assign all fitness components to the actors that cause them." (2018, 119-120). Thus, to calculate an organism's IF one must identify, for each fitness contribution, which organism is "the actor that causes it". This, we suggest, lands us back where we started, i.e. in need of an actor/recipient distinction; it renders the appeal to IF and unity-of-purpose unhelpful for the problems we raise.

To clarify, suppose we are looking at two organisms – a hungry vampire bat, which failed to gain a blood meal on its own; and a satiated next mate, which, having hunted successfully, shares food with the hungry individual. Suppose, for the sake of this discussion, that the successful hunter suffers a decrease in reproductive output while the reproductive output of the hungry bat is increased. Here, if we assume that the successful hunter is the actor controlling these effects and, further, that they are selected for by kin selection, then we may regard the successful hunter as an altruist. If, for some reason, we wished to regard the other (hungry) individual as the actor, i.e. as in control of the fitness effects, then we should regard the interaction as selfish (i.e. the hungry bat would be exploiting its nest mate). What we cannot do, on pain of double counting, is to assign the effects of the behavior to both of them. The crux for our purposes is that the decision as to which individual to attribute the effect to rests on a (logically) prior decision as to which individual is in control of this effect. Or, in other words, on which side is the actor. This decision will, of course, determine whether the interaction is altruistic or selfish. Recall the example of sterility of workers in social insects. In order to count as a contribution to the workers' IF, causal responsibility for sterility should reside in the workers themselves. Yet, as we saw, it is hard to make a case for that. Sterility is, to a large extent, induced by the queen and/or by other workers. It follows that we cannot incorporate sterility into the worker's IF – it is not something the worker is causally responsible for. But if sterility cannot be weighted into a worker's IF, we cannot classify it as altruistic.[7]

5.2.1. *Linear regression to the rescue?*

Here we must look at a relatively technical, but important potential rejoinder.[8] It has been implied that the issue we just raised, concerning the need to independently specify actor and recipient, is obviated when IF is calculated via a linear regression equation. For example, in describing the regression method for calculating IF, Gardner, West & Wild (2011) say that it can be used to "reassign the indirect fitness effects to the actors responsible for them" (ibid, 1024). If this is right, then applying regression analysis circumvents the need to presuppose who the actors for a given effect are; the regression analysis *tells us* who they are.

However, the idea that the regression method *in and of itself* yields information about the causal structure of social interactions is unconvincing. The key point is that the output of a regression analysis is a set of correlations. And, famously, correlations always require further assumptions if they are to be given a causal interpretation – which is necessary to establish an actor/recipient distinction. To bring this out, we rely on recent work by Benjamin Allen, Martin Nowak and E.O. Wilson (2013). The concerns motivating Allen et al. are related but somewhat different from ours.[9] However, we can adapt their critique to our ends. We do so

[7] Grafen (2006) provides a related defense for agential thinking in evolution. He demonstrates that, under certain assumptions, organisms are expected to behave as-if trying to maximize their own IF. This work has been taken as a formal justification for the use of intentional terms like 'altruism' (West et al., 2011). However, as Grafen's analysis critically relies on the concept of IF, and thus presupposes a notion of actor's control, it leads to the same regress we have seen.

[8] We are grateful to a reviewer for this journal for pressing the matter and pinpointing its exact relevance.

[9] Allen et al. (2013) aim to undermine certain claims made about the explanatory power of the regression approach to inclusive fitness theory; our concern is the idea of altruism. We should also

while avoiding technical details – for those interested, Marshall (2015) provides a helpful introduction.

The regression approach effectively re-frames the concept of IF in terms of partial regression coefficients (Birch & Okasha, 2015). This means that the causal definition of IF seen earlier—"re-assigning all fitness components to the actors that cause them" —is in effect replaced by a correlational definition.

Recall that IF has two components: a direct component, accounting for the organism's contribution to its own reproduction, and an indirect component, accounting for its contribution to the reproduction of others. This definition is causal: an organism's IF only takes into account effects for which it is causally responsible. In comparison, in the regression-version of IF, the direct component becomes the partial regression of an organism's genotype on *its own* fitness (controlling for the genotype of its social partners); the indirect component becomes the partial regression of an organism's own genotype on the fitness of *its social partners* (again, controlling for their genotype), weighted by relatedness.

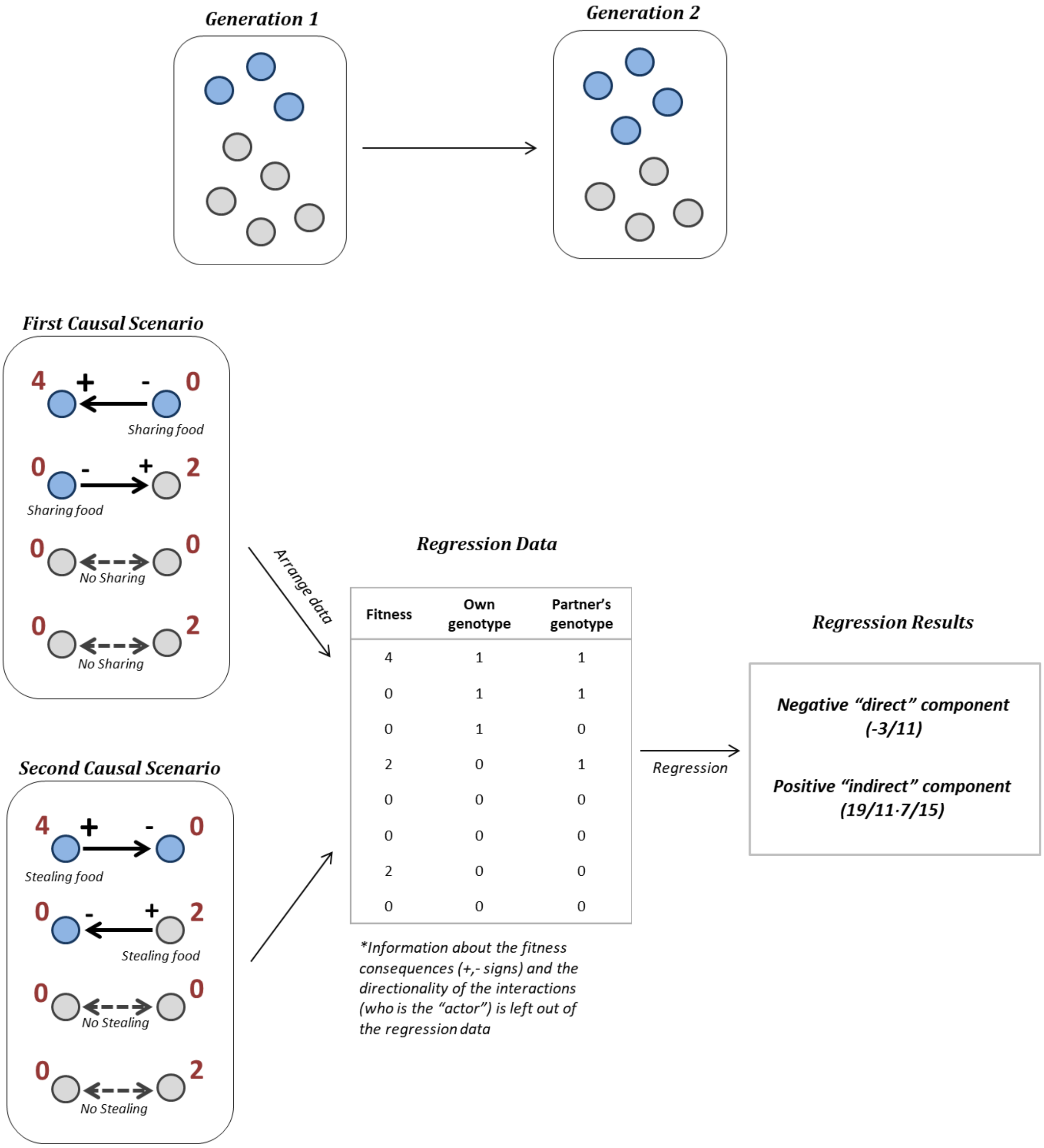
The results of a regression analysis, which define IF in the correlational sense, are sometimes interpreted in the causal sense. For example, according to such an interpretation, if the "direct" component is statistically negative, that is, if there is a negative partial regression between an organism's genotype and its own fitness – this is taken to imply that this genotype causally reduces the fitness of those who express it. However, as Allen et al. (2013) show, such implications depend upon prior assumptions about the underlying causal scenario. In the present context, this means that inferring which organism is the actor and which is the recipient is a matter of extra-statistical interpretation. To show this, we adapt one of their examples, showing that the same regression results are compatible with (a) a scenario in which "altruism" (in the causal sense) is present and (b) a scenario involving "exploitation".

Assume a simple case with a locus involving two alleles – *A* (marked in blue in *Figure 1*) and *B* (marked in grey). The frequency of *A* is increasing from one generation to the next (from 3/8 to 4/8, *Figure 1, upper left panel*). In order to quantify the contributions of the "direct" and "indirect" components to this increase in *A*, we conduct a regression analysis. We have all the necessary data: We know the genotype of each individual, the fitness of each individual, and we also know who interacted with whom (*Figure 1, middle panel*). Notably, the data themselves do *not* include any information about the causal structure of the interactions. The

note that the example we adapt is used by Allen et al. to show that regression analysis approach to IF does not yield causal information, in general. We use it to make a more specific point pertaining to altruism.

results of the regression show that the “direct” component is negative, while the “indirect” component is positive (*Figure 1, rightmost panel*). This means that carrying allele *A* is statistically associated with lower personal fitness, and with having partners with higher fitness. If the results of the regression are taken to correspond to a causal definition of IF, this behavior is classified as altruism; hence, this methodological procedure supposedly allows the identification of altruism without presupposing who is the actor and who is the recipient.

SEE NEXT PAGE FOR FIGURE 1



Fitness	Own genotype	Partner's genotype
4	1	1
0	1	1
0	1	0
2	0	1
0	0	0
0	0	0
2	0	0
0	0	0

**Information about the fitness consequences (+,- signs) and the directionality of the interactions (who is the "actor") is left out of the regression data*

Figure 1. Different causal scenarios may lead to the same regression results. The frequency of the blue allele increases from the one generation to the next (top panel). We construct two hypothetical causal scenarios that underlie this increase (bottom left panel): [+] and [-] signs indicate whether the organism gained or lost fitness; arrows indicate social interaction; red numbers indicate the total number of offspring. In the first scenario there is food "sharing", while in the second scenario there is food "stealing" (see text for details). The data needed for a regression is extracted and arranged in a table, leaving out any information about the causal structure of the interactions (bottom middle panel). The results of the regression show that the "direct" component associated with the increase of the blue allele is negative, while the "indirect" component is positive (bottom right panel). If the results of the regression are taken to correspond to a causal definition of IF, this behavior is classified as altruism. Such an interpretation suits the first scenario, but not the second. Adapted from Allen et al. (2013).

For our purposes, we now further assume the overall fitness consequences of each interaction are also given (marked in [+] and [-] signs): We know whether each organism gained or lost fitness due to the interaction. All the social interactions involving *A* are [+,-] interactions. To be clear, the pattern of fitness consequences is *not* a part of the input of the regression. The data used in the regression omit such causal information. This is why the regression results are compatible with different causal scenarios, as exemplified next (see also *Figure 1*).

Let us imagine an environment in which food is scarce. Alleles *A* and *B* affect behaviors relating to food, and thus have an important effect on fitness.

First Causal Scenario: "Altruism". The individuals expressing allele *A* forage alone, but they subsequently seek out other individuals, and share with them the food they found. If they meet other *A*-individuals, they share more food. Individuals expressing allele *B* also forage alone, but in contrast to *A*-individuals, they save all the food they find to themselves. Under this scenario, the negative "direct" component found in the regression results from *A*-individuals losing fitness by "donating" food; the "indirect" component is positive because the "donation" provides excess-benefits to other carriers of the allele. In this case, individuals expressing *A*-allele drive the relevant fitness outcomes, and interpreting *A* as an allele for "altruism" suits the causal structure of the interaction.

Second Causal Scenario: "*Exploitation*". There is no food sharing, but rather theft. Everyone tries to steal food from their social partners. But, in each interaction which involves stealing, only one side is successful. Individuals expressing allele *A* are better thieves, but those expressing allele *B* are excellent at defending their food. Thus, *B*-Individuals are never the victims of theft. As *A*-individuals are poor defenders, the organisms they encounter might steal their food. Notably, this time, the negative "direct" component found in the regression results from the fact that *A*-individuals are the target of "exploitation"; counter-intuitively, the excess benefits of being better at stealing (from other carriers of the allele) is captured in the positive "indirect" component. In this case, interpreting *A* as an allele for "altruism" seems less suitable than interpreting it as intra-specific parasitism.

Thus, we see that different causal scenarios are compatible with the same regression coefficients. As demonstrated, the fact that a spreading allele is statistically associated with having lower personal fitness and higher partners' fitness does not, as such, entail that the allele induces "altruistic" behavior; for instance, it is also possible that the allele induces parasitizing. The upshot is that the regression approach to IF does not obviate the need to decide, independently of the regression, which side is the actor and which is the recipient. What is more, in our simplistic scenarios, it might seem like there is an answer to the question “who is the actor”, but using the regression method may get it wrong. We should stress again that we think that in many actual cases, there is no “right” answer. In such cases, trying to “re-assign fitness effects to the actors which caused them” may be impossible in practice.

All told, we have seen that classifying an interaction as altruistic presupposes that we can tell which of the interacting organisms is the actor. We are deeply suspicious of this notion, and we think we’ve exhibited why: It is very hard to find a principled way of apportioning causal responsibility such that one side of an interaction comes out as “in control”. The suggestions we have looked at—and we are not aware of further alternatives—are both conceptually problematic and do not match the empirical realities underlying plus-minus interactions.

6. Concluding Remarks

We have argued that the distinction between altruism and other plus-minus interactions presupposes an intentionalist conception of animal behavior, standard definitions notwithstanding. This makes it difficult to tell apart altruism from accidental transfer of fitness and from seeming mistakes. A more acute problem is that absent a way to distinguish an “actor” in a social interaction from a passive “recipient”, we cannot differentiate between altruism and selected-for antagonistic interactions, such as parasitic manipulation.

Attempts to define a relevant notion of agency face substantial conceptual problems and fail to align with the mechanistic facts regarding causal responsibility. It is also worth noting that the cases we have discussed are not borderline or marginal cases by any means. Indeed, they are absolutely central to the study of altruism. Moreover, similar problems are to be expected in many other cases, because the idea of one-sided control of a social interaction is implausible to begin with, given the complex nature of these interactions and the fact that they result from subtle co-evolutionary dynamics.

If our arguments are correct, there appear to be three paths forward. First, it is possible that a clear and general criterion for distinguishing actors from recipients will be found. We are not optimistic in this regard, but we cannot, of course, rule out such progress in advance. Interestingly, Wyatt, West, and Gardner (2013) have recently suggested that the actor/recipient distinction is, to an extent, subjective – although their considerations differ from ours. These authors posit that, at least in some contexts, a modeler may have a choice between allowing individuals of different species to be the “actor” and the “recipient” with respect to each other, and between restricting actor/recipient relations to members of the same species. Such a choice, in turn, determines whether or not a behavior is seen as inter-species (versus intra-species) altruism within an inclusive fitness framework.

A second path is to eliminate the category of altruism altogether. In some ways, this is the most natural response to our argument. But there are definite costs associated with such a move: Altruism has generated a lot of important and exciting research. What shall we do with this body of work if we cease to recognize altruism as a bona fide biological kind?

A third response is to view altruism and research into it in a less objective light. Perhaps treating a social interaction as altruistic carries heuristic value? Perhaps it is a sound employment of the intentional stance (Dennett, 1989; Kornblith, 2002)? We are open to heuristic uses and to related projectivist, as it were, construals of the notion of altruism. Nor does our argument imply that all plus-minus interactions should be handled similarly. That said, it is unobvious why thinking about an interaction between, say, a worker ant and its queen, as altruistic is of greater heuristic value than thinking of it as a case of parasitism (where the queen exploits the worker). While we cannot enter into an extended discussion of this issue here, we surmise that both altruism and selfishness can be utilized in a productive heuristic manner, depending on the context.

Whatever one makes of these options, we hope to have shown that the notion of altruism suffers from a foundational conundrum, threatening its cogency and continued use. We think more work can and should be done to clarify these issues.