The Ways of Altruism[1]

Gualtiero Piccinini, 1 University Blvd, Department of Philosophy, University of Missouri – St. Louis, St. Louis, MO 6121, piccininig@umsl.edu, http://www.umsl.edu/~piccininig/

Armin Schulz, 3101 Wescoe Hall, Department of Philosophy, University of Kansas, awschulz@ku.edu, http://people.ku.edu/~a382s825/

Abstract. We argue that some organisms are altruistically motivated and such altruistic motivation is adaptive. We lay out the *helper's decision problem—determining whether to help another organism.* We point out that there are more ways of solving this problem than most people recognize. Specifically, we distinguish two kinds of altruistic motivations, depending on whether a desire to help is produced for one's own sake or for others' sake. We identify circumstances in which either kind of psychological altruism provides the most adaptive solution to the helper's decision problem. As a result, we show that both kinds of psychological altruism are likely to be instantiated and selected for.

Keywords: egoism, evolution, evolutionary altruism, mechanisms, psychological altruism.

1. Introduction

Some organisms behave in ways that increase the direct (reproductive) fitness of another organism (West et al., 2007). In what follows, we call organisms that behave in this way—regardless of their motive—"helpers." In some species, dispositions to help have been selected for (e.g., Sober & Wilson, 1998; Gardner & West, 2010; West et al., 2007, 2011; Okasha, 2006). This could be because increasing the (direct) fitness of other organisms also increases the (direct) fitness of the helper, as in cases of mutual benefit (West et al. 2007; Sober & Wilson, 1998). There can also be selection for helping where the helping decreases the direct fitness of the helper, as in cases of strict evolutionary biological altruism (West et al., 2007; Okasha, 2006).

Evolutionary biological altruism can be explained in terms of inclusive fitness, which is the sum of an organism's direct fitness (i.e., the expected number of offspring minus the portion of offspring due to the help it receives from others in its population) and its weighted contribution to the direct fitness of every other organism in the population, where the weights are given by the coefficient of relationship between the organisms (Hamilton, 1964; Gardner et al., 2011; Rubin, 2018). There is good reason to

[1] Thanks to Valdenor Brito, Sarah Brosnan, Christine Clavien, Corey Maley, Sarah Robins, Steve Stich, Isaac Wiegman, Felix Warneken, anonymous referees, and our audiences at SPP 2018 and PSA 2016 for helpful comments on previous versions. This material is partially based upon work supported by the National Science Foundation under grant no. SES-1654982 to Gualtiero Piccinini.

think that what natural selection depends on is inclusive fitness, not direct fitness or personal fitness (an organism's expected number of offspring) (Grafen, 2006). This matters, as the inclusive fitness of biologically altruistic behaviors can be positive: the behavior's positive contribution to the personal reproductive success of related organisms can outweigh its negative contribution to the organism's own personal reproductive success.

It remains controversial how (selected for) helping behavior, whether strictly evolutionarily altruistic or not, is *motivated*. In particular, it is not yet clear whether and when helping behavior is motivated in a way that deserves to be called *psychologically altruistic*—rather than selfish or merely reflexive. Equally controversial is whether and when *psychological altruism* is adaptive—i.e., whether it contributes to the inclusive fitness of the helper and, therefore, is subject to evolution and preservation by natural selection. Also, whether psychological altruism is adaptive leaves open whether it is strictly evolutionary biologically altruistic: depending on how it increases the bearer's inclusive fitness, it could also be mutually beneficial or evolutionarily selfish. This will become important below.[2]

Psychological altruism has been heavily debated by ethicists, cognitive neuroscientists, social psychologists, economists, and evolutionary biologists. Some classical and contemporary theories deny that organisms ever have ultimate altruistic desires (Hobbes 1651; La Rouchefoucauld 1665/2007; Bentham 1824: 392-3; Nietzsche 1881/1997: 148; Cialdini et al., 1997; Grant 1997; Miller 1999). Others maintain that some organisms are psychological altruists (e.g., Batson, 1991; Fehr & Camerer, 2007; Stich et al., 2010) and that psychological altruism can be selected for (Sober & Wilson, 1998; Schulz, 2011, 2016, 2018; Clavien, 2011). Some contemporaries argue that identifying psychological mechanisms behind altruism is empirically too difficult, so that we should leave these proximate mechanisms aside and focus on altruistic behavior alone (Wilson 2015, Chap. 5). In summary, there is no consensus on whether psychological altruism occurs and, if it does, whether it is selected for (Stich et al., 2010; Garson, 2016).

In this paper, we propose an enhanced evolutionary framework for investigating psychological altruism. Our framework encompasses both humans and non-human species (cf. Bshary & Raihani, 2017). We argue that some organisms are altruistically motivated and such altruistic motivation is adaptive. In Section 2, we lay out the helper's decision problem—determining whether to help another organism—and point out that it can be extremely difficult to solve. Clarifying the tradeoffs involved in the helper's decision problem allows us to articulate a space of possible strategies for solving it. In Section 3, we note that there are more options than most people recognize. Specifically, we identify four kinds of strategy for solving the helper's decision problem and place them in a matrix based on whether the motivations for a behavior have egoistic or altruistic content and whether they are produced egoistically or altruistically. In Section 4, we argue that psychologically altruistic solutions are likely to be selected for. One reason is that, in some cases, egoistic strategies are impractical for lack of sufficient information and computational resources. Therefore, organisms must resort to altruistic strategies at least some of

[2] Analogous points can also be made within other theoretical frameworks, such as involving neighborhood-modulated fitness or multi-level selection theory.

the time. In Section 5, we bring out some consequences of this discussion for ethics, cognitive neuroscience, and economics.

2. The Helper's Decision Problem

It is widely recognized that organisms must decide when to help and when not. In some circumstances, it might be adaptive for some organisms to help all the time—or to never do so. More commonly, organisms need to determine when helping is biologically called for. We shall call it the helper's decision problem. Four points are important to note.

First, solving this problem does not require that the organism have a concept of adaptiveness or inclusive fitness, nor that it chooses behaviors by calculating their expected (inclusive, direct, or personal) fitness. For one thing, (selected for) helping behavior may evolve and be determined by processes that are not even cognitive, let alone explicitly representational (cf. Strassman and Queller, 2011).[3]

Second, what *is* required is that organisms have some proximate mechanism for choosing helping behaviors in a way that correlates at least reasonably well with (at least) their direct fitness.[4] Organisms cannot use a mechanism that *systematically* picks out maladaptive behaviors. If they did, they would eventually go extinct. In fact, there is empirical evidence that many organisms choose helping behaviors in ways that are by and large adaptive for them (Houston & McNamara, 1999; Jensen, 2012; Chudek et al. 2013, 436-437).[5] Beyond this, no further assumptions about these proximate mechanisms need to be made.

Third, the helper's decision problem is very difficult to solve in the general case. There are a large number of variables that influence whether helping another organism is selected for (e.g., Frank, 1998; West et al., 2007, 2011; Queller, 1985, 1992; Okasha, 2006; Birch & Okasha, 2014; Skyrms, 1996, 2004; Sober & Wilson, 1998; Stevens and Hauser 2004). The adaptiveness of helping depends on how closely

[3] We will remain neutral on the vexed issues of what counts as a representation and how it gets its content. Any reasonable account will do. For an opinionated defense of representational explanation within cognitive neuroscience that doubles as a defense of the notion of neurocognitive mechanism we adopt here, see Boone and Piccinini (2016) and Thomson and Piccinini (2018).

[4] As noted earlier, an organism's direct fitness is the expected number of offspring it has just by itself. The organism's personal fitness is simply its expected number of offspring (this will be greater than its direct fitness if some of its offspring result at least partly from the actions of other organisms). Finally, an organism's inclusive fitness is the relatedness-weighted sum of the organism's own direct fitness and those of the other organisms it is related to (see also Birch, 2017). Note also that the points in the text could be made in terms of other notions of fitness as well (such as neighborhood-modulated fitness or various multi-level notions of fitness—Wilson 2015, 28; Okasha, 2006; Sober & Wilson, 1998, Gardner et al., 2011).

[5] In cases where the relevant organisms are cultural learners, what determines whether a behavior will spread through the population—i.e., whether it is adaptive in the broadest sense—can depend on more than its biological adaptiveness. Gächter et al. (2010) find that culture influences cooperation. For more on this, see e.g., Boyd & Richerson (2005), Richerson et al. (2016), Stich (2016). See also below.

the recipient is biologically related to the agent, the likelihood that the recipient will reciprocate (which in turn may depend on whether there are mechanisms for ensuring reciprocation), the existence of mechanisms for punishing non-helpers and rewarding helpers, and other factors. The value of each of these variables can be onerous to calculate, and combining these values optimally can be computationally highly complex. Because of this, the most that an organism can hope for is a good-enough solution most of the time.

Fourth, given the difficulty of the helper's decision problem, evolution must select shortcuts—heuristics that provide good-enough solutions most of the time (Gigerenzer et al., 1999; Hutchinson and Gigerenzer, 2005). In turn, this implies that different situations are likely to call for different heuristics. Some organisms, such as social insects, have relatively automatic and rigid ways of choosing behaviors, including altruistic ones. Other organisms may sort behaviors in terms of which ones are worth performing and which ones are not—e.g., they might assign different behaviors utility or reward values and then act accordingly (for more on this, see e.g. Morillo, 1990; Schroeder, 2004; Glimcher et al., 2005). Even cognitively sophisticated organisms may sometimes rely on relatively simple, innately[6] hardwired strategies because they are sufficient (e.g., "help offspring in need"—see Schulz, 2016, 2018).

The helper's decision problem and its varied heuristic solutions raise two additional questions. Consider all the different possible ways of determining whether an organism should help another. Which ones deserve to be called psychologically *altruistic*, which *egoistic*, and which deserve neither label? We will answer this question in the next section. The subsequent section will examine which ways of altruism are likely to be selected for.

3. A Matrix of Helping Strategies: Egoistic, Altruistic, and Impersonal

Just because it is adaptive for an organism to help another, it doesn't follow that it is adaptive for its *motives* to be altruistic. Someone can (adaptively) act altruistically for entirely selfish reasons. Our goal in this section is to distinguish more precisely between different motivations for helping.

Drawing these distinctions in a satisfactory way is not easy. The traditional way is roughly in terms of the content of the organism's *ultimate* desires—desires that are *not* instrumentally derived from other desires. To the extent that an organism is motivated by ultimate other-involving desires—i.e., desires directed at increasing others' wellbeing (or happiness, or the like), and thus, by assumption, their fitness—it is deemed a psychological altruist. By contrast, to the extent that an organism is motivated by ultimate self-involving desires—i.e., desires directed at increasing one's own wellbeing—it is deemed a psychological egoist. This implies that, to the extent that an organism is motivated by ultimate *neutral* desires—i.e., desires directed neither at the self nor at another organism—it should be considered

[6] The notion of innateness is controversial. For an account that suits our purposes, see Northcott and Piccinini (2018).

neither an altruist nor an egoist (see also Sober and Wilson, 1998, Stich et al., 2010, Garson, 2015; Schulz, 2016).

This characterization is too simplistic. The main problem is that it neglects how desires are produced. Desires are produced by cognitive mechanisms such as innate dispositions, learning, and instrumental reasoning. Such mechanisms may deserve to be called altruistic or egoistic. To make progress on psychological altruism, this question of production must be addressed.

To begin, we need greater precision on what counts as a desire in the relevant sense. We cannot limit ourselves to paradigmatic desires, namely, propositional representations of a state of affairs that cognizers explicitly deliberate with. On one hand, if we required that psychological altruists be motivated by explicit deliberation that employs propositional desires, we would rule out too much. Someone who acts on an immediate urge *to help another organism*—i.e., a representation of the kind "must help so and so"—without representing the precise state of affairs aimed at and without deliberating, still deserves to be called a psychological altruist. On the other hand, we shouldn't be overly inclusive. For example, an organism that helps another because of a mere reflex—e.g., a cow that gives us milk because its milk ejection reflex is triggered—should *no*t count as a psychological altruist. There is no good reason to classify automatic responses such as reflexes as *psychologically* egoistic or altruistic. Of course, this is not to say that automatic responses to help others are biologically or even morally unimportant—the point is just that they are different in nature from either psychologically altruistic or egoistic helping behaviors.

Therefore, for present purposes we will count as a *desire* any conative state that represents which goals or states of affairs to pursue, without requiring that desires be propositional attitudes that can be deliberately reasoned with (cf. Clavien, 2011). We follow mainstream cognitive neuroscience and assume that not only human beings but also many other organisms are motivated by desires in this broad sense. Desires contrast with reflexes, fixed action patters, motor commands, and other automatic control mechanisms—what we shall call *automatisms* in what follows. While the latter can also trigger behaviors, they do so without representing goals or states of affairs.

When it comes to helping behaviors driven by desires, we can distinguish between behaviors caused by egoistic desires and those caused by (ultimate) altruistic desires. This traditional distinction should be refined. A more fine-grained view of the nature of representational decision-making situates the traditional distinction within a matrix spanned by two dimensions. On one hand, helping behaviors stem from desires with different kinds of *contents*: some are driven by a concern for other organisms (altruistic ones), some by concern for oneself only (egoistic ones), and some by concerns that are neither other-involving nor self-involving (e.g., wanting to play). We will continue to refer to desires with altruistic content as altruistic desires, and desires with egoistic content as egoistic desires. (This dimension is familiar from most of the recent discussion of this topic: e.g., Sober & Wilson, 1998; Stich et al., 2010; Batson, 1991.) On the other hand, helping behaviors stem from desires that are generated in different ways. Here we are not talking solely about ultimate desires. The distinction is between sources of desires regardless of whether the desires are ultimate or instrumental. In particular, desires (whether altruistic, egoistic, or neutral in content) can be *produced* by egoistic, altruistic, or neutral

cognitive mechanisms. This makes for the second dimension of the matrix of helping strategies: it concerns the cognitive mechanisms that *produce* the desire in question.

This brings us to what we mean by egoistic, altruistic, or neutral mechanisms of desire production. By *egoistically produced* desires, we mean desires produced by evolutionarily selfish mechanisms: mechanisms that were selected for increasing their bearer's own reproductive success (i.e., its direct or personal fitness) only, as opposed to that of other organisms (see also West et al., 2007).[7] There are at least three kinds of evolutionarily selfish desire-production mechanisms. First, an organism can have an *innate* disposition to form desires to pursue one's self-interest under certain circumstances. That is to say, the organism's brain is built so that detecting certain circumstances (e.g., low blood glucose levels) triggers the formation of a desire to pursue one's self-interest (e.g., the urge to eat). Second, an organism can *learn* to produce desires. There are a number of different such learning mechanisms available (for an overview, see e.g. Henrich & McElreath, 2007; Boyd & Richerson, 2005; Sterelny, 2012), at least some of which very plausibly evolved for evolutionarily selfish reasons. For example, consider the class of reward-based learning dispositions: the mechanism behind these dispositions—the tendency to seek rewards and avoid aversive stimuli—is evolutionarily ancient and can be found in solitary organisms as cognitively simple as the sea slug *Aplysia californica* (Kandel 2001). There is no reason to think that these reward-based learning mechanisms could not extend to the acquisition of desires. The third example of egoistically produced desires is cases where organisms *internalize* external norms: in cases where certain behaviors are socially rewarded, organisms that can and do internalize the relevant norms might save processing and error costs, and thus benefit themselves (see Gintis, 2003, for a model of altruism based on this).[8]

By contrast, by *altruistically produced* desires, we mean desires produced by mechanisms that are evolutionarily altruistic: mechanisms that were selected, at least partially, for increasing other organisms' reproductive success (i.e., their direct fitness). In other words, evolutionarily altruistic mechanisms increase the expected reproductive success of other organisms (and may or may not increase the organism's own reproductive success).[9] The main instance of this is any selected-for innate disposition to form desires to help others under certain circumstances. That is to say, the organism's brain may have evolved so that detecting certain circumstances (e.g., that a baby is crying) triggers a desire to help others (e.g., the urge to soothe the baby), possibly mediated by an intermediate internal

[7] Put differently, egoistically produced desires are desires produced by mechanisms whose evolution is driven just by the fact that they increase the direct reproductive fitness (i.e., the number of offspring) of the organism in question, while not increasing its indirect fitness (so that the second component of the inclusive fitness calculation here is zero).

[8] This internalization mechanism could also be seen as a form of reward-based learning. However, this would not alter the main point in the text: namely, that desires can be egoistically produced.

[9] We may distinguish further between two kinds of altruistically produced desires. *Strictly* altruistically produced desires are produced by mechanisms solely selected for increasing others' direct fitness—they increase the expected reproductive success of that other organism and do not affect or even decrease the bearer's expected reproductive success. *Broadly* altruistically produced desires are produced by mechanisms selected for increasing others' direct fitness along with their bearer's (what West el., 2007, call "mutual benefit"). This distinction will not play a role in this paper.

state (e.g., empathy[10]) under appropriate background conditions (e.g., bonding between agent and target and sufficient resources). This kind of innate disposition mechanism may be subject to developmental constraints, so that the precise range of others and circumstances that trigger a desire to help depend on environmental factors—e.g., on whom an organism is raised with. For instance, empathy is highly plastic: its intensity is modulated by factors such as whether its target is a member of the group and whether they behaved fairly in the past; as a result, empathy predicts altruistic behavior under certain circumstances (compassion) but not others (empathic distress) (Klimecki, 2015).

Finally, a desire is *neutrally produced* just in case it is neither egoistically nor altruistically produced. It may result from mechanisms that have not been selected for at all, or it may be a by-product of other mechanisms. For example, a desire may be generated in a manner similar to how skills are acquired: for example, after repeatedly helping B, organism A may come to form a desire to help B, much like the skill of riding a bike comes after an organism has practiced this for a while. We are not committed to this being a frequent occurrence, or even for it to be possible at all—we just want to note that our account has room for it if this turns out to be a plausible empirical possibility.

There will also be mixed cases. Most obviously, an altruistic desire may be first produced by an innate mechanism and then reinforced by conditioning. In what follows, we will focus primarily on pure etiologies, but we should not forget that mixed etiologies are possible.

We can now combine the two dimensions—desire content and desire production—to obtain a space of psychological strategies for solving the helper's decision problem. This space includes at least four types of motivation to help others.[11] We label them as follows.

> *Psychological egoism* chooses actions based on desires with egoistic contents.

Psychological egoists need not engage in helping behavior, but they can. If they do, they determine the appropriateness of helping from instrumental reasoning starting from egoistic first principles.

> *Classical psychological altruism* chooses actions based on non-egoistically produced ultimate desires with altruistic contents.[12]

---

[10] De Waal (2008) argues that empathy can play such a role; Klimecki et al. (2016) provide additional supporting evidence. Someone might object that empathy-driven altruistic behavior is *selfishly* motivated, because it improves the agent's emotional state. This objection is confused. It is well established that empathy can lead to either altruistic or selfish desires and, consequently, to either altruistic or selfish behaviors (Schulz, 2017). Here we are considering cases in which empathy leads to ultimate altruistic desires. In such cases, empathy deserves to be considered a component in an *altruistic* source of desires. Any improvement in the agent's emotional state, which may or may not follow the desire's satisfaction, is not the agent's motive—it's just a by-product.

[11] Böckler et al. (2016), Clavien and Chapuisat (2013), and Ramsey (2016) also distinguish different types of altruisms, but they are addressing different questions so their taxonomies are different from ours.

[12] Classical psychological altruism admits of two variants: a *pure* variant, where the ultimate desires with altruistic content are all produced altruistically (whether strictly or broadly), and an impure variant, where the ultimate desires with altruistic content are produced either altruistically or neutrally. We will not consider this subdivision further.

Although we are calling this motivational structure classical psychological altruism, this is not exactly the same as psychological altruism as classically conceived—indeed, it is not entirely clear how psychological altruism *is* classically conceived. For traditional theorists of psychological altruism say nothing at all about whether the ultimate desires are produced altruistically, egoistically, or neutrally. Still, the spirit of psychological altruism as classically conceived is that it's a motivational structure untainted by egoism. Therefore, given our framework, what we call classical psychological altruism is the motivational structure that is closest to the spirit of psychological altruism as classically conceived. Classical psychological altruists need not succeed in helping others, but they aim to help. They have one main source of ultimate altruistic desires: (more or less plastic) innate dispositions.[13]

Apart from these classical motivational structures, there is also a nonclassical variant of altruism:

> *Nonclassical psychological altruism* chooses actions based on egoistically produced ultimate desires with altruistic contents.

Nonclassical psychological altruists need not succeed in helping others, but they aim to help. The main egoistic source of desires with altruistic contents we focus on here is reward-based learning. This kind of learning leads to the reoccurrence of a previously occurring desire by either rewarding its presence or punishing its absence (see also Rachlin, 2002; Erev and Roth, 2014). In cases where this learning operates by rewarding a previously occurring desire with altruistic content, the chain of rewards must end in a desire that is produced by some other means. Thus nonclassical psychological altruists must possess some other source of desires as well.

Finally, there is a motivational structure that is based on desires with contents that are neither altruistic nor egoistic:

> *Impersonal agency* chooses actions based on desires with neutral (i.e., non-egoistic and non-altruistic) contents.

Impersonal agents neither aim to help themselves nor others, but they may help others nonetheless. For example, someone might volunteer to fight a community's enemy to pursue adventure; nevertheless, their action may protect their community. This impersonal helping is most relevant to highly social and cognitively sophisticated organisms subject to cultural pressures, such as human beings. Presumably it piggybacks on other forms of altruism and molds them in light of social pressures. It's not directly relevant to our discussion so we set it aside.

These four motivational structures are summarized in table 1.

[13] As noted earlier, they may have a second source in the form of desires formed by habit. As also noted earlier, we will not consider this further here.

Content of the Desire (⇓)	Origin of the Desire (⇒)	Altruistic	Egoistic
Altruistic		Classical altruism	Non-classical altruism
Neutral		Impersonal agency	
Egoistic		Egoism	

Table 1. Four key motivational structures for helping others.

This taxonomy has some important implications: in particular, we can now see clearly a major deficiency of the traditional definitions of psychological altruism and egoism: these definitions oversimplify the situation and make it appear that there are only two options, when in fact there are (at least) four.[14] This is important, for it overlooks some theoretically and empirically important options that need to be taken into account to develop a proper understanding of the psychological structures that might underlie helping behaviors.

4. The Evolutionary Biology of Psychological Altruism

Now that the theoretical landscape is clearer, we can explore when these ways of making helping decisions are selected for. Specifically, we want to assess the evolutionary pressures that plausibly underlie the different ways of being motivated to help. To prepare the terrain, let's briefly consider the nature of this evolutionary biological methodology.

The empirical plausibility of the different ways of making helping decisions is a (comparative) psychological question. Thus, asking about the evolutionary pressures on these different ways of making helping decisions is not obviously the most straightforward way of proceeding—though hardly unorthodox either (e.g., Sober & Wilson, 1998; Stich et al., 2010; Schulz, 2011, 2015; Garson, 2014; see also Barrett, 2015; Barrett et al., 2002; Buss, 2014; Barkow et al., 1992; Wilson, 2015).

One reason it is particularly valuable here, though, is that other sources of data are not yet available: as the previous section made clear, the questions of psychological altruism has so far been investigated

[14] Schulz (2016, 2018) and Garson (2016) also point out that there are more than two options—though they do not expand the space of helping motivations in the way that we do here.

within an impoverished theoretical framework, so that existing work is unable to discriminate between classical and nonclassical psychological altruism. In turn, this makes it useful to look towards theoretical evolutionary biology to see what it can add to this discussion. (In the next section we return to the concrete empirical implications of the evolutionary framework laid out here.)

We will argue on largely evolutionary grounds that different motivations for helping are instantiated in the ways set out below. Our argument is not conclusive; however, we do hope it is a fruitful starting point for more detailed investigations into how different organisms make helping decisions and how their motivational profiles evolved (see also Schulz, 2011, 2013, 2016).

With this in mind, we can now consider the evolutionary pressures on the different ways of being motivated to help. To begin with, recall that some mechanisms for producing helping behaviors are neither altruistic nor egoistic simply because they involve no desires at all. Specifically, some behaviors are chosen by automatisms, and some of those behaviors are helping ones (whether strictly biologically altruistic or mutually beneficial). In addition, helping behaviors can be caused by entirely noncognitive means. Consider the way cells in multicellular organisms cooperate. Most theorists do not attribute cognitive mechanisms to individual cells in the sense in which they attribute cognitive mechanisms to multicellular organisms. Yet individual cells within multicellular organisms often benefit other cells at their own expense.

This sort of example shows that the default explanation for helping behavior is neither psychological egoism, as many have supposed (cf. the discussion in Sober, 1999, pp. 147-148), nor psychological altruism. At least from an evolutionary biological point of view, *the default explanation for helping behavior is noncognitive mechanisms and automatisms*. This is because there is good reason to think noncognitive mechanisms and automatisms evolved before representational mechanisms, are easier than representational mechanisms for evolution to produce, and are the ancestral states from which representational decision-making evolved—probably more than once. Bacteria, microbes, insects, and many other animals make decisions by relying on automatisms and noncognitive means; the reliance on representational mental states (like desires) to make decisions has evolved after that (Schulz, 2018). Given this, to the extent that one sees any of the above options as the default explanation for altruistic behaviors, it should be automatisms and noncognitive ways of making decisions. This conclusion is important because it frees us to think about the remaining types of mechanism for altruistic behavior in their own right, without presupposing that one is a priori more plausible than the others.[15]

When it comes to cognitively sophisticated organisms, there are many variables that contribute to the evolutionary pressures for helping behavior. The most crucial variables for our purposes are the following: the extent to which a helping behavior is selected over a more egoistic alternative, how easy it is for the organism to recognize a situation in which helping behavior is selected for, and the existence of social structures enforcing reciprocation between members of a group and providing rewards to those who provide help (in the form of resources, status, power, mating opportunities, etc.). These

variables give rise to a space of possible scenarios that favor different strategies for solving the helper's decision problem.

Psychological egoism is the most flexible but most cognitively demanding way to generate altruistic desires. It may be the best strategy in some cases, such as helping someone who will be thereby obligated to reciprocate in light of existing circumstances (quid pro quo). But in many cases it is too cognitively demanding, and hence unfeasible. Would-be altruists often have no reliable way of knowing whether a stranger will reciprocate help in the absence of social structures that enforce reciprocation, let alone whether helping strangers will lead to future rewards through means other than direct reciprocation (see, e.g., Baumard et al., 2013). In many practical circumstances organisms simply lack sufficient information to conduct the relevant instrumental reasoning with any hope of reaching reliable conclusions. Still, no one doubts that psychological egoism plays a role in animal psychology, including—at least in some specialized cases—in generating helping behaviors. What matters here is that, given how difficult the helper's decision problem is in the general case, psychological egoism is unlikely to be the most important way of making helping decisions. This is noteworthy in and of itself.[16]

This takes us to classical psychological altruism (ultimate altruistic desires from non-egoistic sources). Classical psychological altruism is likely to play an important role in animal psychology. For starters, we've seen that helping behaviors can be caused by automatisms, which are neither altruistic nor egoistic. If evolution can select for automatisms that produce helping behavior, there is no reason to rule out that, when more sophisticated heuristics than fixed action patterns are involved in choosing a behavior, ultimate other-involving desires that are generated either altruistically or neutrally may play a role. In other words, there is no reason to rule out that evolution can select for classical altruistic motivations. Specifically, classical psychological altruism is most efficient in cases of reliably adaptive conditions that require little modulation based on social context but still require enough modulation that relying on automatisms would be maladaptive (Schulz, 2018). Examples include helping needy offspring, needy partners, needy family members, and perhaps injured in-group members (see also Alger & Weibull, 2013).

To understand this better, note that organisms are often selected for to help others that are sufficiently closely related to them (Gardner et al., 2011; Taylor & Frank, 1996; Frank, 1998; Queller, 1992; van Veelen, 2009; Birch & Okasha, 2014). Indeed, helping direct descendants often increases an organism's direct fitness; helping other kin often increases an organism's inclusive fitness. This is typically true regardless of social circumstances: it does not matter whether there are mechanisms enforcing reciprocation or punishing freeloading.

[16] Psychological egoists decide whether to help by assessing whether helping increases their own wellbeing. If the psychological variables egoists use as a proxy for wellbeing correlate with personal or direct fitness (as opposed to inclusive fitness), egoists will not choose to help when helping increases their inclusive fitness without increasing their direct or personal fitness. Therefore, psychological egoism will fail to lead to helping behaviors in cases of selected-for evolutionary biological altruism. This further strengthens the conclusion established in the main text. This could be avoided if an egoist's proxy for personal wellbeing correlates with its inclusive fitness, but the biological plausibility of this is low—for reasons related to the ones laid out in the text.

In cognitively sophisticated animals whose behaviors are motivated by desires, this circumstance creates an evolutionary pressure towards an endogenous source of desires to help members of these special groups when they are needy. A mechanism that responds to this pressure must be able to do two things: recognize when a needy organism is a member of these special groups, and then generate a desire to help that organism. As a matter of fact, many cognitively sophisticated organisms help others roughly in proportion to how closely they are genetically related to them (Gardner et al., 2011; Strassman et al., 2011; Kuzdzal-Fick et al., 2011; West et al., 2007; Henrich and Henrich, 2007). The likely explanation is an innate, fitness-enhancing disposition to desire to help their kin: for some organisms, helping their kin cannot be done automatically—for instance, because kin recognition is too complex, or because there are many different ways to help kin, so that an organism benefits from representational reasoning about *how* to help their kin—but it is always adaptive to *somehow* help their kin.[17] This thus leads to the selection of classical altruism.[18]

Helping kin is not the only kind of circumstance where classical psychological altruism is likely to be the evolutionarily favored solution. Classical psychological altruism is likely to be favored under any circumstance with the following characteristics: relatively easy to recognize, relatively low cost, relatively high payoff, but sufficiently complex and variegated to make automatisms too rigid for the job (Schulz, 2016, 2018). One such example is a needy reproductive partner: insofar as an organism is going to reproduce and raise offspring with that partner, helping the partner is also helpful to the self. Another example is the presence of injured in-group members in highly social animals: they are easily recognizable, their recovery benefits us because it strengthens the group and our fitness depends on the group's strength, and helping them can be relatively low cost. Therein may lie a selection pressure for empathy in response to others' pain.

In sum, classical psychological altruism is likely to play a large role in animal psychology. Under any circumstances that are easy to recognize, relatively low cost, and in which representationally driven helping behavior is relatively adaptive, there is the potential for a selection pressure towards classical psychological altruism.

We now turn to nonclassical altruism. Like its predecessors, nonclassical altruism (ultimate altruistic desires from egoistic sources—primarily, reward-based learning) is likely to play a large role in animal psychology. To understand this, recall that the difference between classical and nonclassical altruism is precisely that the latter is based on (reward-based) learning. Reward-based learning is a powerful way to determine when circumstances are appropriate for helping others. Thus, nonclassical altruism is sandwiched between egoism and classical altruism: it is less cognitively demanding than egoism—which needs to derive all helping behaviors from egoistic first principles—but more flexible than classical altruism. This sandwiching makes clear when nonclassical altruism is adaptive: namely, when helping behavior is reliably adaptive in a certain situation, but this adaptiveness depends on social conditions

[17] Of course, for some organisms, helping their kin can be done automatically: see e.g. Strasssman et al. (2011); Kuzdzal-Fick et al. (2011).

[18] Note, though, that this may require complex decisions as to which kin to help (in case there are several options)—including potential future kin. See also Hausfater & Hrdy (1984) and Trivers (1974).

such as the likelihood of reciprocation, which in turn may depend on presence or absence of mechanisms rewarding help or enforcing reciprocation.

Simply put, there are circumstances where it is adaptive for organisms motivated by desires to *learn* when to help even though calculating whether to help in every occasion through instrumental reasoning is unfeasible. More specifically, nonclassical altruism is adaptive if it is inter-generationally variable whether helping is adaptive, but intra-generationally stable: it is not adaptive for organisms to be born with an innate disposition to form desires to help certain other organisms because whether this help is adaptive depends on the precise conditions the organism faces; but, if the conditions are such that helping *is* adaptive, it is also adaptive for the organism not to derive the helping behavior, every time, from egoistic ultimate desires. Rather, the organism learns when forming a desire to help certain other organisms is appropriate.

The existence of circumstances like this is well known: in fact, the evolution of learning is based on it (see e.g. Henrich, 2015; Boyd & Richerson, 2005; Fehr & Fischbacher, 2003). Therefore, it follows straightforwardly that there are circumstances where we should expect the evolution of nonclassical altruism.

Consider cognitively sophisticated social animals, whose behaviors are motivated by desires (as opposed to automatisms). Examples include wolves, vervet monkeys, and vampire bats. These organisms depend on their mutual cooperation with other members of their group for foraging, escaping predators, securing mates, raising offspring, grooming, etc. If all group members cooperate equally, a simple innate disposition to cooperate with in-group members would solve their helper's decision problem. This is what most social insects do. But these animals are capable of both reciprocal cooperation and freeloading. Freeloaders use shared resources without sharing, thus increasing their fitness at the expense of other group members (e.g., Packer and Ruttan, 1988). In addition, such animals can leave or join a group, so that the boundaries of groups are at least somewhat flexible, and members within groups can form alliances that compete with other alliances to some degree. Finally, the same member of such groups may be more or less prone to freeloading depending on circumstances. Therefore, such animals must adjust their degree of helping to circumstances such as encountering new group members and the likelihood that another group member is freeloading. In this context, an innate disposition to cooperate with in-group members may be part of the story but cannot be the whole story because it risks defeat by freeloader invasion.

Thus, social animals who live in groups with flexible boundaries and who are capable of freeloading face several different motivations: (i) to help other group members because a thriving group is also good for them, (ii) to freeload, (iii) to not help (or, even better for the group, to punish) freeloaders (other than themselves). Therefore, social animals of this kind cannot rely solely on simple innate dispositions to help in-group members, except for special circumstances to be discussed below.

A better solution involves mechanisms that generate a sense of reward when individuals act on their desire to help non-freeloading members of their group, so that these desires are reinforced. These mechanisms have two jobs: identifying correct targets for helping behavior (i.e., members of the group,

except for freeloaders) and motivating helping behavior towards the correct targets by rewarding helping behavior and punishing freeloading.

There are a number of mechanisms that appear to play this role. One is ritual-based bonding. This is a process of mutual signaling between individuals, which requires positive feedback loops with its target(s). Bonding involves the limbic system and the release of a set of hormones (oxytocin and vasopressin) and neurotransmitters (dopamine and endorphins) in social situations that are likely to involve members of the group (mates, family members, or members of a larger group). We need not be concerned with the details of the bonding mechanism, except to note that bonding creates trust between partners and a sense of reward in the presence of its target, which means that helping the bonding target is likely to generate a sense of reward—if nothing else, by promoting proximity with the bonding target.[19]

Another type of relevant mechanism is overt rewards and punishments delivered by other members of the group as a function of helping behavior. Many cognitively sophisticated species of social animals live in groups that establish complex social hierarchies. One function of such hierarchies is to maintain balance between the group members' motivation to pursue their direct or personal fitness at the expense of other group members and their motivation to cooperate with the group. This sort of mechanism for generating altruistic desires is primarily exogenous, although in order to work it requires that external rewards and punishments be met by appropriate internal changes, such as reward-based learning.[20] Group members who share food and other resources may be rewarded in the form of acceptance, status, reciprocation, mates, etc.; freeloaders should not be helped and may even be punished in the form of physical aggression, low status, expulsion from the group, etc. The extent to which a cognitively sophisticated social animal exhibits helping behavior is modulated by the extent to which such behavior is appropriately reinforced within a relevant group, which in turn depends on the exact boundaries of the group and the degree to which reinforcing stimuli are elicited during relevant circumstances (Cf. Raihani et al., 2012; Brosnan and de Waal, 2014).

In sum, nonclassical psychological altruism is likely to play a large role in animal psychology. But nonclassical altruism works primarily by *reinforcing* existing altruistic desires; it does not generate them in the first place. Thus, nonclassical altruism needs a way to generate altruistic desires in the first place. As we've already mentioned, though, psychological egoism is an unlikely source of altruistic desires in many circumstances, because there isn't enough information to establish the benefits of altruism for the

[19] Donaldson and Young (2008) review evidence that oxytocin and vasopressin modulate complex social behavior. De Drew (2012) reviews evidence that oxytocin release enables categorization of others into in-group versus out-group members, promotes trust towards in-group members, and motivates cooperation with in-group members and aggression towards out-group members. Frost (2016) provides a formal argument that ritual bonding can promote helping behavior in a way that is consistent with nonclassical psychological altruism. For two types of cooperation that would benefit from this type of mechanism, see Brosnan and de Waal (2002) on symmetry-based reciprocity and attitudinal reciprocity. See also Soares et al, 2010.

[20] An example of inappropriate internal change is instrumental reasoning aimed at avoiding punishment and reaping rewards. Organisms that respond thus are (maladaptive) psychological egoists. We are not considering that sort of response here, although it is certainly a possible one.

agent. Therefore, altruistic desires that originate in those circumstances are likely to be produced via classical psychological altruism.

For simplicity, we have focused on pure versions of the strategies we defined. We should not forget that there might be hybrid strategies. For example, it is possible that an organism wants to help their offspring both because they have learned to do so (e.g., through bonding rituals) *and* because they have an innate disposition to do so. Under certain circumstances, such mixed helping motivations may make cooperative behavior between organisms over time especially stable.

5. Predictions and Consequences

The framework we introduced, to the effect that egoism, classical altruism, and non-classical altruism is likely to be selected for, has important empirical implications.

First, our framework predicts that each of the above three helping motivations is likely to be instantiated, although it may not always be recognized as such. This allows us to reinterpret existing findings in novel and productive ways. For example, as noted above, the disposition to help human offspring in need is widely instantiated. While abortion and even infanticide are common in several cultures, they are universally seen as difficult decisions (Hausfater & Hrdy, 1984). In turn, this suggests that a *desire* to help offspring in need is a virtually universal feature of human life. This can now be recognized as offspring-focused classical altruism. Similarly, financial and other economic interactions among strangers are widely seen to be underwritten by desires to further one's own wellbeing—i.e., psychological egoism. These interactions are widely seen to be egoistic and our framework confirms this. Finally and most interestingly, here's a plausible example of nonclassical altruism. One widely noted effect of military training, discipline, and combat experience is to instill in soldiers a genuine care for their comrades. This can now be recognized as a case of non-classical altruism: the soldiers learn—through individual as well as collective rewards and punishments and the resultant bonding process—to want to their colleagues to do well. This matters, as it shows that the kind of attitude soldiers have towards each other are comparable in content—though not in origin—with those family members have for each other. In turn, this can help us understand the benefits and challenges that come from life in the armed forces.

In this way, the framework laid out in this paper (a) predicts *that* these three motivational structures are instantiated, (b) clarifies the relationships among these different helping motivations, and (c) explains why they are instantiated. So, the reason why non-classical altruism is likely to be instantiated among well-trained soldiers is that these are interactions among non-kin, in which everybody profits from a cooperative relationship, but where the benefits of freeriding can be high. In such a case—as noted earlier—non-classical altruism is likely to be adaptive. This kind of suggestion is worth investigating in more detail.

Second, the above conceptual framework makes novel predictions and suggests new avenues of investigation. In what follows, we lay out several of these predictions and suggestions for the different disciplines studying psychological altruism.

In ethics, two points need to be noted. On the one hand, the arguments of this paper suggest that the value of psychological altruism vis-à-vis egoism must be reassessed. Given the conceptual framework laid out here, the difference between psychological altruism and egoism is much narrower than is traditionally supposed. This is so not only because altruistic behavior can be reliably caused by egoistic motivational structures—as Stich et al. (2010) have argued—but also because even when altruistic behavior is caused by desires with altruistic content, such desires may be produced by egoistic mechanisms. In turn, this makes it plausible that the difference between psychological altruism and egoism is also less ethically significant than commonly supposed (see e.g., Rachels, 2000, p. 81; Schroeder, 2000, p. 396). Of course, fully establishing this last conclusion would require significantly further argumentation; here, we just want to note that our descriptive conclusions put some pressure on ethical positions that are based on a very sharp distinction between psychological altruism and egoism.

On the other hand, the evolutionary framework laid out here may yield valuable insights about how to coax our evolved psychological mechanisms into producing ethically good outcomes. For instance, our account makes clear that there are several different ways to increase the degree of psychological altruism in a population. Most obviously, the plausible existence of non-classical altruism makes clear that people can be *taught* to want to help others. A similar point applies to classical altruism. Given that (as noted above) innate dispositions to form desires to help others may need to be developmentally mediated by a functioning empathy system, ensuring that the latter does indeed develop appropriately can thus increase the prevalence of classically altruistic motivations. Both of these points can help prevent the kind of racist and discriminatory behaviors that are still so common (cf. Greene 2013). Neither of these points has been fully appreciated before: for example, typical efforts to combat bullying and harassment pay little attention to the possibility that people may be trained to no longer *want* to bully or harass.

In cognitive neuroscience, our argument complements recent advances in the subject, which also emphasize that understanding helping behavior requires investigating its underlying neurocognitive structures (cf. Gluth and Fontanesi, 2016, Greene et al., 2016, Hein et al., 2016, Kurzban et al., 2015). More specifically, we have argued that, in order to arrive at a proper understanding of these structures, we must go beyond the content of the relevant motivations and also consider their production—for only then can we develop an adequate view of the ways that organisms are driven to help (whether they are classically altruistic, nonclassically altruistic, egoistic, or impersonal). In turn, this implies that the study of psychological altruism is a diachronic problem: we need to take into account not only how the

organism is psychologically constituted at time t_1 (by assessing the contents of its desires), but also how it was psychologically constituted at time t_0 (by assessing how it produced the relevant desires).[21]

This complicates the empirical investigation of altruism, egoism, and impersonal agency (which was complex to begin with—see e.g. Batson, 1991; Stich et al., 2010). But this is not to say that this topic is intractable. There are ways to investigate the distinction between classical and nonclassical altruism empirically. For example, we have hypothesized that unconditioned empathy and bonding between kin are primary mechanisms behind classical altruism, whereas conditioned empathy and bonding to non-kin group members are primary mechanisms behind nonclassical altruism. This is a hypothesis worth investigating. Indeed, this hypothesis yields novel, testable predictions: for example, increasing bonding mechanisms among non-kin—e.g., through administering oxytocin—and rewarding a desire to help others can increase stable altruistic helping dispositions beyond just increasing trust in an economic interaction (Kosfeld et al., 2005)—these helping dispositions are likely to be maintained for long periods even if no longer encouraged.

In economics, the adaptiveness of different forms of altruism is also important. To see this, note that it is a typical assumption in much economic modeling that people are egoistically motivated. This assumption is not required by economic theory—which generally leaves the form of an agent's utility function open—but it is still often made (Kalenscher & van Wingerden, 2011; Fehr & Camerer, 2007; Falk et al., 2003; Fehr & Gaechter, 2000). Recently, though, some authors have argued that this assumption is misguided: it appears that at least sometimes, some people are motivated to help others (Rand, 2016; Fehr & Camerer, 2007; Falk et al., 2003; Fehr & Gaechter, 2000; Fehr & Schmidt, 1999; Clavien and Chapuisat 2016). What our discussion does is use evolutionary biological considerations to expand this latter position further: there are good reasons to think that human beings (among other organisms) are in fact *frequently* altruistically motivated, and for deep evolutionary reasons. Specifically, the above framework predicts that, for many humans, *very many* economic interactions—even beyond those with kin—have an altruistic component: after all, many humans live in circumstances that favor one or the other form of altruism. In turn, this prediction needs to be further considered, as it can call for a major reevaluation of much consumer behavior.

Finally, in evolutionary biology, some theorists say that all that matters for evolution is whether a behavior is altruistic, not whether it is motivated altruistically (e.g., Wilson, 2015). On the contrary, we have argued that different sorts of motivational profiles are likely to be involved in solving different portions of the helper's decision problem. As we have argued, different motivational profiles may be favored by different selection pressures, which make them adaptive under different circumstances. Therefore, to reach a deeper understanding of the ways of altruism, it is incumbent on evolutionary theorists to consider which evolutionary pressures favored one or another of the possible causes of altruistic behavior (cf. Brosnan and Bshary, 2010; Bshary & Raihani, 2017).

[21] There is an asymmetry here, in that we do not need to take this diachronic perspective when it comes to psychological egoism and impersonal agency: as noted above, these are defined just by the contents of the relevant conative states. Still, the cognitive neuroscience of helping behavior cannot ignore the diachronic perspective, on pain of missing the distinction between classical and nonclassical altruism.

6. Conclusion

The upshot is this. First, the problem of deciding when to help others—*the helper's decision problem*—is generally difficult to solve and requires heuristic solutions, which often must be selected for. Second, the solutions that are easiest to be selected for are those that involve noncognitive mechanisms and automatisms (reflexes, fixed action patterns, and the like). Therefore, the default explanation of altruistic behavior involves noncognitive mechanisms and automatisms, not psychological egoism as is often assumed. Nevertheless, many organisms are motivated by desires (including urges); this cognitive motivational structure gives rise to the distinction between psychological egoism (egoistic desires) and psychological altruism (desires to help). Third, there are two importantly different kinds of altruistic motivations: *classical psychological altruism*, which generates desires to help for others' sake, and *nonclassical psychological altruism*, which generates desires to help for one's sake. Fourth, calculating whether to behave altruistically is egoistically desirable is unfeasible or inefficient in many cases; therefore, either classical or nonclassical psychological altruism are more efficient and hence adaptive solutions of the helper's decision problem than psychological egoism. Classical altruism is most efficient when altruistic behavior is reliably adaptive without requiring much modulation based on social context. Nonclassical psychological altruism is most efficient when altruistic behavior is reliably adaptive but this adaptiveness depends on social conditions that can be learned. Thus, both kinds of psychological altruism are likely to be instantiated and selected for. Fifth, we hypothesize that unconditioned empathy and bonding between kin are primary mechanisms behind classical altruism, whereas conditioned empathy and bonding to non-kin group members are primary mechanisms behind nonclassical altruism. We submit that grounding this theory of psychological altruism in neurocognitive mechanisms, testing it empirically, and exploring its normative implications would be a fruitful interdisciplinary research program.

References

Alger, I., & Weibull, J. W. (2013). Homo Moralis; preference evolution under incomplete information and assortative matching. *Econometrica, 81*(6), 2269–2302.

Barkow, J., Cosmides, L., & Tooby, J. (Eds.). (1992). *The Adapted Mind*. Oxford: Oxford University Press.

Barrett, H. C. (2015). *The Shape of Thought: How Mental Adaptations Evolve*. Oxford: Oxford University Press.

Barrett, L., Dunbar, R., & Lycett, J. (2002). *Human Evolutionary Psychology*. Princeton, NJ: Princeton University Press.

Batson, D. (1991). *The Altruism Question: Toward a Social-Psychological Answer*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Baumard, N., André, J. B., & Sperber, D. (2013). A Mutualistic Approach to Morality. *Behavioral and Brain Sciences*, 36(1), 59-122.

Birch, J., & Okasha, S. (2014). Kin Selection and Its Critics. *BioScience*.

Böckler, A., Tusche, A., & Singer, T. (2016). The Structure of Human Prosociality: Differentiating Altruistically Motivated, Norm Motivated, Strategically Motivated, and Self-Reported Prosocial Behavior. *Social Psychological and Personality Science, 7*(6), 530-541.

Boyd, R., & Richerson, P. (2005). *The Origin and Evolution of Cultures*. Oxford: Oxford University Press.

Brosnan, S. F., Bshary, R. (2010). Cooperation and deception: from evolution to mechanisms. *Phil. Trans. R. Soc. B*, 365: 2593–2598.

Brosnan, S. F., and de Waal, F. B. M. (2002). "A Proximate Perspective on Reciprocal Altruism." *Human Nature*, 13 (1), 129–152.

Brosnan, S. F., & de Waal, F. B. (2014). Evolution of responses to (un)fairness. *Science* 346: 1251776.

Bshary, R., & Raihani, N. J. (2017). Helping in humans and other animals: a fruitful interdisciplinary dialogue. *Proc. R. Soc. B*, 284: 20170929. http://dx.doi.org/10.1098/rspb.2017.0929

Buller, D. (2005). *Adapting Minds*. Cambridge, MA: MIT Press.

Buss, D. M. (2014). *Evolutionary Psychology: The New Science of the Mind* (5th ed.). Boston: Allyn & Bacon.

Butler, J. (1726 / 1965). Fifteen Sermons upon Human Nature. In L. A. Selby-Bigge (Ed.), *British Moralists* (pp. 1-227). New York: Dover.

Chudek, Maciek; Zhao, Wanying, and Henrich, Joseph (2013). Culture-Gene Coevolution, Large Scale Cooperation, and the Shaping of Human Social Psychology. In K. Sterelny, R. Joyce, B. Calcott, and B. Fraser (eds.). *Cooperation and Its Evolution*. Cambridge, MA: MIT Press, pp. 425-457.

Churchland, P. (2010). *Braintrust*. Princeton, NJ: Princeton University Press.

Cialdini, R. B., Brown, S. L., Lewis, B. P., Luce, C., & Neuberg, S. L. (1997). Reinterpreting the Empathy-Altruism Relationship: When One Into One Equals Oneness. *J Pers Soc Psychol, 73*(3), 481-494.

Clavien, C. (2011). "Altruistic Emotional Motivation: An Argument in Favour of Psychological Altruism." in K. Plaisance & T. Reydon (eds.), *Philosophy of Behavioral Biology. Boston Studies in Philosophy of Science*, *Volume 282*. Springer, 275-296.

Clavien, C. & Chapuisat, M. (2013). "Altruism across disciplines: one word, multiple meanings." *Biology & Philosophy*, 28 (1), 125–140.

Clavien, C. & Chapuisat, M. (2016). "The evolution of utility functions and psychological altruism." *Studies in the History and Philosophy of the Biological and Biomedical Sciences*, 56, 24-31.

De Dreu, C. K. (2012). Oxytocin modulates cooperation within and competition between groups: an integrative review and research agenda. *Hormones and Behavior*, 61: 419–428.

de Waal, F. B. M. (2008). "Putting the Altruism Back into Altruism: The Evolution of Empathy." *Annual Review of Psychology*, 59: 279-300.

Donaldson, Z. R., & Young, L. J. (2008) Oxytocin, vasopressin, and the neurogenetics of sociality. *Science*, 322, 900–904.

Erev, I., & Roth, A. E. (2014) Maximization, learning, and economic behavior. *Proc. Natl Acad. Sci. USA*, 111: 10 818–10 825.

Falk, A., Fehr, E., & Fischbacher, U. (2003). On the Nature of Fair Behavior. *Economic Inquiry, 41*(1), 20-26.

Fehr, E., & Camerer, C. F. (2007). Social neuroeconomics: the neural circuitry of social preferences. *Trends Cogn Sci, 11*(10), 419-427.

Fehr, E. & Fischbacher, U. (2003). The nature of human altruism. *Nature, "*, 785-791

Fehr, E., & Gaechter, S. (2000). Fairness and Retaliation: The Economics of Reciprocity. *The Journal of Economic Perspectives, 14*, 159-181.

Fehr, E., & Schmidt, K. M. (1999). A Theory of Fairness, Competition, and Cooperation. *The Quarterly Journal of Economics, 114*(3), 818-868.

Frank, S. A. (1998). *Foundations of Social Evolution*. Princeton: Princeton University Press.

Frost, K. (2016). "Coevolutionary Dynamics of Costly Bonding Ritual and Altruism." doi: http://dx.doi.org/10.1101/060624.

Gächter, S., Herrmann, B., & Thöni, C. (2010). Culture and cooperation. *Phil. Trans. R. Soc. B*, 365: 2651-2661.

Gardner, A., & West, S. A. (2010). Greenbeards. *Evolution, 64*(1), 25-38.

Gardner, A., West, S. A., & Wild, G. (2011). The genetical theory of kin selection. *J Evol Biol, 24*(5), 1020-1043.

Garson, J. (2014). *The Biological Mind: A Philosophical Introduction*. Routledge.

Garson, J. (2016). Two Types of Psychological Hedonism. *Stud Hist Philos Biol Biomed Sci., 56*, 7-14.

Gigerenzer, G., Todd, P. M., & ABC Research Group (1999). *Simple heuristics that make us smart*. Oxford, UK: Oxford University Press.

Hutchinson, J. M., Gigerenzer, G. (2005). Simple heuristics and rules of thumb: where psychologists and behavioural biologists might meet. *Behavioral Processes*, 69: 97–124.

Glimcher, P. W., Dorris, M. C., & Bayer, H. M. (2005). Physiological utility theory and the neuroeconomics of choice. *Games Econ Behav, 52*(2), 213-256.

Gluth, S., & Fontanesi, L. (2016). "Wiring the Altruistic Brain." *Science* 351(6277), 1028-1029.

Godfrey-Smith, P. (1996). *Complexity and the Function of Mind in Nature*. Cambridge: Cambridge University Press.

Godfrey-Smith, P. (2008). Varieties of Population Structure and the Levels of Selection. *British Journal for the Philosophy of Science, 59*, 25-50.

Gintis, H. (2003) – The Hitchhiker's Guide to Altruism: Genes, Culture, and the Internalization of Norms. *Journal of Theoretical Biology,* 220, 407-418.

Grafen, A. (2006). Optimization of Inclusive Fitness. *J Theor Biol, 238*, 541-563.

Green, D. M., & Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. New York: Wiley.

Greene, J. D. (2013). *Moral Tribes*. Penguin.

Greene, J. D., Morrison, I., & Seligman, M. E. P. (2016). *Positive Neuroscience*. Oxford: Oxford University Press.

Hausfater, G., & Hrdy, S. B. (Eds.). (1984). *Infanticide: Comparative and Evolutionary Perspectives*. Chicago: Aldine Transactions.

Hein, G., Morishima, Y, Leiberg, S., Sul, S., & Fehr, E. (2016). "The Brain's Functional Network Architecture Reveals Human Motives." *Science* 351(6277), 1074-1078.

Henrich, J. (2015). *The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter*. Princeton, NJ: Princeton University Press.

Henrich, N., & Henrich, J. (2007). *Why Humans Cooperate: A Cultural and Evolutionary Explanation*. Oxford: Oxford University Press.

Henrich, J., & McElreath, R. (2007). Dual-Inheritance Theory: The Evolution of Human Cultural Capacities and Cultural Evolution. In R. Dunbar & L. Barrett (Eds.), The Oxford Handbook of Evolutionary Psychology (pp. 555-570). Oxford: Oxford University Press.

Hobbes, T. (1651). *Leviathan*.

Houston, A. I., & McNamara, J. M. (1999). *Models of Adaptive Behaviour: An Approach Based on State*. Cambridge: Cambridge University Press.

Jensen, K. (2012). Who Cares? Other-Regarding Concerns—Decisions with Feeling. In P. Hammerstein & J. R. Stevens (Eds.), *Evolution and the Mechanisms of Decision Making* (pp. 299-317). Cambridge, MA: MIT Press.

Kalenscher, T., & van Wingerden, M. (2011). Why we should use animals to study economic decision making – a perspective. *Frontiers in Neuroscience, 5*, 1-11.

Kandel, E. (2001). "The Molecular Biology of Memory Storage: A Dialogue Between Genes and Synapses." *Science*, 294 (5544): 1030-8.

Klimecki, O. M. (2015). "The Plasticity of Social Emotions." *Social Neuroscience*, 10(5): 466-473.

Klimecki, O. M., Mayer, S. V., Jusyte, A., Scheeff, J. & Schönenberg, M. (2016). "Empathy promotes altruistic behavior in economic interactions." *Scientific Reports*, 6: 31961. DOI: 10.1038/srep31961.

Kosfeld, M., Heinrichs, M., Zak, P. J., Fischbacher, U. and Fehr, E. (2005). "Oxytocin increases trust in Humans", *Nature*, 435 (7042): 673–6.

Kurzban, R., Burton-Chellew, M. N., & West, S. A. (2015). "The Evolution of Altruism in Humans." *Annual Review of Psychology***,** 66: 575-599.

Kuzdzal-Fick, J. A., Fox, S. A., Strassmann, J. E., & Queller, D. C. (2011). High relatedness is necessary and sufficient to maintain multicellularity in Dictyostelium. *Science, 334*, 1548-1551.

Morillo, C. (1990). The reward event and motivation. *Journal of Philosophy, 87*, 169-186.

Nagel, T. (1970). *The Possibility of Altruism*. Princeton: Princeton University Press.

Northcott, R. & Piccinini, G. (2018). Conceived this Way: Innateness Defended. *Philosophers' Imprint* (forthcoming).

Okasha, S. (2006). *Evolution and the Levels of Selection*. Oxford: Oxford University Press.

Okasha, S., "Biological Altruism", *The Stanford Encyclopedia of Philosophy* (Fall 2013 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2013/entries/altruism-biological/>.

Packer, C., & Ruttan, L. (1988). The evolution of cooperative hunting. *American Naturalist*, 132: 159–198.

Queller, D. C. (1985). Kinship, Reciprocity and Synergism in the Evolution of Social Behavior. *Nature, 318*(28), 366-367.

Queller, D. C. (1992). Quantitative Genetics, Inclusive Fitness and Group Selection. *American Naturalist, 139*, 540-558.

Raihani, N. J., Pinto, A. I., Grutter, A. S., Wismer, S., & Bshary, R. (2012). Male cleaner wrasses adjust punishment of female partners according to the stakes. *Proc. R. Soc. B*, 279: 365–370.

Rachels, J. 2000. Naturalism. In H. LaFollette (ed.), *The Blackwell Guide to Ethical Theory*. Oxford: Blackwell Publishing, pp. 74-91.

Ramsey, G. (2016). Can altruism be unified? *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 56*, 32-38.

Rand, D. (2016). Cooperation, Fast and Slow: Meta-Analytic Evidence for a Theory of Social Heuristics and Self-Interested Deliberation. *Psychol Sci*.

Rachlin, H. (2002). "Altruism and Selfishness." *Behavioral and Brain Sciences*, 25, 239–296.

Richardson, R. (2007). *Evolutionary Psychology as Maladapted Psychology*. Cambridge, MA: MIT Press.

Richerson, P., Baldini, R., Bell, A. V., Demps, K., Frost, K., Hillis, V., Mathew, S., Newton, E. K., Naar, N., Newson, L., Ross, C., Smaldino, P. E., Waring, T. M., & Zefferman, M. (2016). "Cultural group selection plays an essential role in explaining human cooperation: A sketch of the evidence." Behavioral and Brain Sciences. doi:10.1017/S0140525X1400106X

Rubin, H. (2018). The Debate over Inclusive Fitness as a Debate over Methodologies. *Philosophy of Science, 85*(1), 1-30.

Schroeder, T. (2004). *Three Faces of Desire*. Oxford: Oxford University Press.

Schroder, W. 2000. Continental Ethics. In H. LaFollette (ed.), *The Blackwell Guide to Ethical Theory*. Oxford: Blackwell Publishing, pp. 375–399.

Schulz, A. (2011). Sober & Wilson's Evolutionary Arguments for Psychological Altruism: A Reassessment. *Biology and Philosophy, 26*, 251-260.

Schulz, A. (2013). The benefits of rule following: A new account of the evolution of desires. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 44*(4, Part A), 595-603.

Schulz, A. (2016). Altruism, Egoism, or Neither: A Cognitive-Efficiency-Based Evolutionary Biological Perspective on Helping Behavior. *Stud Hist Philos Biol Biomed Sci, 56*, 15-23.

Schulz, A. (2017). The Evolution of Empathy. In H. Maibom (ed.). *Routledge Handbook of the Philosophy of Empathy.* London: Routledge, pp. 64-73.

Schulz, A. (2018). *Efficient Cognition: The Evolution of Representational Decision Making.* Cambridge, MA: MIT Press.

Skyrms, B. (1996). *Evolution and the Social Contract*. Cambridge: Cambridge University Press.

Skyrms, B. (2004). *The Stag Hunt and the Evolution of Social Structure*. Cambridge: Cambridge University Press.

Soares, M. C., Bshary, R., Fusani, L., Goymann, W., Hau, M., Hirschenhauser, K., & Oliveira, R. F. (2010). Hormonal mechanisms of cooperative behavior. *Phil. Trans. R. Soc. B*, 365: 2737–2750.

Sober, E. (1994). The adaptive advantage of learning and a priori prejudice. *Ethology and Sociobiology, 15*(1), 55-56.

Sober, E. (1999). Psychological Egoism. In: H. LaFollette (Ed.), *The Blackwell Guide to Ethical Theory* (pp. 129-148). Oxford: Blackwell.

Sober, E. (2001). The Two Faces of Fitness. In R. Singh, D. Paul, C. Krimbas & J. Beatty (Eds.), *Thinking about Evolution: Historical, Philosophical, and Political Perspectives* (pp. 309-321). Cambridge: Cambridge University Press.

Sober, E., and Wilson, D. S. (1998). *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Cambridge, MA: Harvard University Press.

Stevens, J. R. and Hauser, M. D. (2004). " Why be nice? Psychological constraints on the evolution of cooperation." *Trends in Cognitive Sciences*, 8 (2) 60–65.

Stich, S. (2007). Evolution, Altruism and Cognitive Architecture: A Critique of Sober and Wilson's Argument for Psychological Altruism. *Biology and Philosophy, 22*, 267-281.

Stich, S. (2016). "Why there might not be an evolutionary explanation for psychological altruism." *Studies in the History and Philosophy of the Biological and Biomedical Sciences*, 56, 3–6.

Stich, S., Doris, J., and Roedder, E. (2010). Altruism. In J. Doris and the Moral Psychology Research Group (eds.). *The Moral Psychology Handbook*. Oxford: Oxford University Press, pp. 147-205.

Strassmann, J. E., & Queller, D. C. (2011). Evolution of cooperation and control of cheating in a social microbe. PNAS, 108, 10855–10862.

Strassmann, J. E., Gilbert, O. M., & Queller, D. C. (2011). Kin discrimination and cooperation in microbes. *Annu Rev Microbiol, 65*, 349-367.

Taylor, P. D., & Frank, S. A. (1996). How to Make a Kin Selection Model. *J. Theor. Biol., 180*, 27-37.

Thomson, E. & Piccinini, G. (2018). Neural Representations Observed. *Minds and Machines* 28 (1), 191-235.

Trivers, R. (1974). Parent-Offspring Conflict. *American Zoologist, 14*, 247-262.

van Veelen, M. (2009). Group Selection, Kin Selection, Altruism, and Cooperation: When Inclusive Fitness is Right and When It Can Be Wrong. *J Theor Biol, 259*, 589-600.

West, S. A., Griffin, A. S., & Gardner, A. (2007). Social semantics: altruism, cooperation, mutualism, strong reciprocity and group selection. *J Evol Biol, 20*(2), 415-432.

West, S. A., El Mouden, C., & Gardner, A. (2011). Sixteen common misconceptions about the evolution of cooperation in humans. *Evolution and Human Behavior, 32*(4), 231-262.

Wilson, D. S. (2015). Does Altruism Exist? Culture, Genes, and the Welfare of Others. Yale University Press.