BEHAVIORAL AND BRAIN SCIENCES (2002) **25,** 239–296
Printed in the United States of America

Altruism and selfishness

Howard Rachlin
Psychology Department, State University of New York, Stony Brook, NY 11794-2500
howard.rachlin@sunysb.edu

Abstract: Many situations in human life present choices between (a) narrowly preferred particular alternatives and (b) narrowly less preferred (or aversive) particular alternatives that nevertheless form part of highly preferred abstract behavioral patterns. Such alternatives characterize problems of self-control. For example, at any given moment, a person may accept alcoholic drinks yet also prefer being sober to being drunk over the next few days. Other situations present choices between (a) alternatives beneficial to an individual and (b) alternatives that are less beneficial (or harmful) to the individual that would nevertheless be beneficial if chosen by many individuals. Such alternatives characterize problems of social cooperation; choices of the latter alternative are generally considered to be altruistic. Altruism, like self-control, is a valuable temporally-extended pattern of behavior. Like self-control, altruism may be learned and maintained over an individual's lifetime. It needs no special inherited mechanism. Individual acts of altruism, each of which may be of no benefit (or of possible harm) to the actor, may nevertheless be beneficial when repeated over time. However, because each selfish decision is individually preferred to each altruistic decision, people can benefit from altruistic behavior only when they are committed to an altruistic pattern of acts and refuse to make decisions on a case-by-case basis.

Keywords: addiction; altruism; commitment; cooperation; defection; egoism; impulsiveness; patterning; prisoner's dilemma; reciprocation; reinforcement; selfishness; self-control

1. Introduction

1.1. Biological compatibility

Altruism and selfishness, like free will and determinism, seem to be polar opposites. Yet, as with free will and determinism (Dennett 1984), the apparent incompatibility may be challenged by various forms of compatibility. From a biological viewpoint selfishness translates into survival value. Evolutionary biologists have been able to reconcile altruism with selfishness by showing how a biological structure mediating altruistic behavior could have evolved. (The next section will briefly summarize one such demonstration.) This structure is assumed to be more complex than ordinary mechanisms that mediate selfish behavior but in essence is no different from them. The gazelle that moves toward the lion (putting itself in danger but showing other gazelles where the lion is) may thus be seen as acting according to the same principles as the gazelle that takes a drink of water when it is thirsty. The desire to move toward the lion stands beside the desire to drink.

Evolutionary biologists do not conceive of behavior itself as being passed from generation to generation; rather, some mechanism, in this case an internal mechanism – a structure of nervous connections in the brain – is hypothesized to be the evolving entity. Altruism as it appears in behavior is conceived as the action of that mechanism developed over the lifetime of the organism. Tooby and Cosmides (1996, p. 125) compare the structure of the altruism mechanism to that of the eye: "We think that such adaptations will frequently require complex computations and suspect that at least some adaptations for altruism may turn out to rival the complexity of the eye."

This biological compatibility makes contact with modern cognitive and physiological psychology (Sober & Wilson 1998). Cognitive psychology attempts to infer the mechanism's principles of action (its software) from behavioral observation and manipulation, while physiological psychology attempts to investigate the mechanism itself (its hardware).

From the biological viewpoint, altruistic acts differ from selfish acts by virtue of differing internal mediating mechanisms; altruism becomes a motive like any other. In this view, a person leaves a tip in a restaurant to which he will never return because of a desire for fairness or justice, a desire generated by the restaurant situation and the altruistic mechanism within him, which is satisfied by the act of leaving the tip. Similarly, he eats and drinks at the restaurant because of desires generated by internal mechanisms of hunger and thirst. For the biologist, Person A's altruistic behavior (behavior that benefits others at a cost to A) would be fully explained if Person A were shown to possess the requisite internal altruistic mechanism. Once the mechanism were understood, no further explanation would be required.

The problem with this conception, from a behavioral viewpoint, is not that it postulates an internal mechanism as

HOWARD RACHLIN obtained a PhD at Harvard University in 1965. He is currently a Distinguished Professor of Psychology at the State University of New York at Stony Brook. He has written six books including *Behavior and Mind* (1994) and *The Science of Self-Control* (2000). He has published three previous target articles in BBS: "Maximization theory in behavioral psychology," with John Kagel, Ray Battalio, and Leonard Green (1981), "Pain and behavior" (1985), and "Self-control: beyond commitment" (1995).

© 2002 Cambridge University Press 0140-525X/02 $12.50

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

such. (After all, no behavior is possible without internal neural structure.) The problem is that in focusing on an inherited internal mechanism, the role of learning over an organism's lifetime tends to get ignored. To develop normally, eyes have to interact with the environment. But we inherit good eyesight or bad eyesight. If our altruism mechanisms are like our visual mechanisms we are doomed to be more or less selfish depending on our genetic inheritance. This is a sort of genetic version of Calvinism. Experience might aid in the development of altruistic mechanisms. Environmental constraints imposed by social institutions – family, religion, government – might act on selfish motives (like glasses on eyesight) to make them conform to social good. But altruistic behavior as such, according to biological theory, would depend (as eyesight depends) much more on genes than on experience.

The present article, does not deny the existence of such mechanisms. A large part of human altruism and a still larger part of nonhuman altruism may well be explained in terms of inherited mechanisms based on genetic overlap. However, the mechanisms underlying these behaviors would have evolved individually. The mechanism responsible for the ant's self-sacrifice in defense of a communal nest would differ from that responsible for a mother bear's care for her cubs. There remains some fraction of altruistic action, especially among humans, that cannot be attributed to genetic overlap. For the remainder of this article I will symbolize such actions by the example of a woman who runs into a burning building to save someone else's child. I mean this example to stand for altruistic actions not easily attributed to genetic factors. Biological compatibility attributes such an act not to a specific mechanism for running into burning buildings, but to a general mechanism for altruism itself. The present article will argue that it is unnecessary to postulate the existence of such a general mechanism. I claim, first, that altruism may be learned over an individual's lifetime and, second, that it is learned in the same way that self-control is learned – by forming particular acts into coherent patterns of acts. The woman who runs into a burning building to save someone else's child does so not by activating an innate self-sacrificing tendency but by virtue of the same learning process she uses to control her smoking, drinking, or weight.

1.2. Behavioral compatibility

For biological compatibility, selfishness translates into survival value; for behavioral compatibility, selfishness translates into reinforcement.[1] From a behavioral viewpoint, an altruistic act is not motivated, as an act of drinking is, by the state of an internal mechanism; it is rather a particular component that fits into an overall pattern of behavior. Given this, the important question for the behaviorist is not, "What reinforces a particular act of altruism?" – for this particular act may not be reinforced; it may never be reinforced; it may be punished – but, "What are the patterns of behavior that the altruistic act fits into?"

To explain why a woman might risk her life to save someone else's child, it would be a mistake to look for current or even future reinforcers of the act itself. By definition, as an altruistic act, it is not reinforced. In economic terms, adding up its costs and benefits results in a negative value. Some behavioristic analyses of altruism have tried to explain particular altruistic acts in terms of delayed rather than immediate reinforcement (Ainslie 1992; Platt 1973). But delayed reinforcers, after being discounted, may have significant present value, even for nonhumans (Mazur 1987). If the present value of a delayed reward is higher than the cost of the act, it is hard to see how the act can be altruistic. It is certainly not altruistic of the bank to lend me money just because I will pay them back later rather than now. If the woman who risked her life to run into the burning building to save someone else's child were counting on some later reward or sequence of rewards to counterbalance her risk (say ten million dollars, to be paid over the next ten years, offered by the child's parents), her action would be no more altruistic than that of the bank when it lends me money.

This narrow behavioral view of altruism has been criticized by social psychologists (e.g., Edney 1980) but the criticism focuses mostly on the behaviorism rather than on the narrowness of the view. These critics have merely replaced, as an explanatory device, the present action of delayed rewards with the present action of internal mechanisms. I argue here that it is a mistake to look for the cause of a specific altruistic act either in the environment or in the interior of the organism. Rather, the cause of the altruistic act is to be found in the high value (the reinforcing value, the survival value, the function) of the act as part of a pattern of acts, or as a *habit* (provided habit is seen as a pattern of overt behavior extended in time rather than, as sometimes seen in psychology, as an internal state). According to the present view, a woman runs into a burning building to save someone else's child (without the promise of money) not because she is compelled to do so by some internal mechanism, nor because she has stopped to calculate all costs and benefits of this particular act; if she did stop to calculate she would arrive at a negative answer and not do the act. Rather, this act forms part of a pattern of acts in her life, a pattern that is valuable in itself, apart from the particular acts that compose it. The pattern, as a pattern of overt behavior, is worth so much to her that she would risk dying rather than break it.

Biological compatibility says that a particular altruistic act is itself of high value by virtue of an inherited general altruistic mechanism. Learning would enter into the development of altruism, according to biological compatibility, only in the minimal sense that a baby has to learn how to eat. The mechanism is there, the biologist says; you need only to learn how to use it. Behavioral compatibility says, on the other hand, that the altruistic act itself is of low value and remains of low value. What is highly valued is a temporally extended pattern of acts into which the particular act fits. The role of the hypothesized internal altruistic mechanism in biological compatibility – to provide a motive for otherwise unreinforced particular acts – is taken, in behavioral compatibility, by the highly valued pattern of acts. Learning of altruism, the behavioral compatibilist says, is learning to perform relatively low valued particular acts as part of a highly valued pattern. Thus, from a behavioral viewpoint, particular altruistic acts are not in themselves fundamentally selfish; rather, an altruistic act is selfish only by virtue of the high value of the pattern.

1.3. Teleological behaviorism

The kind of behaviorism that this view embodies is called "teleological behaviorism" (Baum 1994; Rachlin 1994; Stout 1996). Aristotle's psychology and ethics are behavioristic in this teleological sense: For Aristotle, a particular action has no meaning by itself; the meaning of an action resides in

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

habits of overt behavior as they are played out in time, not in internal mechanistic or spiritual events; whether a particular act is good or bad depends on the habit into which it fits. In Aristotle's conception of science, habits are *final causes* of the particular acts that comprise them. While a particular ethical act may be caused (in the sense of *efficient cause*) by the action of an internal mechanism, it is caused (in the sense of *final cause*) by an abstract pattern of overt behavior. It is the final cause that determines whether the particular act is good or bad, altruistic or selfish.

Teleological behaviorism retains Aristotle's final-cause system of explanation in psychology. For example, it explains motives in terms of habits rather than habits in terms of motives. It is at least arguable that we will not be able to uncover the mechanisms underlying altruistic behavior until we gain a clear idea of what altruistic behavior is in its own terms – as a kind of habit. That is the purpose of this target article.

1.4. Outline

Altruism and selfishness were introduced in section 1 as apparently contradictory but nevertheless compatible behaviors. Particular altruistic acts are compatible with a larger selfishness – selfishness on a more abstract level. The introduction is followed in section 2 by a discussion of group selection, a biological compatibility between altruism of the individual relative to other members of a group and selfishness (increased survival) of group members relative to those of other groups. section 3 draws an analogy between group selection and self-control; just as particular acts of self-sacrifice are compatible with a more abstract benefit to a group of individuals, so particular unreinforced acts are compatible with a more abstract long-term benefit to the individual. Section 4 tightens the analogy with more formal definitions of both self-control and altruism. Whether a person acts impulsively or selfishly, on the one hand, versus temperately or altruistically, on the other, depends on the degree to which that person structures particular acts in patterns. Such structuring is discussed in section 5 on commitment. If the analogy between self-control and altruism reflects a fundamental correspondence, altruism may be explained as self-control has been explained – as a choice between high valued particular acts and higher valued patterns of acts. Section 6 describes how the principles of reinforcement and punishment, which have been used to determine the value of self-control alternatives, may apply to social cooperation. Section 7 presents an experiment showing that behavior in a laboratory social-cooperation game depends strongly on the game's context. Sections 8 and 9 deal with potential objections. Section 8 claims that altruism cannot be fully explained in biological terms, without the concept of reinforcement. Section 9 claims that altruism cannot be fully explained in Skinnerian terms, without the concept of intrinsic reinforcement of behavioral patterns. Section 10 concludes that altruism as well as self-control involves organization of behavior in patterns and choosing among patterns as wholes.

2. Group selection

Biologists have speculated that the degree of common interest between organisms is fundamentally reflected in their shared genes (Dawkins 1976/1989). The innate tendency of any organism to sacrifice its own interests for those of another organism would then depend on the degree to which their genes overlapped. To the degree that closeness of familial relationship correlates with genetic overlap, innate altruism should be greatest within families and decrease as overlap decreases in the population. The behavior of a mother who ran into a burning building to save *her own* child would thus be explained. But the many documented cases of altruism with respect to strangers (that of saints, heroes, and the like) would not be explained. Why would a mother ever run into a burning building, risking her own life (100% genetic overlap with herself), to save *someone else's* child?

Some principle other than genetic overlap seems to be necessary to explain the inheritance of an altruism that goes beyond the family. Recently, Sober and Wilson (1998) described such a principle – "group selection of altruism." To understand group selection you first have to understand a kind of social contingency called, "The Prisoner's Dilemma." An example of a prisoner's dilemma game (in this case, a multi-person prisoner's dilemma) is a game that I have, for the last ten years or so, been playing with the audience whenever I present the results of my research at university colloquia or conferences. I begin by saying that I want to give the audience a phenomenal experience of ambivalence. Index cards are then handed to ten randomly selected people and the others are asked to imagine that they had gotten one of the cards. They choose among hypothetical monetary prizes by writing either *Y* or *X* on the card. The rules of the game (projected on a screen behind me while I talk) are as follows:

1. If you choose *Y* you get $100 times *N*.
2. If you choose *X* you get $100 times *N* plus a bonus of $300.
3. *N* equals the number of people (of the 10) who choose *Y*.

Then I point out the consequences of each choice as follows: "You will always get $200 more by choosing *X* than by choosing *Y*. Choosing *X* rather than *Y* decreases *N* by 1 (Rule #3), costing you $100; but if you chose *X*, you also gain the $300 bonus (Rule #2). This results in a $200 gain for choosing *X*. Logic therefore says that you should choose *X*, and any lawyer would advise you to do so. The problem is that if you all followed the advice of your lawyers and chose *X*, $N = 0$, and each of you would get $300; while if you all ignored the advice of your lawyers and chose *Y*, $N = 10$ and each of you would get $1,000." Sometimes, depending on the audience, I illustrate these observations with a diagram like Figure 1 (bold labels).

Then I ask the ten people holding cards to make their choices, imagining as best they can what they would choose if the money were real, and letting no one else see what they have chosen. Then I collect the cards and hold them until I finish my lecture. I have done this demonstration or its equivalent dozens of times with audiences ranging from Japanese psychologists to Italian economists. The result is an approximately even split between cooperation (choosing *Y*) and defection (choosing *X*), indicating that the game does create ambiguity. Although the money won by members of my audiences is entirely hypothetical, significant numbers of subjects in similar experiments in my laboratory, with real albeit lesser amounts of money, have also chosen *Y*.[2]

Figure 1 (labels in bold typeface) represents the contin-

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press



Figure 1. **Bold typeface:** Contingencies of 10-person prisoner's dilemma experiment. *Italic typeface:* Contingencies of self-control, "primrose path", experiment (1 player, successive choices) to be described later. [In brackets]: Contingencies faced by an alcoholic. In all three cases, particular choices of X [having a single drink] are always worth more than particular choices of Y [refusing a single drink], yet on the average it is better to choose Y [to drink at a low rate].

gencies of the prisoner's dilemma game that I ask my audience to play. Point *A* represents the condition where everyone cooperates. Point *C* represents the condition where everyone defects. The line from *A* to *C* represents the average (hypothetical) earnings per person at each value of *N* (the inverse of the x-axis). Clearly, the more people who cooperate, the greater the average earnings. But, as is shown by the two lines, *ABC* (representing the return to each player who defects) and *ADC* (representing the return to each player who cooperates), an individual always earns more by defecting than cooperating.

Suppose, instead of hypothetically giving money to each player, I instead pooled the money each player earned (still hypothetical) and donated it to the entertainment fund of whatever institution I lectured at. Given this common interest it would now pay for every individual to choose *Y;* a choice of *Y* by any individual would increase *N* by 1 for all ten players, gaining $1,000 at a cost of the individual player's $300 bonus, for a net gain to the pool of $700. A common interest thus tends to reinforce cooperation in prisoner's dilemma games.

Group selection relies on common interest. A highly simplified version of group selection runs as follows: Consider a population of organisms divided into several relatively isolated groups (tribes, for example). Within each tribe there are some altruists and some selfish individuals ("egoists") interacting with each other repeatedly in multi-person prisoner's dilemma-like games such as the one with which I introduce my lectures, except, instead of monetary reward, the players receive more or less fitness – that is, ability to reproduce. In these games, the altruists tend to cooperate while the egoists tend to defect. Within each group (as in the prisoner's dilemma) altruists always lose out to egoists. However, those groups originally containing many altruists grow much faster than those originally containing many egoists – because cooperation benefits the group more than defection does.

Consider the case of teams, such as basketball teams, playing in a league. It is commonly accepted that, all else being equal, teams with individual players who play unselfishly will beat teams with individual players who play selfishly; however, within each team, the most selfish players will score the most points. Imagine now, that instead of scoring points and winning or losing games, the teams competed for reproductive fitness. Then the number of players on teams with a predominance of unselfish players would grow rapidly while that of teams with a predominance of selfish players would grow slowly or (in competition for scarce resources) shrink – the group effect. Although, within each team, selfish players would still increase faster than unselfish ones (the individual effect), this growth could well be overwhelmed by the group effect. As time goes on, the *absolute* number of unselfish individuals (altruists) could increase faster across the whole population than the *absolute* number of egoists even though within each group the *relative* number of altruists decreases. If the groups remained rigidly divided, eventually, because the relative number of altruists is always decreasing within each group, the absolute number would begin to decrease as well. However, if, before this point is reached, the groups mixed with each other and then re-formed, the process would begin all over again and altruists might maintain or increase their gains. Again, this is a highly simplified version of the argument. But the essential point is that while individual altruists may always be at a disadvantage relative to egoists, groups of altruists may be at an advantage relative to groups of egoists.

Nothing in the present article argues against group selection. Organisms may be born with greater or lesser biological tendencies to be altruistic. But, it does not follow from group selection that altruistic behavior is incompatible with a larger individual selfishness. Sober and Wilson (1998) consider only two forms of human selfishness: the selfishness which desires maximization of consumer goods and that which desires (immediate) "internal, psychological benefits" (p. 2). They do not consider individual behavior in the long run and in the abstract. They leapfrog over behavioral contingencies that may cause behavioral change (contingencies analogous to the group selection processes they have just developed) and proceed directly to "delve below the level of behavior" (p. 194) to an internal cognitive mechanism hypothesized to mediate between the biological selective process and altruistic behavior. Their cognitive psychology may well be correct but it is not clear how (or even whether), according to their psychology, altruism might emerge from selfishness over an organism's lifetime. If it implies that we are born with fixed proportions of selfish and altruistic motives, and that experience cannot teach us to alter those proportions, then their theory is not as optimistic as Sober and Wilson seem to think; it will not be of much use to those of us trying, despite our weaknesses, to live a better life.

3. Altruism and self-control

The contingencies of my lecture demonstration of ambivalence in a social prisoner's dilemma situation correspond to

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

those of "primrose-path" experiments with individual subjects facing an intertemporal dilemma (Herrnstein 1991; Herrnstein & Prelec 1992; Herrnstein et al. 1986; Heyman 1996; Kudadjie-Gyamfi 1998; Kudadjie-Gyamfi & Rachlin 1996). In the prisoner's dilemma situation illustrated in Figure 1 (bold typeface) each of many subjects make a single choice between *X* and *Y*. In primrose path experiments, on the other hand, a single subject makes repeated choices between *X* and *Y*. The rules of the primrose path experiment, usually not told to the subjects, parallel those of the social cooperation experiments. A typical set of rules follows:

1. Each choice of *Y* gains *N* points (convertible to money at the experiment's end).
2. Each choice of *X* gains *N* points plus a bonus of 3 points.
3. *N* equals the number of *Y* choices in the last 10 trials.[3]

Figure 1 (labels in italic typeface) illustrates these contingencies in a corresponding way to social cooperation. The reward for choosing *X* is always greater than that for choosing *Y* but overall reward (proportional to the ordinate of line *AC*) would be maximized by repeatedly choosing *Y*. Ambivalence (reflected in social cooperation dilemmas as non-exclusive choice between *X* and *Y* across subjects) would be reflected, in primrose path experiments, as non-exclusive choice by individual subjects across trials. Indeed, in these experiments, subjects generally distribute choices non-exclusively across *X* and *Y*.

Complex as it is, Figure 1 is a highly simplified picture of real-world complexity. Lines *AD* and *BC* need not be parallel or straight or even monotonic (Rachlin 1997; 2000). High rates of consumption, harmful in one context, may be not harmful, or may be beneficial, in others. Nevertheless, the ambivalence represented by Figure 1 is real and captures everyday-life problems of self-control as well as everyday social dilemmas.

The labels in brackets in Figure 1 illustrate the application of this model to alcoholism. Let us say that point *A* represents a low rate of drinking (one or two glasses of wine with dinner). Dinner would be more enjoyable, however, with three glasses of wine and perhaps a cocktail beforehand (point *B*). But this much drinking every evening might interfere with sleep, or cause a hangover the next morning, or be slightly damaging to health. That is, notwithstanding the distinct pleasure of the extra drinking, the average value of the drinker's state over time (line *AC*) would be ever so slightly lower as the rate of drinking moves one unit to the right. Further increases in the number of drinks before, during, or after dinner (or instead of dinner) would always be immediately preferable to continuing at the lower rate but, if repeated day after day, would bring average value over time lower and lower (moving to the right on line *AC*). Eventually, at point *C*, drinking would serve only to prevent the misery of descent to point *D*. In other words, *positive reinforcement*, in going from point *A* to *B* by the social drinker having an extra drink, would have been replaced by *negative reinforcement* (avoidance of point *D*) in staying at point *C* by the alcoholic continuing to drink at a high rate.

The model of alcoholism as represented in Figure 1 is highly simplistic. Social drinking may be more valuable than teetotaling even in the long run. As noted above, lines *AD* and *BC* may not be parallel or even straight (see Herrnstein & Prelec 1992; Rachlin 1997; 2000, for discussion of more complex cases). Nevertheless, the model has suggested several methods of bringing behavior back from addiction (from point *C* to *A*). These include formation of temporally extended behavior patterns (Rachlin 1995a; 1995b), substitution of a "positive addiction" such as social activity for a negative addiction (Rachlin 1997), and manipulation of discriminative stimuli so as to signal changes in overall value (Heyman 1996; Rachlin 2000).

The existence of conflicting reinforcement at the level of particular acts versus that of patterns of acts makes it at least conceivable that a particular unreinforced act such as a mother's running into a burning building to save someone else's child may nevertheless be reinforced as part of a pattern of acts. A group of such acts, every one of them unreinforced (altruistic in the strict sense), may nevertheless form a highly reinforced – a maximally reinforced – pattern.

Just as group selection theory postulates more than one level of selection, so there may be more than one level of reinforcement – reinforcement of particular acts and reinforcement of groups, or patterns, of acts. Just as the behavior maximizing benefit to the individual may conflict with the behavior maximizing benefit to the group (which is what generates ambivalence in prisoner's dilemma situations), so a maximally reinforced act may conflict with a maximally reinforced pattern of acts. I have argued (Rachlin 1995a; 2000) that this latter type of conflict epitomizes many problems of self-control. I call this conflict *complex ambivalence,* as opposed to *simple ambivalence* in which one response leads to a smaller more immediate reward while an alternative response leads to a larger more delayed reward.[4]

Platt (1973) pointed out the relation between "temporal traps" and "social traps." Temporal traps are conflicts in the individual between smaller-sooner and larger-later rewards – situations of simple ambivalence. Social traps are conflicts between rewards beneficial to the individual and rewards beneficial to the group. Platt speculated that social traps could be understood as a subclass of temporal traps. But the correspondence between the two kinds of traps breaks down when attention is focused on particular choices (Dawes 1980; Messick & McClelland 1983). These authors point out that prisoner's dilemma problems, such as the one in my class demonstration, involve immediate conflicting consequences for the individual versus the group. The people in the audience are faced with only one momentary choice. Where is the temporal trap? The answer is that there is no temporal trap as long as temporal traps are limited to conditions of simple ambivalence. However, the correspondence of altruism and self-control is based not on simple ambivalence but on complex ambivalence; single choice exists in a vacuum. Assuming that their hypothetical choices are those they would make in a real situation, the members of my audience are making only one in a series of choices extending to their lives outside of the lecture hall. Messick and McClelland say (footnote 1, p. 110), "Obviously, a repeated Prisoner's Dilemma game requires a temporal component [that is, it can be explained in terms of self-control] but the opposition that characterizes a social trap exists without such repetition." This assertion highlights a crucial difference between teleological behaviorism and cognitive psychology. For the teleological behaviorist there can be no social trap without repetition. All prisoner's dilemmas are repeated. If a person were born yesterday, played one prisoner's dilemma game, cooperated in that

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

game, and then died today, it would be impossible to say whether the person's cooperation were truly altruistic or just an accident or really, in some other conceivable game, a defection.

4. Definitions of self-control and altruism

Moral philosophers, at least since Plato, have claimed that there is a relationship between self-control and altruism.[5] The fundamental issue addressed by ancient Greek philosophy was the relation between particular objects and abstract entities: abstract ideals for Plato; abstract categories for Aristotle (Rachlin 1994; Stout 1996). The problem of self-control in cases of complex ambivalence is a conflict between particular acts such as eating a caloric dessert, taking an alcoholic drink, or getting high on drugs, and abstract patterns of acts strung out in time such as living a healthy life, functioning in a family, or getting along with friend.

Neither self-control nor altruism is a class of particular movements, operants, or acts. Moreover, while self-control and altruism are both relative terms, depending on alternatives rejected as well as alternatives chosen, neither term refers to a particular choice independent of its context. For example, an alcoholic's particular choice of ginger ale over scotch and soda cannot be self-controlled unless it is embedded in a context of similar choices; if a person chooses scotch and soda 99 times to each choice of ginger ale, the choice of ginger ale is in no way self-controlled. The person might have been extremely thirsty at the moment when ginger ale was chosen, or might have been trying to hide his alcoholism at that moment, or might have made a mistake in his choice. The alcoholic's verbal claim that he intended to control his drinking at that moment would be taken as valid by the behaviorist only in the light of consistent future choices of ginger ales over scotch and sodas. And this criterion would hold regardless of the state of his nervous system, regardless of the activity or lack of activity of any internal mechanism. For the behaviorist, self-control as such has to lie wholly in choice behavior – but need not lie in any particular act of choice.

Similarly, no particular act is altruistic in itself – even a woman's running into a burning building and saving a child. If the woman were normally selfish we would look for other explanations (perhaps she was just trying to save her jewelry and only incidentally picked up the child). *A truly altruistic act is always part of a pattern of acts (highly valued by both the actor and the community) particular components of which are dispreferred by the actor to their immediate alternatives.* Altruistic patterns of acts are thus subsets of self-controlled patterns. The particular components of an altruistic pattern, like those of a self-controlled pattern, are less valuable to the actor than are their immediate alternatives; however, in the case of altruistic acts, they are also more valuable to the community than are their immediate alternatives.

Self-control may be defined more formally as follows: If two alternative activities are available, a relatively brief activity lasting t units of time, and a longer activity lasting T units of time, where $T = nt$ and n is a positive number greater than one, a self-control problem occurs when two conditions are satisfied:

1. The whole longer activity is preferred to n repetitions of the brief activity, and

2. The brief activity is preferred to a t-length fraction of the longer activity.

By "brief activity" and "long activity" I mean classes of activities perhaps not identical in topography but classified functionally, as Skinner (1938) defined operant class. For example, eating a steak dinner at a restaurant and drinking a "malted" at a lunch counter might be counted as repetitions of the same brief activity – eating high-calorie food. The long activity would be going through a period of time (a day, a month, a year) without eating high-calorie foods. The choice of the longer activity over a series of choices of the shorter activity is self-control.[6]

According to this definition, the "self" underlying self-control is not an internal entity, spiritual or mechanistic, containing a person's mental life (including a more or less powerful "will"). Such an entity would imply what Parfitt (1971) calls "personal continuity," a concept he believes we would be better off abandoning. Rather, the self is conceived as existing contingently in a series of overlapping temporal intervals during which behavior occurs in patterns (what Parfitt calls "contingent personal interactions"). People's "selves" would thus evolve and change over their lifetimes, as these patterns evolved and changed, as a function of social and non-social reinforcement.

Social cooperation situations may now be seen as a subcategory of self-control situations. A social cooperation situation exists when, in addition to Conditions 1 and 2:

3. A group benefits more when an individual member chooses a t-length fraction of the longer activity than it does when the individual chooses the brief activity.

An altruistic act is defined as a choice of the t-length fraction of the longer activity over the brief activity under Conditions 1, 2, and 3. The size of the group may range from only two people to the population of the world. The cost of the altruistic act may be a true cost, as when one anonymously donates money to charity, or an opportunity cost – the loss of the preferred brief alternative. Note that in this definition a particular altruistic act need not be reinforced, either presently or in the future. Reinforcement of altruism is obtained only when such acts are grouped in patterns that are, as a whole, intrinsically valuable. Thus, the woman's act of running into a burning building to save someone else's child is reinforced only insofar as it is part of a highly valued pattern. It may not itself ever be reinforced and may be punished by injury or death. If the woman died in the attempt, the act may still have been worth doing, since not doing it would have broken a highly valued pattern.[7]

This way of thinking about altruism and self-control may seem strange but it is not at all unusual. It is what Plato meant when he held Socrates' life (and death) to be both good (ethical) and happy. It is what many thinkers about ethics, before and since, have been saying. In 20th century psychology, the gestalt psychologists emphasized that the whole could be greater than the sum of its parts. They intended this maxim to apply to motivation or value as much as to perception (Lewin 1936). Consider listening to a symphony (assuming you enjoy this activity) on a CD that you just bought. Your enjoyment apparently begins when the music begins and ends when the music ends. Now suppose, after listening to the first 57 minutes of the symphony, you discover that the final three minutes of the 60-minute piece are missing from the CD. Is your enjoyment of the music just reduced by 3/60 of what it would have been if the whole symphony were played? Or is the breaking of the pattern so

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

costly that the missing three minutes ruins the whole experience? In my own case, the latter would be true. Readers who do not agree may imagine some other temporally extended activity that would be ruined for them by interruption late in the sequence.

The meaning of a single instrumental act can be found only in a context of other acts. Conditions 1, 2, and 3 place the act in such a context. For the cognitive psychologist, on the other hand, the meaning of a single act is to be found in the mechanism that immediately and efficiently caused the act. Thus, for the cognitive psychologist, a single act may be altruistic or not independent of other acts. Obviously, both cognitive and behavioral investigations need to be pursued. I am not saying that one is any more valuable or important than the other. But I do believe that it makes more sense to say that the behaviorist studies altruism itself while the cognitive psychologist studies the mechanisms behind it, than it does to say that the cognitive psychologist studies altruism itself while the behaviorist studies only its behavioral effects.

It seems clear that a person may be self-controlled without being altruistic. That is, Conditions 1 and 2 may obtain while Condition 3 does not. Although, given our strong social dependencies, there is usually some social benefit when a person stops drinking or smoking or overeating or gambling, such benefits are arguably incidental. The opposite question, whether a person may be altruistic without being self-controlled, however, is the one that concerns us here. This question is important because its answer determines whether people need a special mechanism for altruism, aside from whatever mechanism mediates self-control. Most demonstrations of altruistic behavior without egoistic incentives have focused on particular acts (Caporael et al. 1989). But it is not possible to determine that a separate altruism mechanism exists by the absence of reinforcement (immediate or delayed) of particular altruistic acts. The question is rather: Are there altruistic acts under Conditions 2 and 3 above where Condition 1 does not obtain? This is a difficult question to answer because Condition 1 does not specify the appropriate context (the longer activity, *T*) for a particular act. Is there *any* context (any relatively long-duration activity, *T*) in which a given altruistic act would also be a self-controlled act? I believe that it will always be possible to find such a context. This makes altruism a relative concept; in some contexts a given act will be altruistic and in some contexts, not. Where it is altruistic it will also be self-controlled (although the reverse may not be true).

The relativity of the concept of altruism should not be disturbing. First, it does not imply a moral relativism. Many Nazi soldiers behaved altruistically in the context of their military units but immorally in a larger context. Morality does not depend on altruism any more strictly than it depends on self-control. A moral code may approve of some kinds of altruism but disapprove of others, just as it may approve of some kinds of self-control and disapprove of others.

Secondly, whether an act is self-controlled or impulsive is no less contextually dependent than whether it is altruistic or selfish. Even a hungry rat rewarded by food for pressing a lever is to an extent controlling itself. The pattern of pressing the lever and eating takes longer (necessarily) than the act of pressing the lever alone. Pressing the lever, considered alone, is dispreferred to just sniffing in the corner of the cage; hence pressing the lever for food to be delivered within a fraction of a second is an instance of self-control. Correspondingly, even a slug may be said to exhibit self-control – on a microscopic level. At the other extreme, strict sobriety may be narrow relative to a still more complex pattern of social drinking.

There is a sense in which all acts (of choice) are selfish; the same sense in which all instrumental acts are reinforced and, for the economist, all behavior maximizes utility. These are assumptions of theory, or rather methods of procedure, not empirical findings. But this does not mean that selfishness is a meaningless concept (any more than reinforcement or utility maximization is). The sense in which an altruistic act is selfish (as part of an ultimately selfish pattern) differs from that in which a non-altruistic act is selfish. And this distinction is an empirical one.

Behavioral psychology has not been able to trace every particular act to a particular reinforcer – immediate or in the future. Organized patterns of acts occur despite the existence within them of unreinforced particular acts. What then reinforces the patterns? In psychology, theories of reinforcement based on "pleasure" or "need" or "drive" have not been able to explain particular acts. Such theories have proved to be circular – "pleasures," "needs," and "drives" proliferated about as fast as the behaviors they were supposed to explain. It is often not possible to use these concepts to predict behavior in one choice situation from behavior in another one. But Premack's (1965) wholly behavioral theory and the economic theories based on it (Rachlin et al. 1981) are predictive and noncircular. These theories use the choices under one set of behavioral contingencies or constraints to estimate the values of the alternatives (or the parameters of a utility function) and then use those values or parameters to predict choice under other sets of contingencies or constraints.

This method serves to explain choices among patterns of acts as well as particular acts. And, it answers the social-cooperation question, "Why is friendship rewarding?" as well as the self-control question, "Why is sobriety rewarding?" The answer in both cases, for the behavioral psychologist, is that in a choice test between each of these patterns *as a whole* and their respective alternative patterns *as a whole*, friendship would (at least in some cases) be chosen over loneliness, and sobriety would (at least in some cases) be chosen over drunkenness.[8]

5. Commitment

No amount of calculation by the mother who runs into a burning building to save someone else's child will bring the benefits-minus-risks of this activity *considered by itself* into positive territory. But over a *series* of actions, a *series* of opportunities to sacrifice her own benefit for the benefit of others, the weightings may change. As we have seen (Fig. 1) social and individual decisions may individually be completely negative, their only value appearing when they are grouped.[9] The problem is that life ordinarily faces us not with groups of decisions but with particular decisions that must be made. It is up to us to group decisions together, and we do this by means of various commitment devices – contracts, agreements, buying tickets to a series of concerts or plays, joining a health club, and so on.

These commitments may work by instituting some pun-

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

ishment (such as loss of money or social support) should we fail to carry them through. Green and Rachlin (1996) have shown that pigeons prefer, A: a future choice between (1) a small, immediate reward followed by punishment and (2) a larger, delayed reward to, B: the same future pair of alternatives but without the punishment. Only by the present choice of the future pair of alternatives involving punishment will they avoid being tempted later by the smaller immediate reward and obtain the larger reward that they prefer at the present time. Another kind of commitment shown by pigeons (Siegel & Rachlin 1996) is "soft commitment." At an earlier time the pigeon begins a pattern of behavior, such as rapidly pecking a fixed number of times on a lit button. This pattern is difficult for the pigeon to interrupt. Then, in the midst of this pattern, the tempting alternative (the smaller, immediate reward) is presented. Only by continuing and completing the previously begun pattern of behavior will the larger reward be obtained. By beginning and continuing the pattern the pigeon avoids the temptation and obtains the larger reward. The further along the pigeon is into the pattern, the more likely it is that the tempting small reward will be avoided.

In a primrose-path experiment (italicized labels of Fig. 1) in my laboratory (Kudadjie-Gyamfi & Rachlin 1996) human subjects chose the self-control option (*Y*) more when choices were clustered in threes (patterned) than when they were evenly spaced. Within a group of three choices, the probability of self-control on the first choice was high but, given self-control on the first choice, the conditional probability of self-control on the second choice was higher and, given self-control on the first two choices, the probability of self-control on the third choice was higher still. Similarly, in a repeated prisoner's dilemma situation, playing against tit-for-tat (a strategy that mimicked, on a given trial, the subject's choice to cooperate or defect in the previous trial), human subjects cooperated more when trials were clustered in fours than when they were evenly spaced out; moreover, as in the self-control experiment, conditional probability of cooperation increased as the sequence progressed (Brown 2000).

Soft commitment with pigeons is a model, on a narrow temporal scale, for successful self-control by humans, on a much wider temporal scale (Rachlin 2000). The alcoholic, for example, resolves to stop drinking, and refuses one drink. At that point he is vulnerable to the offer of another drink. But if he refuses 10 drinks he is less vulnerable and if he refuses 100 drinks he is still less vulnerable. He refuses the later drinks not because their value is reduced (their value is actually enhanced as deprivation increases) but because he has already begun a pattern of refusal that involves some cost to break. As he repeatedly refuses drinks (climbs up line DA in Fig. 1) the long term rewards that sobriety entails – better health, social support, better job performance – grow apace.

In experiments on repeated prisoner's dilemmas, some subjects cooperate and continue to cooperate regardless of whether other subjects cooperate with them (Brann & Foddy 1988). These people may be said to cooperate out of a sense of moral duty or for ethical reasons or because they are more altruistic than others. But these sorts of explanations do not say why such people behave as they do. To understand their behavior, the laboratory experiment has to be seen not as an isolated situation but in the context of everyday life. Many experimental subjects are willing and able to separate decisions made in a psychology experiment from those they make in everyday life. But others are not able or not willing to do so. They have decided to cooperate in life and continue to do so in the experiment, not necessarily because of some innate tendency to be altruistic, but because altruism is generally valuable and they would not act altruistically if they made decisions on a case-by-case basis. The experiment is merely one case, one situation out of many in their lives. Moral duty, ethical concerns, and altruism are apt descriptions of their behavior. But these qualities do not come from nowhere. They are highly valued patterns of behavior – just as moderation in eating, moderation in drinking, and moderation in sexual activity are highly valued patterns.

6. Reinforcement and punishment in The Prisoner's Dilemma

Current discussions of altruism and selfishness in philosophy, biology, economics, and psychology are generally united by reference to strategies of play in prisoner's dilemma situations. The present analysis does not deny the interest or importance of strategies. Rather, as patterns of behavior, it sees them as crucial. The difference between the present behavioral analysis and cognitive analyses is that, in determining what underlies a strategy, the behaviorist looks for contingencies of reinforcement and punishment rather than internal mechanisms. Thus, it is important to show that the prisoner's dilemma incorporates reinforcement and punishment contingencies and that prisoner's-dilemma behavior is sensitive to those contingencies.

Consider the contingencies of the two-person prisoner's dilemma diagramed in Figure 2a. If both players cooperate, each gets 5 points (convertible to money at the experiment's end); if both defect, each gets 2 points; if one cooperates

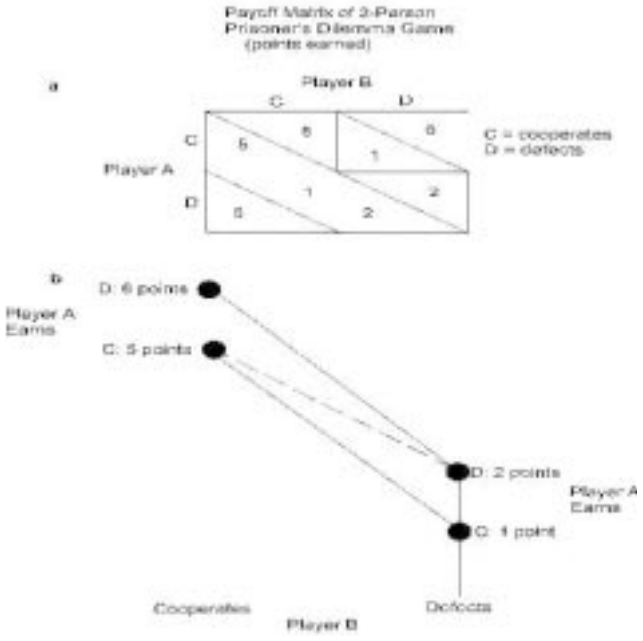


Figure 2. (a) Payoff matrix of two-person prisoner's dilemma game. (b) Same game. Player A's earnings for cooperation (lower black dot) and defection (upper black dot) as a function of Player B's choice.

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

while the other defects, the cooperator gets 1 point while the defector gets 6 points. Figure 2b diagrams the game in a corresponding way to Figure 1, revealing the ambivalence. As in Figure 1, defection results in a higher immediate reward and a lower long-run reward while cooperation results in the reverse. Regardless of the other player's choice, it is always immediately better to defect than to cooperate; if the other player has cooperated then a player will gain 6 points by defecting and 5 points by cooperating; if the other player has defected then a player will gain 2 points by defecting and only 1 point by cooperating. If communication between players is against the rules, if the game could be played only once (and no similar cooperative tasks were ever expected to be undertaken with the other player), then the motive to defect should predominate. However, if there was some way to get the other player to cooperate, then whatever it takes to do this should predominate over defection because the gain from the right to the left vertical line in Figure 2b averages 4 points while the gain from the lower to the upper line (from cooperation to defection) averages 1 point. The best set of circumstances would be to defect while the other player cooperates, earning 6 points. This is an unlikely scenario since the other player would then earn only 1 point. However, if communication was within the rules, it would be possible to compromise by agreeing to mutual cooperation, earning 5 points each (the highest pooled score). Or, if the game were to be played many times, it would be possible to reinforce the other player's cooperation by cooperating, and to punish the other player's defection by defecting. This strategy is called "tit-for-tat." The dashed line shows average points gained in repeated trials against tit-for-tat with a distribution of choices proportional to the distance between the vertical lines. For example, alternation of cooperation and defection (halfway between the vertical lines) yields 6 points and 1 point alternately for an average of 3.5 points per trial against tit-for-tat. The highest point on the dashed line (hence the best strategy against tit-for-tat) is to cooperate on all trials. Tit-for-tat has indeed been highly effective in generating cooperation and maximizing pooled scores in several situations: computer simulations of prisoner's dilemma games (Axelrod 1997); 2-person games with human subjects (Brown & Rachlin 1999; Rapoport & Chammah 1965; Silverstein et al. 1998); with a single subject playing against a computer programmed to play tit-for-tat (Komorita & Parks 1994).

The crucial variable influencing cooperation in two-person games seems to be reciprocation (Komorita & Parks 1994; Silverstein et al. 1998). This is also true in games with more than two players such as illustrated in Figure 1 (Komorita et al. 1993). The tit-for-tat strategy imposes a strict reciprocation and thus engenders cooperation. Prior communication enhances reciprocation and thus has the same effect. On the other hand, when reciprocation is low or nonexistent, as when the other player plays randomly or always cooperates or always defects, cooperation deteriorates (Silverstein et al. 1998). Baker and Rachlin (2001) found that a player's probability of cooperation in a two-person prisoner's dilemma game varied directly with the other player's probability of reciprocation.

7. Context

As Tversky and Khaneman (1981) showed, context, or "framing," strongly influences probabilistic choice behavior. Context is likewise a strong determinant of self-control. Heyman (1996) cites a study by Robins (1974) of American soldiers who became addicted to heroin in Vietnam. The majority of these addicts easily gave up their addiction when they came home to a different environment. Heyman argues that the boundary line separating local from non-local events (the duration of the chosen activity) may vary over a wide range (depending on the salience and relevance of environmental stimuli), thereby explaining how humans and nonhumans may act impulsively in one situation and self-controlled in another. A second experiment by Baker and Rachlin (2002) demonstrates a similarly strong influence of context in a social cooperation experiment with human subjects.

Tit-for-tat is a *teaching* strategy. A computer, playing tit-for-tat against a player, invariably follows the player's cooperation by cooperating on the next trial and invariably follows the player's defection by defecting on the next trial. Since the computer's cooperation is much more valuable to the player than its defection, the computer's cooperation reinforces the player's cooperation and its defection punishes the player's defection. Thus, the computer "teaches" the player to cooperate.

Another strategy that has been successful in computer tournaments (dominating tit-for-tat) is called Pavlov (Fudenberg & Maskin 1990; Nowak & Sigmund 1993). Pavlov is a *learning* strategy. Using Pavlov, the computer's choice on the present trial, whether cooperation or defection, is repeated in the next trial if the player cooperates and changed in the next trial if the player defects. Against tit-for-tat, the player cannot successfully punish the computer's defection; the computer would respond to defection by defecting itself. Using Pavlov, however, the computer would respond to defection by changing its choice in the next trial: if it had defected, it would now cooperate; if it had cooperated, it would now defect. The computer using Pavlov would respond to cooperation by repeating its choice in the next trial; if it had defected, it would defect again; if it had cooperated, it would cooperate again. That is, the computer would behave as if its choice were reinforced by the player's cooperation and punished by the player's defection. Thus, the computer, playing Pavlov, "learns" from the player.

In this experiment, four groups of subjects (Stony Brook undergraduates) played 100 trials of a prisoner's dilemma game. Against each subject in two groups, the computer played a modified form of tit-for-tat. Against each subject in the other two groups, the computer played a modified form of Pavlov.[10] One of the tit-for-tat groups and one of the Pavlov groups saw a spinner on the computer screen and were correctly informed that the computer's responses were determined by that spinner. The other two groups believed that they were playing the game against another player rather than against a computer. They did not see a spinner but they did see the "other player's" reward matrix (and reward presumably received) as well as their own.[11]

The results of the experiment are shown in Figure 3. The context of the game – whether or not the subjects were led to believe that they were playing against another subject – had a strong effect on their behavior, but the context effect was opposite for the two computer strategies. When subjects believed that they were playing against a computer, they cooperated more against tit-for-tat (where the computer reinforced and punished the players' cooperation and defection) than they did against Pavlov (where the com-

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

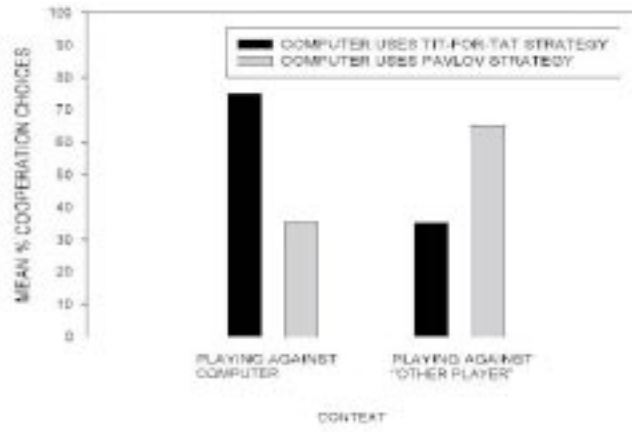


Figure 3. Results of Baker and Rachlin's (in press) experiment. Average of last 15 of 100 trials.

puter's choices were reinforced and punished by the players' cooperation and defection). On the other hand, when subjects believed that they were playing against a human being, they cooperated more against Pavlov than against tit-for-tat. This result may be attributed to the fact that subjects' histories of interacting with machines (unlikely to be responsive to reinforcement and punishment) differed from their histories of interacting with other people (more likely to be responsive). When the relatively global histories matched the relatively local set of contingencies (the computer's strategies) subjects cooperated; when the global histories contradicted local contingencies they defected. (In all cases, however, under the most narrowly local contingencies, defection was immediately reinforced.) Choice in prisoner's dilemma situations, therefore, like choice in self-control situations, may be understood in terms of global as well as local reinforcement.

Taken together with the previously discussed experiments of Kudadjie-Gyamfi and Rachlin (1996) and Brown (2000), in which patterning choices over time increased human subjects' self-control and prisoner's-dilemma cooperation, the experiment described above shows that, at least in laboratory studies, self-control and social cooperation are similarly responsive to reinforcement contingencies and similarly sensitive to context.

Laboratory models, however, are necessarily diminished representations of everyday-life processes. The reinforcers in all of these experiments – points convertible to money – were extrinsic to the subjects' choices. If, as argued here, the reinforcers of real-life self-controlled and altruistic behavior are intrinsic in the patterns of those behaviors, and if those patterns are extended over long durations – months and years, – real-life rewards will never be duplicated in a 30-minute laboratory experiment.

The experiment described above partially gets around this limitation by varying verbal instructions so as to bring the brief laboratory experiment into differing long-term, real-life contexts. Moreover, an economic extension of Premack's conception of reinforcement (Rachlin et al. 1981) sees all reinforcement as intrinsic (even that of a rat's lever press reinforced by food; the rat is seen as choosing the pattern of lever pressing plus eating over not pressing lever plus not eating).

Nevertheless, there remains a vast difference in scale between laboratory experiments and real life. The point of the experiments is to show that, on a small scale, self-control and altruism are sensitive to reinforcement and punishment. In the case of self-control there is ample evidence that large-scale, real-life behavior is similarly sensitive (Bickel & Vuchinich 2000). If, as is argued here, there is no essential difference between self-control and altruism, the same behavioral laboratory studies that have proved useful in developing real-life self-control techniques may be equally useful in developing real-life altruistic behavior.

8. Can altruism be explained without reinforcement?

Does this way of thinking put more weight on reinforcement than it can bear? Can the job be done entirely by internal mechanisms with reinforcement playing no part whatsoever? The issue is this: There are some particular acts, especially by humans, that we normally classify as done through a sense of altruism, of duty, of principle. No biologist claims that a separate inherited mechanism exists for each of the infinitude of possible acts that fall within these categories. To explain such actions as inherited, the biologist must hypothesize the existence of a general mechanism for altruism which is somehow aroused by situations such as the game I play with my audiences illustrated in Figure 1 (bold labels). It seems to me that the postulation of such a mechanism as inherited – like blue or brown eyes – puts far too heavy a load on inheritance; we have no idea how such a mechanism could work.

On the other hand, it is generally agreed that self-control may be taught at some level even to nonhumans. The crucial issue then is whether or not altruism is a subcategory of self-control. If it is, there is no need to postulate an innate altruistic mechanism; the job can be done by whatever mechanism we use to learn self-control – an innate mechanism to be sure, but an innate learning mechanism.

This is hardly an original idea. Plato and Aristotle both claimed that self-control and altruism were related concepts. The experiments described in this article illustrate the correspondence. However, perhaps the argument is ultimately not empirical. It rests on two assumptions: (1) Habitual altruism is a happier mode of existence than habitual selfishness and (2) Particular altruistic acts (together with their consequences) are less pleasurable (even for saints) than particular selfish acts (together with their consequences). If you accept both of these propositions, altruism must be seen as a kind of self-control.

9. Can altruism be explained wholly in terms of extrinsic reinforcement?

How are patterns of behavior learned and how are they maintained? Consider the following set of cases. Four soldiers are ordered to advance on the enemy. The first and second advance; the third and fourth do not. Of the two who advance, the first is just obeying orders; he advances because he fears the consequences of disobedience more than he fears the enemy. The second is not just obeying orders; he advances because he believes it is his patriotic duty to advance. Of the two who do not advance, the third soldier remains in his foxhole out of fear of the enemy; he weighs the aversive consequences of disobeying orders less than

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

the aversive consequences of advancing. The fourth soldier does not advance because he believes that the orders are immoral.

No one, neither the biologist, the cognitivist, the Skinnerian behaviorist, nor the teleological behaviorist, denies that there are important differences between the two soldiers who advance and between the two soldiers who do not advance. But the biologist and cognitivist alike see all the differences in thought, feeling, moral sentiment of the soldiers, as contemporary with their current behavior. Behaviorists do not disagree that internal differences exist but their focus is rather on noncontemporary events; the Skinnerian behaviorist is concerned to discover crucial differences in the soldiers' *extrinsic* reinforcement histories. The teleological behaviorist is concerned to discover the patterns of behavior of which each soldier's present act forms a part (*intrinsic* reinforcement). Note, however, that even the concept of extrinsic reinforcement must rely at some point on intrinsic reinforcement. According to Premack's theory, for example, eating reinforces lever pressing because eating is (intrinsically) of high value and lever pressing is (intrinsically) of lower value. I am claiming here that an abstract pattern of behavior may be (intrinsically) of high value while the sum of the values of its particular components are of (intrinsically) lower value. Value, in either case, would be determined by a choice test.

Let us first consider extrinsic reinforcement. By careful selection, with humans, it is possible to reinforce members of a set of particular acts belonging to a wide or abstractly defined class of acts (a rule) so that particular acts that have never been reinforced, but that obey the rule, are performed. That is, humans are able to generalize across instances of complex rules and, with simple rules, nonhumans are also able to do so. Behavior thus learned is said to be *rule-governed.* Imitation (of certain people) and following orders (in certain circumstances) are two such kinds of rules. There is no space here to discuss the several techniques developed for generating rule-governed behavior with extrinsic reinforcement (see Hayes 1989, for a collection of articles on the subject), nor to discuss current disputes about whether language precedes complex rule-following or whether rule-following precedes language (Sidman 1997).

The behavior of the first soldier, who advances because he fears the consequences of disobeying orders more than he fears the enemy, and that of the third soldier, who fails to advance because he fears the enemy more than the consequences of disobeying orders, may be explained in terms of conflicting rules. Regardless of the complexity of the relation between the consequences of the present act and those of past acts, it is the weighting of the extrinsic consequences of the present act (the magnitudes, probabilities, and delays of enemy fire versus those of punishment for disobedience) that determines the behavior of these two soldiers.

Moreover, it may be possible to account for the *initial learning* of ethical rules and principles, such as those that govern the altruistic behavior of the second and fourth soldiers, in terms of extrinsic social reinforcement at home or school or church. But extrinsic reinforcement cannot account for the *maintenance* of altruistic behavior. An altruistic act may never be reinforced. The second and fourth soldiers (as well as the woman who runs into the burning building to save someone else's child) are as capable of weighing the immediate consequences of their acts as are the first and third soldiers. But those consequences are ignored by these two soldiers. The second and fourth soldiers, both of whose behavior has been brought under the control of highly abstract principles (we are assuming), are surely capable of discriminating between the extrinsic consequences of their present acts and the extrinsic social approval or disapproval of their past behavior at home, school or church where the principles were learned. A person capable of bringing his or her behavior into conformance with an abstract principle by means of extrinsic reinforcement, and of transferring the application of that rule across situations, could not fail to discriminate the present context (where social approval is dwarfed by the possibility of death) from situations where the rule-governance may have been initially learned. Yet the altruistic act is performed anyway.

Such acts must be maintained not by extrinsic reinforcement but by intrinsic reinforcement. The patterns of those acts (patriotic, ethical, altruistic), perhaps supported during their formation by a scaffold of extrinsic reinforcement, must be highly valuable in themselves. If they depended on extrinsic reinforcement for maintenance they would not be maintained.

In Premack's (1965) terms, valuable patterns would be chosen if offered as whole patterns in a free choice situation. In cases such as the patriotic and ethical soldiers and the woman saving a child, imagine a giant concurrent-chain schedule with years-long terminal link alternatives: heroism versus timidity, reverence for life versus toleration of killing, kindness versus cruelty. Because of their intrinsic value the chosen patterns are final causes of their component acts and may themselves be effects of still wider final causes: a coherent concept of self; living a happier life, living a better life.

Most of us would indeed choose to be heroes rather than cowards, to revere life rather than to kill, to be kind rather than cruel. We realize that the former alternatives of each pair are actually patterns of happy lives and the latter, of unhappy lives. But these alternatives are rarely offered to us as wholes. Rather, we are faced with a series of particular choices with outcomes of limited temporal extent. The altruists among us, however, have chosen such more extended patterns as wholes; they are the patterns most of us would choose if we could choose them as wholes. But to do this we would need to evaluate particular alternatives not by their particular consequences but rather by whether or not they fit into the larger patterns. This of course is a problem of self-control.

10. Conclusions

Some particular altruistic acts are profitable some of the time. Giving to charity is often observed and frequently rewarded by society. But patterns of behavior may be maintained without extrinsic rewards. For example, on a relatively small scale, activities such as solving jigsaw or crossword puzzles are valuable in themselves. People, like me, who like to do crossword puzzles, find value in the whole act of doing the puzzle. When I sit down on a Sunday morning to do the puzzle I am not beginning a laborious act that will be rewarded only when it is completed. Yet, despite the lack of extrinsic and intrinsic reward for putting

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

in that last particular letter, completing the puzzle is, for me, a necessary part of its value. Like listening to symphonies, the pattern is valuable only as a whole. Extrinsic rewards may initially put together the elements of these patterns but the patterns, once formed, are maintained by their intrinsic value. The cost of breaking the pattern is the loss of this value – even that of the parts already performed. On an infinitely larger scale, living a good life is such a pattern. This is why the woman runs into the burning building to save someone else's child without stopping to calculate the cost of this particular act, why Socrates chose to die rather than violate the sentence of the Athenian court.

It is not possible to tease apart the individual and social benefits of such acts. High degrees of altruism are infrequent, not because most people lack an internal altruism mechanism, not because they are selected by evolution to be egoists rather than altruists, but because of the highly abstract nature of the valuable patterns. The relation between particular acts of altruism and the intrinsic reward of the pattern is vague and indistinct. Altruism for most of us (like sobriety for the alcoholic) is not profitable and would not be chosen considering only its case-by-case, extrinsic reinforcement. Consequently, the way for most of us to profit from altruism (and the way for an alcoholic to profit from sobriety) is to pattern our behavior abstractly – to choose to be an altruistic (or a sober) person. But in order to pattern our behavior in this way (and reap the rewards for so doing) we must forego making decisions on a case-by-case basis. Once we abandon case-by-case decisions, there will come times in choosing between selfishness and altruism when we will be altruistic even at the risk of death.

ACKNOWLEDGMENTS

The research reported in this article and the preparation of the article were supported by grants from the National Institute of Mental Health and the National Institute on Drug Abuse. Some sections of the article are rewritten versions of sections of the author's book, *The Science of Self-Control,* published by Harvard University Press (Rachlin 2000).

NOTES

1. These are very wide conceptions of selfishness. Usually, by "selfishness," we mean explicit rejection of a clearly altruistic alternative; so the word has a socially negative connotation. However, in popular explanations of biology, "selfishness" has lost its negative sense. It just stands for survival value (as in "selfish gene"). Similarly, I use the term here to stand for reinforcement value.

2. What counts seems to be how the problem is presented – whether I emphasize the group or the individual benefit – rather than who the players are (Italian economists, Japanese psychologists, Stony Brook undergraduates, and so forth) or whether the amounts of money won are large and hypothetical or small and real.

3. Other versions of the primrose path manipulate delays rather than amounts (with inverse contingencies). In some experiments subjects are given more or less explicit instructions about the contingencies in effect. In others, the base number of trials determining *N* (rule number 3) is varied. In still others, trials are grouped in temporal patterns. These manipulations have systematic effects on the proportion of *X*s and *Y*s chosen (over a typical session of about 100 trials), but none results in exclusive choice of *X* or *Y,* showing that the contingencies retain their essential ambivalence.

4. The social prisoner's dilemma, in which a single person's interests conflict with the common interests of a group, is analogous to a single person's intertemporal dilemma, in which the person's interests over a narrow time range conflict with the common interests of that same person over a wide time range. Ainslie (1992) pointed out that the prisoner's dilemma among groups of individuals corresponds to that within an individual at different times. The difference between Ainslie's view of self-control and mine is my conception of common interests reinforcing behavioral patterns (analogous to group selection) versus Ainslie's conception of internal bargaining among a person's temporally distant interests. Underlying this is a difference in our conceptions of simple versus complex ambivalence. Ainslie believes that complex ambivalence – where abstract rewards such as good health reinforce behavioral patterns such as daily exercise – may be reduced to the sum of discounted values of particular rewards acting on each particular act of exercise. That is, Ainslie believes that complex ambivalence may be reduced to multiple cases of simple ambivalence. I believe that complex and simple ambivalence are essentially different. Where simple ambivalence opposes larger but more delayed rewards to smaller but less delayed rewards, complex ambivalence opposes larger but more abstract (and temporally extended) rewards to smaller, particular rewards.

5. And many times since. Ainslie (1992), Platt (1973), and Schelling (1971) have recently stressed this correspondence.

6. It is sometimes supposed that in a perfect world there would be no conflict between immediate desires and long-term values. The image of a natural human being living a natural life has this sort of framework – a place where our immediate desires are in harmony with our long-term best interests. But, as Plato pointed out (*Philebos,* 21c), life in such a world would be the life of a slug. In such a world we would have no need to behave in conformance with more abstract environmental contingencies; therefore we would have no ability to do so.

7. As previously noted, however, people often ignore valuable long-term patterns and focus on particular present costs and benefits. In economic terms, this implies that you need to be very careful in determining which previously incurred costs are really "sunk costs" and which are investments that, if pursued (at a present additional cost), may still pay off.

8. This is as far as the behavioral psychologist can go. For the evolutionary biologist, the answer to, "Why is this pattern valuable?" is that it has contributed to survival in the past. I am not arguing that the behavioral psychologist's answer is better than the evolutionary biologist's answer but rather that a correspondence between self-control and social-cooperation is no less consistent with an evolutionary biological approach to behavior than it is with a teleological behavioral approach.

9. As the Gestalt psychologists pointed out, we perceive patterns (like melodies) directly rather than as the sum of their parts. Similarly, the value of a pattern (like the enjoyment of listening to a melody) may be far greater than the sum of the values of its parts (the enjoyment of listening to particular notes).

10. The game was modified to make the computer's responses probabilistic rather than all-or-none. When a strategy would ordinarily dictate cooperation, the computer increased its probability of cooperation by .25 (and decreased its probability of defection by .25) but keeping probability between 0 and 1. When a strategy would ordinarily dictate defection, the computer increased its probability of defection by .25 (and decreased its probability of cooperation by .25).

11. There were two other groups whose results are not presented here. Those groups did not see a spinner on the computer screen but neither did they see another reward matrix and they were not led to believe that they were playing against another subject.

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

Open Peer Commentary

Commentary submitted by the qualified professional readership of this journal will be considered for publication in a later issue as Continuing Commentary on this article. Integrative overviews and syntheses are especially encouraged.

Altruism is a primary impulse, not a discipline

George Ainslie[a] and Nick Haslam[b]

[a]*Department of Veterans Affairs Medical Center, Coatesville, PA 19320;* [b]*Department of Psychology, University of Melbourne, Parkville, VIC 3010, Australia.* **www.picoeconomics.com** **George.Ainslie@med.va.gov.**

Abstract: Intertemporal bargaining theory based on the hyperbolic discounting of expected rewards accounts for how choosing in categories increases self-control, without postulating, as Rachlin does, the additional rewardingness of patterns per se. However, altruism does not seem to be based on self-control, but on the primary rewardingness of vicarious experience. We describe a mechanism that integrates vicarious experience with other goods of limited availability.

Utility theory is frequently read as declaring altruism to be irrational. Rachlin offers one of many current rebuttals of this counterintuitive conclusion (see, e.g., Batson & Shaw 1991; Field 2001). His argument is that altruism is a kind of self-control, overriding one's current impulse for the sake of a longer-range good. He depicts the mechanism as learning to see choices not as isolated instances but as part of overall patterns. We agree that such overall interpretation is a key mechanism of self-control; but Rachlin's mechanism entails the unnecessary assumption that the reward for a pattern of self-control is greater than the sum of rewards for the choices that make up the pattern. Furthermore, we do not agree that altruism is motivated mainly by the incentives that exist for self-control.

Rachlin confronts two conceptual problems, both of which can be solved more specifically by assuming only hyperbolic discounting of delayed rewards, the additivity of discounted sequential rewards, and the dependence of total available appetite on rate of satiation. The first problem is how patterns of choices come to be preferred oppositely from individual members of the pattern. The second is how an organism comes to be rewarded by another organism's experience.

1. Rachlin's Note 4 concisely characterizes the differences between his and Ainslie's theories of "complex ambivalence." In Rachlin's model, the value of a pattern of being sober, say, or being altruistic, is greatly reduced by being intoxicated or selfish just today, just as the value of a symphony plummets if notes are taken out; but the lapse does not make the remaining pattern unavailable. In Ainslie's intertemporal bargaining model the lapse is attractive but is avoided, when it is avoided, because of the risk that it will break off the pattern. The advantage of the bargaining theory is that it accommodates the widely reported urge to duck out of a pattern as well as the urge to maintain it, without postulating more than the rewards literally available in each choice: When a person sees each choice in a category as a test case for her continuing cooperation in an intertemporal prisoner's dilemma, she will face both an incentive to distinguish the present choice from the pattern, that is, to rationalize a defection vis-à-vis her future selves, and a growing incentive to preserve her expectation of future prudence, that is, to cooperate with these selves. Rachlin's alcoholic is less vulnerable to temptation after refusing 100 drinks than after refusing one, not because he sees a pattern of sobriety as any more desirable than before, but because it has become more believable that he will attain it if he does not slip this time. To get the pattern-keeping effect, he must still see each refusal of alcohol as necessary to this believability; if someone offers him a drink under circumstances that he does not expect to reduce this believability – a really good rationalization, say, like coercion or a rare occasion – he is apt to welcome it.

Hyperbolic discounting and its implication of intertemporal bargaining would be a good theory of the incentives for pattern-following even without independent evidence that it exists. In fact, there is not only overwhelming evidence that all organisms discount rewards in single-shot choices hyperbolically (e.g., Kirby 1997), but good evidence that they reverse preferences from smaller-earlier to larger-later rewards when they choose between a whole sequence of pairs at once, a phenomenon that would not occur with conventional (exponential) discounting (Ainslie & Monterosso, in press; Kirby & Guastello 2001). Softer evidence that willpower is based on the perception of current choices as test cases predicting such sequences comes from thought experiments (Ainslie 2001, pp. 117–38) and experimental bargaining analogs (Monterosso et al. 2002).

2. Certainly choosing according to principle increases self-control. The assertion that the same mechanism overcomes selfish motives by creating altruistic ones is more tenuous.

Rachlin does not make any attempt to say why altruistic patterns of choice should be rewarding when their components are not, merely assuming that "habitual altruism is a happier mode of existence than habitual selfishness" (target article, sect. 8, para. 3). But the same appeal to common experience argues that the vicarious feeling of other people's emotions is a primary good.

From early childhood on we spontaneously put ourselves in other people's shoes, a phenomenon that has been demonstrated in nonhuman animals and for which neurophysiological substrates have been found (Preston & de Waal 2002).

The vicarious reward that empathy supplies is often so compelling that people make efforts to discipline it – to rein in altruistic tendencies, for example, suppressing sympathy toward those whom one expects to request costly help later (Shaw et al. 1994). It is quite believable that the woman who saves the stranger's child from a burning building may do so impulsively, in violation of a perceived duty to her own family, because she cannot tolerate imagining a burning baby or an anguished mother. Furthermore, if altruism were dependent on farsightedness learned, like civility, over the person's lifetime, then altruism would be unknown among young children and would increase with age. The evidence contradicts both of these implications (e.g., Fiske 1991; Frank et al. 1993). Empathy is a robust and early-developing process that underpins prosocial behavior (altruism) as well as antisocial behavior (retribution, gloating).

Two issues have apparently kept utility theorists from accepting vicarious experience as a primary good: the lack of a physical need for stimuli from other people in order to have a positive emotional response to them, and, conversely, the difficulty of avoiding negative emotional responses to information about other people's painful experiences. Conventional theory does not tell us why we want to know that others are happy, or why we allow ourselves to be moved when they are not. Folk psychology depicts our responses as unmotivated, perhaps classically conditioned. However, the fact of hyperbolic discounting predicts that I will be impatient in consuming a reward that is at my free disposal, as emotional reward apparently is. The consequent premature satiation should reduce emotional patterns to the quality of daydreams, unless I learn to cue my emotional behaviors with occasions that are outside of my control and adequately surprising. Vicarious experience represents a rich source of such occasions, which may thus come to govern my emotions almost as if they were stimuli for reflexes. The same hyperbolic curves may also cause vivid aversive experiences to seduce my attention, just as over a slower time course these curves may cause addictive substances to lure me into unrewarding choices. Thus, even anguish need not be seen as either imposed by a process like classical conditioning or accepted through the kind of discipline Rachlin proposes, but rather as a good that can compete in the free market of choice (see Ainslie 1995; 2001, pp. 161–86).

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

Behaviorism and altruistic acts

J. McKenzie Alexander

Department of Philosophy, Logic and Scientific Method, London School of Economics and Political Science, London WC2A 2AE, United Kingdom.
jalex@lse.ac.uk

Abstract: Rachlin's idea that altruism, like self-control, is a valuable, temporally extended pattern of behavior, suggests one way of addressing common problems in developing a rational choice explanation of individual altruistic behavior. However, the form of Rachlin's explicitly behaviorist account of altruistic acts suffers from two faults, one of which questions the feasibility of his particular behaviorist analysis.

Rational choice explanations of altruistic behavior tend to flounder when they try to reconcile individual maximization of expected utility with the fact that altruistic behavior confers a lower utility than other available choices. Incorporating other-regarding interest into the conception of self-interest makes this reconciliation easier to achieve, but at the expense of a notion of self-interest that can reasonably be attributed to most individuals. In this light, Rachlin's explanatory strategy of accounting for altruistic acts by seeing them as particular instances of a highly valued pattern of behavior – the value of which overrides the value of nonaltruistic acts – seems to be the right way to proceed. However, as promising as Rachlin's strategy is, I find the extent to which he adopts a behaviorist position troubling. Indeed, his overarching idea that altruistic acts belong to a general pattern of conduct does not commit him to behaviorism, and his final suggestions regarding how patterns of behavior are maintained in fact undermine an explicitly behaviorist account.

To begin, it seems that Rachlin's definition allows for inconsistent classifications of acts. Consider the following example: A father of two children enjoys spending a short time with them at home, but because of their young age prefers being at work all day to being at home all day. Conditions 1 and 2 are thereby satisfied. Then, one day the father contemplates staying at work an extra hour before going home and, reluctantly, decides to stay at work. Is this act altruistic? Because Condition 3 leaves the identity of the group benefitted by the father's choice a free variable, whether the act is altruistic or not depends on which group we select. Because the company benefits from the father's choice, by Rachlin's definition the act is altruistic. On the other hand, if we consider the family and see that the father's choice does not benefit it, the act can simultaneously be seen as not altruistic. Note that this is a different point from saying that altruistic acts are context-dependent, a point Rachlin accepts, believing that one can always find contexts that render acts altruistic.[1] While I readily admit that different act-tokens of the same act-type may be differentially classified as altruistic or not depending on the context in which the particular act-token occurs, it does not make sense for the same act-token to be identified as either altruistic or not on the basis of how *we* carve up the world into groups that are or are not affected beneficially by the act.

More importantly, though, I find that in Rachlin's definition of an altruistic act a tension obtains between the behaviorist account of altruism as choices violating individual preference and the underlying behaviorist account of preference presupposed by Conditions 1 and 2. Consider whether a coherent behaviorist gloss may be given to Conditions 1 to 3. Condition 3 poses no difficulty because choice can be defined operationally, but what about the references to individual *preferences* in Conditions 1 and 2? Can these be given a suitable behaviorist interpretation? One behaviorist response might adopt the traditional economist view, which says individual preferences are revealed through choice. However, this explanation of what it means to talk of individual preferences in Conditions 1 and 2 proves difficult to reconcile with Condition 3: if an individual chooses a t-length fraction of the longer activity over the brief activity (i.e., Condition 3 obtains), in what sense can one say that the individual *prefers* the brief activity to a t-length fraction of the longer activity (i.e., Condition 2 obtains)? The preference is not revealed through the choice of the individual, because the choice runs directly counter to the supposed preference. Moreover, one may not even be able to say that in *previous* instances, the individual has revealed a preference for the briefer activity, for this instance may be the first time that the individual is presented with the choice opportunity. It seems that behavioral evidence supporting Condition 3 provides evidence *against* Condition 2. How, then, can a behaviorist determine when Conditions 1 to 3 obtain?

It is important to note that these criticisms only target Rachlin's particular definition of altruistic acts and their relation to acts of self-control. They have little impact on his primary observation that *because* altruism "for most of us . . . is not profitable and would not be chosen considering only its case-by-case, extrinsic reinforcement," altruistic behaviour is best explained by appealing to benefits conferred by our choosing to adopt abstract patterns of behaviour. In choosing to follow such patterns, we "forego making decisions on a case-by-case basis" even to the point of being altruistic "at the risk of death" (target article, sect. 10). This seems right, yet need not commit one to a behaviorist position. Moreover, this approach to understanding altruistic behavior raises important questions for future research. How do people acquire preferences regarding these valued abstract patterns of behaviour, and why do they choose to maintain them? Rachlin acknowledges that "extrinsic social reinforcement . . . at home or school or church" may explain the initial acquisition of such pattern. Yet when Rachlin says that "such acts must be maintained not by extrinsic reinforcement but by intrinsic reinforcement" (sect. 9), one wishes for more. The transition from extrinsic to intrinsic reinforcement asks for further explanation, while simultaneously underlining the need to move away from an explicitly behaviorist understanding of altruistic acts.

NOTE

1. A "context" for Rachlin seems to involve only the specification of the longer activity T, another free variable in his account: "[c]ondition 1 does not specify the appropriate context (the longer activity, T) for a particular act. Is there *any* context (any relatively long-duration activity, T) in which a given altruistic act would also be a self-controlled act? I believe that it will always be possible to find such a context." This suggests that the context of an act is solely determined by specifying the long-duration activity. This passage is somewhat confusing because it is not clear how one should understand the expression "altruistic act" appearing within it. I assume that should be read as referring to Rachlin's account. Yet in the sentence immediately preceding the quote, Rachlin asks "Are there altruistic acts under Conditions 2 and 3 above where Condition 1 does not obtain?" According to his definition, this is impossible because "an altruistic act is defined as a choice of the t-length fraction of the longer activity over the brief activity under Conditions 1, 2, and 3." Condition 1 must obtain for an altruistic act *by definition.*

Rationality and illusion

Jonathan Baron

Department of Psychology, University of Pennsylvania, Philadelphia, PA 19104-6196. **baron@psych.upenn.edu**
http:\\www.psych.upenn.edu\~baron

Abstract: Commitment to a pattern of altruism or self-control may indeed be learnable and sometimes rational. Commitment may also result from illusions. In one illusion, people think that their present behavior causes their future behavior, or causes the behavior of others, when really only correlation is present. Another happy illusion is that morality and self-interest coincide, so that altruism appears self-interested.

Consider two patterns for yourself, behaving selfishly and behaving altruistically. Behaving altruistically can be seen as a commitment, what Irwin (1971) called an "extended act." Rachlin may well be right in arguing that the altruistic pattern is better in terms of your long-run self-interest and that you can learn this in a variety of ways.

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

But a third pattern might be to attempt to find somewhat more complex rules about whether altruism is in your long-run self-interest. You could then apply these rules case by case. Why can't such a calculated pattern be cost-effective? It might be better for you, in the long run, than indiscriminate altruism.

Possibly, the attempt to discover and apply such complex rules is too time-consuming, difficult, and/or error prone. If you knew that the effort is not cost-effective, then you would rationally adopt the simpler rule of indiscriminate altruism.

On the other hand, it is easy to imagine cases in which the calculation could be quite cost-effective. For example, it seems safe to assume that altruism is worthwhile when your behavior is observed by others with whom you will interact in the future. But the long-run benefits of altruism are typically absent in large-scale social dilemmas, especially when your behavior is not easily monitored and rewarded by others in the next round. Political behavior is often like this. Voting is secret, so voting for the good of others, a cooperative and altruistic act, is difficult for them to discover and reward. You could rationally calculate that voting on the basis of effects on others is an exception to the pattern of altruism that is otherwise rewarded. If you were rational, you would decide not to vote on the basis of others' welfare, although if you were consistently altruistic you might do this anyway (Baron 1997a).

If altruistic or prudent behavior patterns are not always rational, they could be supported by illusions. In one sort of illusion, people may confuse correlation and causality, thinking that, because their cooperation is correlated (across situations) with the behavior of others, their choice can influence others. Often this is true, but people may overgeneralize to cases in which it is false. Quattrone and Tversky (1984) found that subjects were more willing to vote for their favored political party when they were part of a group considered crucial than when they were not, even though they had one vote in each case. Presumably, the subjects reasoned, "I'm just like everyone else; so if I cooperate, they will cooperate." The logic implies correlation (as a result of variation of external factors that affect everyone) but not causality. Shafir and Tversky (1992) found that subjects were likely to defect in a two-person social dilemma when they knew that their partner had defected, or when they knew that their partner had cooperated. But, when they did not know their partner's choice, more subjects cooperated, behaving as if they could influence their partner.

The same kind of illusion could affect self-control. The drinker may reason, "Tonight is just like any other night. If I drink to excess, then I am a drunkard, and I will continue to drink to excess." It is true that individual differences in the drunkard trait will influence behavior tonight and in the future, but this does not imply that the drinker's choice tonight will affect his future.

His choice may affect his future choices for other reasons. For example, he may save effort in decision making by using his past behavior as a guide. If he knows that he does this, then he has reason to consume moderately. This by itself may not be enough reason, and it may need the support of an illusion. The idea of setting a precedent may have both rational and illusory components. Putting this another way, the division of options into "get drunk tonight and forever" and "be moderate tonight and forever" neglects the possibility of a third option, "get drunk tonight and then be moderate." The third option is real, however difficult it may be to pull off in fact.

A second type of illusion that causes cooperation is the "illusion of morality as self-interest" (Baron 1997b). People seem to deny the existence of the conflict between self and others, the conflict that defines a social dilemma. When presented with a scenario about a social dilemma, in which it was apparent that cooperation was not in one's self-interest, many subjects said that it was, even to the point of saying that cooperators would make more money. (The result with money shows that subjects were not interpreting *self-interest* to include benefits resulting from altruism or from emotional responses.) Although the scenarios explicitly denied the possibility of influencing others, many subjects also showed the voter's illusion, arguing that if they cooperated then others would cooperate as well. The two illusions reinforced each other.

An analogous illusion may promote self-control. People may convince themselves, through wishful thinking, that they really do not want that third beer. In other words, they may see their immediate self-interest as coinciding with their long-term self-interest. In this case, because the desire in question *is* short term, it is difficult to distinguish self-deception from a real change in desire.

Illusions that promote altruism and self-control can be beneficial, but the self-interest illusion is also dangerous, because it is exacerbated when a person is a member of a group and cooperation is on behalf of the group. People often sacrifice their self-interest for a group to which they belong, even when outsiders are harmed so that the sacrifice has no net benefit. Baron (2001) found that people who cooperate on behalf of a group in such a parochial way (at the expense of outsiders) are more prone to the self-interest illusion. The illusion is in part the result of not making the calculations. One experiment found that the illusion (and the resulting parochial cooperation) was reduced when subjects were required to calculate all gains and losses.

Can't we all just be altruistic?

Gwen J. Broude

Cognitive Science Program, Vassar College, Poughkeepsie, NY 12604-0152.
Broude@vassar.edu

Abstract: Neither evolutionary theory nor behavioral evidence is consistent with Rachlin's view of altruism as a learned, domain-general learned habit displayed because of its intrinsic value. But human beings can be psychologically motivated by altruism while still reaping a genetic benefit from their altruistic actions.

Rachlin's stated purpose in writing this target article is to explain altruism "in its own terms – as a kind of habit" (target article, sect. 1.3). Rachlin is responding to the dogma from evolutionary psychology that biological organisms, including human beings, do not habitually engage in genuine altruism, because sacrificial acts would not be favored by natural selection. This message seems to offend Rachlin who, we gather, thinks that the world would be a better place if, given the choice between an altruistic and a selfish act, we would all choose altruism "even at the risk of death" (sect. 10, "Conclusions").

Rachlin views his behaviorist take on altruism as complementary to an evolutionary theory of altruism. So let us see how easily his version of altruism lives beside evolutionary predictions. Rachlin's interpretation of altruism as a habit unpacks into two main claims, which we can evaluate in turn.

Claim 1. Altruism is not a domain-specific phenomenon. Rather, it is a kind of learned self-control. Rachlin argues that altruism cannot be a product of a special-purpose mechanism, as cognitively oriented evolutionary theorists insist, because such a mechanism would put "too much of a load on inheritance." But then, doesn't the attribution of altruism to some domain-general learning mechanism place too much of a load on a domain-general mechanism? If Rachlin's theory of altruism is going to remain consistent with the logic of evolutionary theory, then we want to avoid positing a mechanism that would incline a person to learn any old thing about altruism (or anything else). This would make evolutionary nonsense, as organisms would be liable to learn to engage in sacrificial behavior at the expense of their fitness.

Further, the idea is not supported by evidence. Research on moral development shows that the moral behavior of children tends to be situation-specific. With regard to altruism, sharing a toy is not correlated with sharing food, or a spot on a picnic blanket, and so on in the actions of the same child (Harper 1989). This kind of situation-specificity of behavior is not surprising to an evolutionary theorist, who would expect biological organisms to behave in any

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

specific context consistent with a cost-benefit analysis. Nor is there any reason to believe that these discriminations are being made by children because they have somehow learned to be generous with their food but not a toy or vice versa. The fact that context-sensitivity is displayed by children at least suggests that we have a deeply rooted domain-specific cognitive function operating here.

Cosmides and Tooby's (1992) work on cheating algorithms, which suggests that human beings are especially sensitive to being cheated, also suggests the existence of a special mechanism dedicated to a vulnerability relevant to altruism in particular. Again, we see evidence of the mechanism in the actions of very young children, who resist giving away or sharing their possessions without some concrete return on their "investment," as any parent knows. Across cultures, although adults attempt to teach children to share food, prized objects, and other resources, youngsters object, lie, hide their belongings, and so on, in defiance of adult urging (Broude 1995).

All of this suggests that decisions about when to display altruism are a basic feature of human cognition. If we agree that human beings generally behave in a way that is consistent with evolutionary predictions, then a domain-general learning mechanism needs to carry the load of explaining how we get untutored situation-specific choices regarding altruism, even in toddlers, across cultures, without benefit, and in defiance of instruction from the environment. This is a heavy load indeed. Better to posit inborn mechanisms mediating evaluations regarding altruism.

This is not to say that no learning occurs with regard to altruism. Human beings need to learn the value of acts, possessions, and so on, that are not species-specific. Otherwise, there would be no basis on which to make choices. But it is not the tendency to display or withhold altruism, but rather the value of the "commodities," that is likely to be learned.

Claim 2. Altruism can be taught if the learner can be made to view altruistic acts as comprising a pattern of behaviors with high value to the actor. Here, the evolutionary psychologist is left to wonder where the high value of the habit of altruism comes from. Sometimes, Rachlin implies that the value comes from reinforcement for performing altruistic acts. But then, of course, we don't have altruism. Rachlin also claims that high value can come to be abstract and intrinsic. But then, how does such a pattern of behavior that has no concrete benefit to the actor, even over the long haul, get selected?

In the final analysis, Rachlin is arguing that human beings can become altruists by learning to ignore any short-term costs of an act in favor of viewing the lifelong habit of altruism as something to be valued in and of itself. The idea that human beings can perform short-term acts of altruism without rewards is perfectly consistent with evolutionary theory. Such tendencies allow individuals to reject smaller short-term payoffs in favor of longer long-term benefits. We might say that such a person is exercising self-control. This does not appear to be enough for Rachlin, for whom even long-term cost-benefit analyses are morally repugnant.

Rather, what Rachlin seems to want is a species that also engages in long-term altruism. But this is impossible if evolutionary thinking is right. Imagine a world in which we were all altruists. You and I are both hungry. We come upon a single apple, enough to feed just one of us. "You take the apple," I say. "No. You do," you respond. "No, you do," I insist. And so on. Hmmm. . . .

Theorists like Rachlin who wish to salvage altruism in our species might want to take a page from Midgley (1978) whose simple distinction between genetic and psychological altruism lets us eat our cake and have it too. As Midgley points out, evolutionary theory only predicts *genetic* selfishness, by which it is meant that behaviors must on the whole promote the survival of the genes of the actor. But acts that promote the actor's genes can be motivated by genuine altruism at the level of psychological motivation. I can genuinely wish to help you, with no expectation of return. If that induces you to help me in the future, my genes reap the reward, but I was certainly not looking for one. Indeed, Rachlin's woman who runs into the burning building to save someone else's child is acting out of psychological altruism. But an evolutionary theorist would observe that the act may ultimately promote the survival of her genes if the grateful parents return the favor.

"Choice" and "emotion" in altruism: Reflections on the morality of justice versus the morality of caring

Ross Buck

Department of Communication Sciences and Psychology, U-1085, University of Connecticut, Storrs CT 06269-1085. **Buck@uconnvm.uconn.edu**
http://www.coms.uconn.edu/people/faculty/rbuck/index.htm

Abstract: Rachlin uses the word "choice" 80 times, whereas "emotion" does not appear. In contrast, "Empathy: Its ultimate and proximate bases" by Preston and de Waal, uses the word "emotion" 139 times and "choice" once. This commentary compares these ways of approaching empathy and altruism, relating Rachlin's approach to Gilligan's Morality of Justice and Preston and de Waal's to the Morality of Caring.

The impetus for this response to Rachlin's "Altruism and selfishness" comes from an objective comparison of that paper and a related article by Preston and de Waal (2002) "Empathy: Its ultimate and proximate bases," which appears in the previous *BBS* issue. A word count reveals the word "choice" 80 times in the body of Rachlin's article, whereas the word "emotion" does not appear at all. In comparison, Preston and de Waal have used "emotion" 139 times and "choice" only once in their paper. Clearly, these articles represent fundamentally different ways of looking at the phenomena of empathy and altruism, with neither one addressing the positions of the other. Rather, each is talking as if the other point of view did not exist. The aim of this commentary is to consider how the emotional controls of behavior emphasized by Preston and de Waal might relate to Rachlin's approach.

Rachlin uses the actions of a woman who runs into a burning building to save someone else's child as a symbol for the kinds of actions he wishes to explain. He argues that such a behavior can be explained as a consequence of a commitment to an altruistic pattern of acts learned over the course of an individual's lifetime. He compares this to the learning of self-control, in that particular acts are formed into coherent patterns of acts, and suggests that this is particularly relevant to human behavior. I agree with all of these points, and add that the development of such coherent behavioral patterns requires the linguistic structuring of behavior.

Linguistic competence makes possible a system of behavior control unique to human beings: "[O]nly in humans does behavior come so completely under the control of principles that are mediated by language, including logic, reason, and social rules" (Buck 1985, p. 406). Moreover, language is the basis of culturally patterned systems of behavior control that are "functionally independent of biology and fundamentally different from anything seen in animals" (Buck 1988, p. 30). These include principles of logic and reason on which general moral rules and judgments are based. Linguistic competence is necessary for most cognitive consistency and attribution processes, and the sense of self, as well as for abstract moral judgments. Indeed, language is the behavior control system underlying the notion of the "rational soul" of Plato and Aristotle.

However, I disagree with Rachlin's argument that the existence of this learned coherent pattern of behavior makes it unnecessary to postulate the existence of a general mechanism for altruism. There are mechanisms for the control of human behavior other than linguistic ones, and these are shared with other animals. In this regard, Thomas Aquinas equipped humans with a "sensitive soul" similar to those in animals, as well as a unique "rational soul." Descartes' mind-body dualism similarly distinguished between animal behavior, which could be explained by purely mechanical forces, and human behavior, which was partly mechanical and partly based on a nonmechanical soul. Ryle (1949) ridiculed this

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

theory as the "dogma of the ghost in the machine," but arguably language actually does function in some ways like a ghost in a machine. Linguistic control systems enable human beings to transcend individual experience and allow the contemplation of possibilities that never have been, and never could be, experienced. It also allows the symbolic sharing of experiences with others, including others long dead: Plato, Aristotle, Aquinas. Ghosts, of a kind.

The Western tradition has often viewed logic and reason as somehow superior to the passions we share with animals: Indeed, "right conduct" is often viewed as involving the control of animal passions. However, reason, logic, and organized social rules have been at the core of some of the most violent and destructive of human behaviors, including the official directives, chains of command, and orderly bureaucratic procedures of the Holocaust. An alternative is to view prosocial emotions of attachment and bonding as being the truly effective counter to aggression and violence (Buck 1988; 1999).

There is evidence of the importance of emotional bonds in mediating a variety of behaviors with moral implications: fostering cooperation and altruism and reducing aggression and conflict. Examples of such emotional controls of behavior have been found in conflict resolution among monkeys and apes, as observed by de Waal and colleagues (e.g., de Waal 1996; de Waal & Aureli 1997). In human beings, there is considerable evidence that feeling empathy for a needy person leads to altruism, that is, to unselfish tendencies to help that person (Eisenberg & Fabes 1991; Eisenberg & Miller 1987; Hoffman 1975; 1976). C. Daniel Batson and colleagues (Batson & Oleson 1991; Batson & Shaw 1991) reviewed evidence for the role of selfish motives in altruistic behavior and advanced the empathy-altruism hypothesis (EAH): that the expression of needs by the other naturally evokes empathic emotions of sympathy and compassion that motivate altruistic responses.

In a larger sense, the gulf between the approach of Rachlin and that of Preston and de Waal reflects the gap between the Piaget-Kohlberg analysis of moral judgment and the morality of caring emphasized by Carol Gilligan. Gilligan and colleagues argued that there are two fundamental moral orientations. The justice perspective "holds up an ideal of reciprocity and equal respect," whereas the care perspective "holds up an ideal of attention and response to need" (Gilligan & Attanucci 1988, p. 73). Either or both of these perspectives can be active during moral choice. Moreover, there are suggestions that choice behavior per se is more emotional than previously believed (Lowenstein et al. 2001).

Human behavior is multiply determined, and it arguably is an error to dismiss another point of view because it is "not necessary" to explain a phenomenon. All aspects of human empathy and altruism cannot be explained by the principles developed by Preston and de Waal (2002). However, an exclusive focus on choice can leave out the emotional controls that may set the basic agenda for human morality in general and altruism in particular.

The need for proximal mechanisms to understand individual differences in altruism

Gustavo Carlo and Rick A. Bevins

Department of Psychology, University of Nebraska-Lincoln, Lincoln, NB 68588-0308. **gcarlo@unl.edu**

Abstract: There are three concerns regarding Rachlin's altruism model. First, proximal causal mechanisms such as those identified by cognitive neuroscientists and behavioral neuropharmacologists are not emphasized. Second, there is a lack of clear testable hypotheses. And third, extreme forms of altruism are emphasized rather than common forms. We focus on an overarching theme – proximal mechanisms of individual differences in altruism.

Rachlin proposes a theory of altruism that focuses on self-control as the central explanatory mechanism. Other notable aspects to this model include the emphasis on temporal patterns of altruistic behaviors and the connections made to evolutionary theory. Although Rachlin is to be commended for elevating the importance of these aspects to explain altruism, there are a number of gaps in the model that seriously limit this theoretical perspective. These limitations will likely decrease any impact his theory may have on the field. We briefly outline some of these concerns and propose avenues for future theoretical and empirical pursuit.

Psychological debate and research on altruism have often focused on the existence of altruism. Assuming the existence of altruism moves the debate to the nature of altruism. Accordingly, it is important to place the issue of altruism in its broader context. Altruism is considered to be a subset of the larger set of prosocial behaviors (i.e., behaviors that benefit others) which include behaviors that *primarily* benefit others, often incurring a cost to the self (Carlo & Randall 2001; Eisenberg & Fabes 1998). This definition requires one to acknowledge the presence of much variation in the forms of altruism. Our working definition is, of course, subject to debate; but for the purpose of the present commentary, it is necessary to make explicit. That is, our subsequent comments regarding Rachlin's model may be partly attributed to definitional differences.

We have three recurring concerns regarding Rachlin's altruism model. First, virtually no time is spent on proximal causal mechanisms such as those that might be provided by areas like cognitive neuroscience and behavioral neuropharmacology. Second, there is a lack of clear testable hypotheses that follow from this self-control model. Finally, the model appears to be built around extreme forms of altruism (e.g., woman saving unknown baby) rather than the more regularly occurring forms that vary widely in occurrence across and within individuals. From our perspective, these concerns are interrelated. Thus, this commentary will address these issues by focusing on an overarching theme – proximal mechanisms of individual differences in altruism.

An account of individual differences in altruism requires the consideration of more proximal causal mechanisms. According to some theorists (Carlo & Randall 2001; Eisenberg & Fabes 1998; Hoffman 1991), individual differences in altruism can stem from differences in cognitive, emotive, and social context (e.g., culture-related socialization experiences). We will focus on a specific set of emotive variables relevant to altruism. A number of investigators posit that empathy is the primary motivator associated with altruism (Batson 1998; Carlo & Randall 2001; Eisenberg & Fabes 1998; Hoffman 1991). Empathy can be defined as an other-oriented matching emotion that results from vicariously observing another's distress. There are two processes that may stem from empathy: sympathy and personal distress. Sympathy is an other-oriented vicarious emotional response that results in feelings of sorrow or concern for a needy other. In contrast, personal distress is a self-focused vicarious emotional response that results in aversive, uncomfortable feelings. The difference between these two empathy-related responses is critical because while sympathy can result in prosocial behaviors (including altruism), personal distress often results in avoidance behaviors (but see Batson 1998).

The characteristics of individuals who exhibit these different responses in distress situations are distinct. For example, empathy and sympathy responses reflect moderate sympathetic arousal whereas personal distress reflects over-arousal. Derryberry and Rothbart (1988) proposed a temperament theory that identifies two major dimensions: physiological reactivity and self-regulation. Physiological reactivity refers primarily to affective arousal and motor activity. It includes an assessment of emotional and behavioral threshold, latency, intensity, and rise and fall time. Self-regulation refers to behavioral and emotional control. Specific aspects of this dimension include attentional processes, approach-withdrawal, soothability, and behavioral inhibition (Rothbart et al. 1994). More important, Rothbart proposes that empathic and prosocial tendencies are best explained by examining the interaction between these two internal processes. Following Rothbart's model, personal distress responses are the joint consequences of

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

individuals who have difficulty modulating their physiological arousal and have overactive physiological reactivity; there is evidence to support these assertions (Rothbart et al. 1994; see also Eisenberg & Fabes 1998). Thus, Rothbart's approach represents a set of more proximal causal mechanisms that might account for individual differences in altruism, and it also respects the wide variations in altruism.

In contrast to Rothbart's model, we are unable to find in Rachlin's proposal an explicitly defined set of analogous proximal causal agents. This situation is unfortunate in that such proximal mechanisms would help account for the wide variation seen in the forms of altruistic behaviors, the individual differences seen between and within individuals, and, arguably the most important point, provide researchers with clear and testable hypotheses that avoid circularity (Panksepp 1998). For example, Rothbart's notion of distress as a functional interaction between physiological reactivity (e.g., arousal) and self-regulation (e.g., emotional control) suggests distinct neurophysiological processes (e.g., hypothalamic pituitary adrenal axis versus frontal cortical areas, respectively). Notably, individual difference in distress, and hence likelihood of behaving altruistically, becomes a product of these processes that vary with evolutionary and individual history. The task of identifying all neurobiological factors that mediate the presence or absence of an altruistic behavior at time *x* is daunting. However, it is an obtainable goal that already has a basis from which to start. Powerful animal models exist that could be used to explore the processes posited to mediate altruism (e.g., distress). There are numerous papers concerning rodent models, indicating that different experiences in early development (e.g., naturally occurring maternal care) can differentially impact later sensitivity to distress (e.g., alteration in hypothalamic pituitary adrenal axis, see Liu et al. 1997; for other examples, see Boksa et al. 1998; Dellu et al. 1996; Kehoe et al. 1998). An exciting possibility would be to merge the work on individual difference with an animal model of self-control that attempts to measure choice that includes prosocial options (see Poulos et al. 1998 for an example of assessing individual differences using a self-control [impulsivity] preparation).

ACKNOWLEDGMENTS

Funding support to the first author was provided by a grant from the John Templeton Foundation and the Mayerson Foundation. The second author was partially supported by USPHS grant DA11893 while preparing this commentary. Correspondence may be addressed to Gustavo Carlo, Department of Psychology, University of Nebraska-Lincoln, Lincoln, NB, 68588-0308, e-mail: gcarlo@unl.edu.

Learning to cooperate: Reciprocity and self-control

Peter Danielson

Centre for Applied Ethics, University of British Columbia, Vancouver, V6T 1Z2, Canada. **pad@ethics.ubc.ca** **http://www.ethics.ubc.ca/~pad**

Abstract: Using a simple learning agent, we show that learning self-control in the primrose path experiment does parallel learning cooperation in the prisoner's dilemma. But Rachlin's claim that "there is no essential difference between self-control and altruism" is too strong. Only iterated prisoner's dilemmas played against reciprocators are reduced to self-control problems. There is more to cooperation than self-control and even altruism in a strong sense.

As Rachlin points out, the analogy between self-control and altruism has a long history. For example, Sidgwick (1893) and Nagel (1970) use the analogy as a bridge from the natural appeal of prudential motivation to a justification of altruism. These authors argued that since we stand to others as to our own future selves, we should not treat the two cases differently. Rachlin moves the normative discussion from this weak appeal to rationality as symmetry, to a productive blend of formal and empirical modeling methods.

Rachlin makes two significant contributions to the study of social cooperation. First, in a literature that has emphasized rationality and evolution, he focuses on agents learning to cooperate. Second, he develops the analogy between learning self-control and learning social cooperation by showing that two problems in these areas share a common structure. The primrose path (PP) to addiction is a problem for self-control and the prisoner's dilemma (PD) is a problem for cooperation because immediate reinforcement is higher for the addictive choice, X, or the noncooperative choice, D, respectively. Thus, each presents learners with a local optimum trap. Rachlin argues that learners who structure reinforcement for patterns of acts rather than single acts can avoid these traps.[1]

Although he offers much to build on, Rachlin's claim that "there is no essential difference between self-control and altruism" (target article, sect. 7, last para.) is too strong. Self-control may be necessary for cooperation, but it is not sufficient; he underemphasizes the crucial factor of reciprocity. That is, although Rachlin puts great weight on reciprocity, the complexities of dealing with human self-control and cooperation obscure some issues. We propose to clarify the discussion by focusing on simple artificial agents and games.

The main problem is that whereas a temporally extended agent is the benefactor from her own self-control, this need not be the case for cooperation. In the best case, cooperators share the benefits with others; in the worse case, unreciprocated cooperation harms the agent compared to her opponent. So we suspect that self-control might be learnable by reinforcement under weaker conditions. To test our intuition, we construct the simplest learner, who tries two alternative acts and selects the one with the higher immediate reinforcement; it will choose addictive X in the PP. Even if we allow reinforcement to be remembered over trials, the result is the same (Fig. 1). But if we constrain choice (and therefore reinforcement) to the simplest patterns, that is, to sets such as XXX and YYY, Y is chosen (Fig. 2).[2]

Turning to the PD, we get parallel results. The act chooser will learn to defect and the pattern chooser to cooperate. In Figure 3 the dark bars show Rachlin's PD game iterated for 40 rounds with a Learner playing against Tit-For-Tat. Performance increases with pattern length up to 3, then decreases (because the longer defection trial is costly). But we will get this parallel only under restrictive conditions. First, the game must be iterated; in the one-shot PD, D remains dominant. Rachlin notes that his subjects "are making only one in a series of choices extending to their lives outside of the lecture hall" and claims that "for the teleological behaviorist there can be no social trap without repetition. All prisoner's dilemmas are repeated " (sect. 3, last para.). Be this as it may for humans, we can impose the restriction explicitly for automata and focus on repeated PD games. Second, learning to cooperate

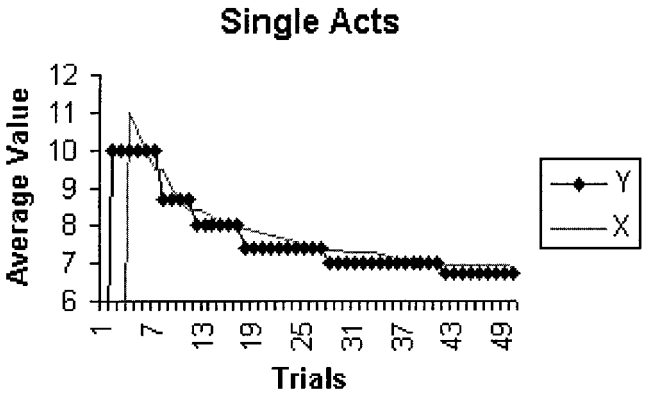


Figure 1 (Danielson).

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

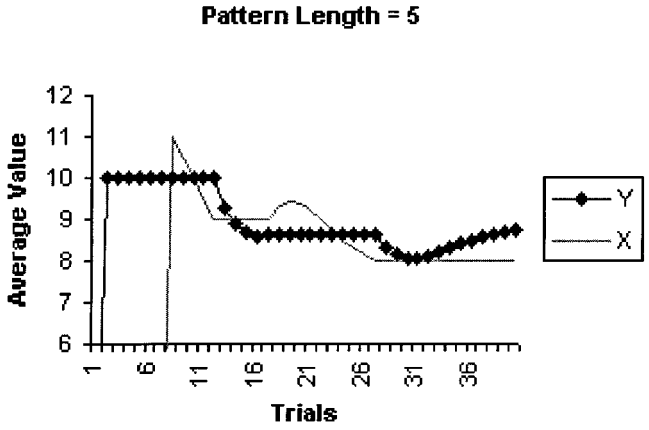


Figure 2 (Danielson).

only follows from self-control if one is facing a particular kind of agent: a reciprocator. To see this, suppose instead that one faced All-D. Our learner learns to D; this is good. But suppose it faces another similar learner. Now the outcome is due to chance. Half the time they coordinate their tests, trying, for example, CCC against CCC and learning to cooperate, otherwise they try CCC against DDD and learn to defect.[3] In any case, they never learn to reciprocate.

Indeed, given the assumptions of repetition and reciprocity, Rachlin's focus on *altruism* rather than *cooperation* is problematic. Cooperating with a reciprocator in an iterated PD, demands only minimal altruism, as payback is almost certain and only delayed two or three moves.[4] In general, Rachlin's emphasis on altruism is potentially misleading. Following biologists, he uses *altruism* where economists use *cooperation*, merely to denote the choice of the dominated C strategy. So altruism adds nothing about motivation. However, Rachlin's traditional example of the dangerous rescue of another's child seems to go beyond cooperation to demand strong altruism, that is, valuing another's welfare. However, even this stronger sort of altruism alone will not produce stable cooperation in social dilemmas; reciprocity is also necessary (Danielson 2002; Levine 1998; Sethi & Somanathan 2001).

Summing up, there is an analogy between self-control and cooperation; both require that learning apply to patterns rather than to isolated acts. But the PP is simpler than the PD. In addition to patterned learning, the PD requires reciprocity (to maintain various non-D equilibria) and likely something else to select socially better equilibria. Self-control only takes us part of the way.

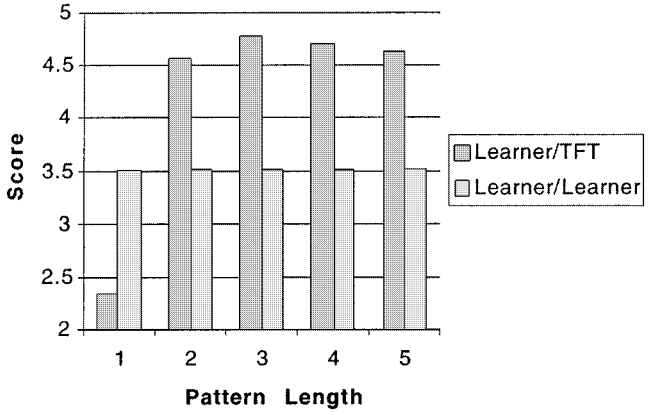


Figure 3 (Danielson).

NOTES

1. This strategy is related to the "rule rationality" of Gauthier (1986) and McClennen (1998).

2. Note that even patterned learning is not straightforward in the PP; the Y value dips under the X. The reason is that so long as the agent remains open to learning, the second group of Y-choices falls under the shadow of the first X trials. The pattern must be long enough that X shows how bad it can get. (In our experiment, the hangover has a length 5; initial Y reward is 10; X reward 11, declining 1 point for each of the next 5 X choices. A pattern length of 5 will just barely escape the PP, 6 very easily (after 1 shot learning), and 3 not at all.

3. See Marinoff (1998) for more complex cases.

4. This is the reason Gauthier (1986) denies that the repeated PD is relevant to morality.

The role of negative reinforcement; or: Is there an altruist in the house?

Edmund J. Fantino and Stephanie J. Stolarz-Fantino

Department of Psychology, 0109, University of California, San Diego, La Jolla, CA 92093-0109. **efantino@ucsd.edu** **sfantino@psy.ucsd.edu**

Abstract: We agree with Rachlin's argument that altruism is best understood as a case of self-control, and that a behavioral analysis is appropriate. However, the appeal to teleological behaviorism and the value of behavioral patterns may be unnecessary. Instead, we argue that altruism can generally be explained with traditional behavioral principles such as negative reinforcement, conditioned reinforcement, and rule-governed behavior.

In the past, Rachlin's empirical and theoretical papers persuasively argued that self-control could be understood as a specific case of choice behavior, abiding by the same general laws and principles (e.g., Baum & Rachlin 1969; Rachlin & Green 1972). Arguing in the same vein, he now makes the case that altruistic behavior may be understood as a kind of self-control. He contrasts this behavioral viewpoint with accounts of altruism from cognitive and biological perspectives. We believe that he presents convincing arguments for the behavioral approach. But we are not persuaded that his version of behaviorism, teleological behaviorism, is necessarily superior to the version he rejects: what he refers to as "Skinnerian behaviorism." We submit that a behaviorist account of altruism is possible without recourse to the language of response patterns and teleology. We have two general points. First, can we really rule out contemporary reinforcers of the altruistic act (what Rachlin refers to as "current or even future reinforcers of the act itself")? If not, as Rachlin suggests, we may not need the concept of altruism after all. Second, if we grant that altruism does occur, may we not explain instances of it more parsimoniously in traditional behavioral terms?

The central example in the target article is that of the woman risking her life to save someone else's child. The choice facing our hypothetical woman is to risk her life and try to save the child, or to move on in safety and have the knowledge that she allowed the child to burn. Both choices involve negative reinforcement (escape or avoidance of an aversive outcome). By fleeing, the woman avoids personal injury; by entering the building, she may prevent the child's death. But there is an additional powerful component of negative reinforcement propelling the woman toward the altruistic act: By risking her life she also avoids living with the knowledge that she allowed the child to die. Most people have had some experience of feeling regret at having been unable to prevent some occurrence that resulted in harm to others. Acting to prevent a tragedy involving the baby would avert the possibility of experiencing that feeling. That some women may behave "altruistically" in this situation may, thus, in part reflect the role of negative reinforcement. Indeed, many (perhaps most) examples of altruistic behavior may be explained in terms of the negative reinforcement that they provide. In addition to future regrets, there are the

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

immediate aversive effects of seeing the baby in peril. Acting to put an end to this situation could involve fleeing the scene, as noted above; however, it could also be accomplished by saving the baby. Based on comments in the target article we can assume that Rachlin would say that these are not cases of altruism. Then what cases are?

Let's assume, however, that we can rule out contemporary reinforcers of the altruistic act, including negative reinforcement. In other words, let's assume that there are cases of what Rachlin would regard as true altruism. Can teleological behaviorism really provide a better or more parsimonious account of true altruism than traditional behaviorism? In distinguishing between these two types of behaviorism, Rachlin argues,

> Behaviorists do not disagree that internal differences exist but their focus is rather on noncontemporary events; the Skinnerian behaviorist is concerned to discover crucial differences in the soldiers' *extrinsic* reinforcement histories. The teleological behaviorist is concerned to discover the patterns of behavior of which each soldier's present act forms a part (*intrinsic* reinforcement). Note, however, that even the concept of extrinsic reinforcement must rely at some point on intrinsic reinforcement. (target article, sect. 9)

In regarding cases of true altruism, the traditional behaviorist would examine the history of the individual. For example, the woman attempting to save the baby may well have been instructed on the value of assisting others (rule-governed behavior) and may have been rewarded in the past for self-sacrificing behavior (contingency-shaped behavior). These behaviors will most likely have been given a common linguistic label ("helping," for example) that will make them more obviously members of a class. In addition, the behaviorist has traditional principles such as conditioned reinforcement and generalization that may readily explain the woman's behavior. And, as we noted previously, the woman may have learned that not to act will inevitably lead to the disapproval of others, including self-disapproval. By acting she avoids these aversive consequences.

Consider the issue of what makes people more or less likely to behave altruistically. What kind of behavioral history would more likely promote altruism? It is difficult to discuss the example of the imperiled baby without recalling the events of September 11, 2001. On that day scores of firefighters who were off duty rushed to the disaster and to their deaths. The behaviorist would say that they had a long reinforcement history of behaving in a similar manner. Indeed, all the principles we mentioned here are likely to help account for the firefighters' altruism (rule-governed and contingency-shaped behaviors, negative reinforcement, conditioned reinforcement, generalization, and stimulus control). Many of the contingencies in the firefighters' contemporary and historical environments prepared them for their acts of altruism. But perhaps this is all that Rachlin means when he claims that the altruist is following "an abstract pattern of behavior" that "may be (intrinsically) of high value" (sect. 9). If so, the distinction between traditional behaviorism and teleological behaviorism is blurred indeed. In any event, we suggest that the firefighters' behavior (whether technically altruistic or not) can be explained at least as directly by traditional principles as by an appeal to valued patterns of behavior (see Fantino & Stolarz-Fantino 2002 for a discussion of Rachlin's teleological behaviorism).

An important point raised by Rachlin's article is the possibility that altruism could be learned in the same way that self-control can be learned. Rachlin has already made the case (Rachlin 2000) that self-control can be enhanced by viewing one's acts as part of a larger temporal pattern. It remains to be empirically determined whether such an approach to the learning and maintenance of altruistic behavior would add to what can be accomplished by an appeal to rules and reinforcement alone.

ACKNOWLEDGMENT

Preparation of commentary supported by NIMH Grant MH57127.

Altruism and emotions

Herbert Gintis

Department of Economics, University of Massachusetts, Amherst, MA 01060.
hgintis@attbi.com **http://www-unix.oit.umass.edu**

Abstract: If altruism requires self-control, people must consider altruistic acts as costly, the benefits of which will only be recouped in the future. By contrast, I shall present evidence that altruism is dictated by emotions: Altruists secure an immediate payoff from performing altruistic acts, so no element of self-control is present, and no future reward is required or expected.

In most social species, cooperation is facilitated by a high level of relatedness of conspecifics. Human groups, however, sustain a high level of cooperation despite a low level of relatedness among members. Three approaches have been offered to explain this phenomenon: long-run self-interest, in the form of reciprocal altruism (Axelrod & Hamilton 1981; Trivers 1971), cultural group selection for prosocial traits, (Boyd & Richerson 1985; Cavalli-Sforza & Feldman 1981), and genetically based altruism (Lumsden & Wilson 1981; Simon 1993; Sober & Wilson 1998). Rachlin's account falls under the second category, because he considers altruism as non-self-interested and learned, as opposed to flowing from an internal (i.e., genetic) mechanism.

Rachlin argues that the habitual practice of helping others at a cost to oneself leads to a "happier mode of existence" than habitual selfishness, but conformance to a norm of altruism involves self-control because the pleasures of selfishness are immediate, whereas the returns to altruism manifest themselves only over long periods of time. Having reviewed the moral precepts of most of the world's great religions, I can attest that this attitude toward altruism is characteristic of most theological writings. The assertion of a deep harmony among the happy life, the ethical life, and the prosocial life is practically a religious universal. Moreover, if it were true, most of the problems involved in explaining the emergence and sustenance of altruism in society would disappear: The "costs" of helping others would fall short of the benefits, and altruism would be a fitness-enhancing, as well as long-run welfare-enhancing, behavior.

I want to explore one problem with Rachlin's analysis. The idea that altruism requires self-control makes sense only if people consider altruistic acts as *costly,* the benefits of which will only be recouped in the future. By contrast, I shall present evidence that altruism is dictated by emotions, not reasoned self-control: People secure an immediate payoff from performing altruistic acts, so no element of self-control is present.

The emotional basis of altruism lies in our possessing certain *prosocial emotions,* including empathy, shame, and guilt. As Adam Smith noted long ago in the *Theory of Moral Sentiments* (1759), people become uncomfortable when confronted with the dis-ease of others, and most will expend some effort to correct the source of dis-ease. People save babies from fires, then, because they empathize with the plight of the infant and pity the distress of its parents, not because they believe that altruism has a long-run personal benefit. Moreover, people would feel *ashamed* if discovered in the cowardly act of ignoring the baby's plight; even if no observers were present, many would engage in the altruistic act to avoid the *guilt* they would carry with them knowing the selfishness of their behavior.

Rachlin treats doing good deeds as unpleasant duties that bring future rewards in knowing one has had a life well lived. Experimental evidence, by contrast, indicates that personally costly prosocial acts are motivated by immediate emotional satisfaction. Consider the following public goods experiment, reported by Ernst Fehr and Simon Gächter in *Nature* (Fehr & Gächter 2002). A total of 240 subjects participated in a public goods game, the payoffs being real money (we'll call the money Monetary Units, or MUs for short). Subjects were assigned to four-person groups, and each person was given 20 MUs. Each was permitted to contribute (anonymously) some, none, or all of this amount to a "group proj-

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

ect" that gave a return of 0.4 MUs for each MU contributed to each member of the four-member group. Note that if each member acted selfishly, each would end up with 20 MUs. If all acted altruistically, however, each would contribute 20 MUs to the group project, so each would receive 0.4% x 80 = 32MUs. However, no matter what the others did, each would receive the highest payoff by contributing nothing. For example, if three players contributed everything and the fourth nothing, the three would earn 0.4% x 60 = 24 MUs, whereas the selfish noncontributor would earn 0.4% x 60 + 20 = 44 MUs.

Subjects made their investment decisions anonymously and simultaneously, and once these decisions were made, each was informed about the investment decisions of the other three players in the group. Subjects were then allowed to "punish" other group members by assigning between zero and 10 points to the punished member. Each point cost the punisher 1MU and the punishee 3 MUs. All punishment decisions were made simultaneously.

This game was repeated six times for each subject, but group membership was changed after each trial period, to eliminate the possibility that punishment could be self-interested by inducing partners to increase their contributions in future periods.

The public goods game is an excellent model of cooperation in human society because it involves cooperation among several people; the group gains from the cooperation of its members, yet each member has an incentive to free-ride on the effort of the others. Cooperation is thus an altruistic act. If we add the possibility of punishment, and if enough people are public spirited and actually punish free-riders, even selfish types will cooperate, so cooperation is no longer altruistic. But punishing free-riders is an altruistic act – personally costly, but beneficial to the group because it induces others to contribute.

Fehr and Gächter found that, when punishment was not allowed, the level of cooperation in the first period was about 50 (people invested about half of their 20 MUs), but cooperation deteriorated steadily until at round six, when cooperation was only 10. However, when punishment was allowed, cooperation was 60 in the first period and *rose* to about 75 in later periods. Clearly, altruistic punishment allowed cooperation to be sustained in this public goods game.

According to Rachlin, altruistic punishers should have felt it costly to punish, but punished anyway because they had a high level of self-control and perceived the long-run personal gains of behaving morally. In fact, however, subjects felt *anger* directed toward the free-riders and punished them to achieve the immediate emotional reward of venting this anger on an appropriate individual. Fehr and Gächter ascertained this fact by postgame interviews with subjects, in which players' motivations were elicited through questionnaires.

ACKNOWLEDGMENT

I would like to thank the John D. and Catherine T. MacArthur Foundation for financial support.

Can altruism be understood in terms of socially-discounted extrinsic reinforcement?

Randolph C. Grace, Anthony McLean, and Orn Bragason

Department of Psychology, University of Canterbury, Christchurch, New Zealand. **{r.grace; a.mclean; o.bragason}@psyc.canterbury.ac.nz**
http://www.canterbury.ac.nz

Abstract: Altruism can be understood in terms of traditional principles of reinforcement if an outcome that is beneficial to another person reinforces the behavior of the actor who produces it. This account depends on a generalization of reinforcement across persons and might be more amenable to experimental investigation than the one proposed by Rachlin.

Rachlin's behavioristic treatment of addiction is neatly recast to explain another important aspect of human behavior, altruism. But addiction arises out of a conflict between tangible consequences: gains for individual acts of consumption versus long-run costs for persistent consumption. The same account applied to altruism is perhaps less convincing because it seems to oblige us to invent the reinforcers for altruistic behavior that counteract the costs associated with individual acts of altruism. Our main concern here is whether an account based on traditional principles of extrinsic reinforcement may be more parsimonious than the one proposed by Rachlin.

As Rachlin points out, an explanation of altruism without reinforcement faces a problem: An internal, inherited mechanism must be somehow aroused by situations in which costs and benefits to oneself are weighted against costs and benefits to others. But his learning-based mechanism must, similarly, be somehow aroused if it is to explain altruistic acts. Presumably that job would be done by the discriminative stimuli that signal the existence of contingency structures like the ones he outlines in Figure 1. It appears to us that discriminative stimuli are needed in the development and maintenance of a behavior pattern that is, as a whole, intrinsically valuable. We are not sure what place discriminative stimuli play in Rachlin's account.

We would like to question Rachlin's assumption that altruistic acts are always less preferred than non-altruistic acts. We could understand altruism in terms of traditional principles of reinforcement if an outcome that is beneficial to another person reinforces the behavior of the actor who produces that outcome. The behavior of a woman who runs into a burning building to save a child is reinforced by the benefit to the child (i.e., life being saved). In other words, the woman may be choosing the most highly valued immediate alternative on a case-by-case basis. This account depends on a generalization of reinforcement across persons. Presumably, a positive outcome that accrues to oneself is maximally valued. The value of positive outcomes that occur to others would be inversely related to the social distance between the actor and the other person. So, for example, saving the life of a spouse or family member might be similarly valued to saving one's own. Saving the life of a complete stranger would be less valued. Reinforcement may thus be socially discounted; that is, the reinforcing efficacy of a positive outcome occurring to another person may decrease as social distance increases. Ultimately, it may be possible to explain socially discounted reinforcement in terms of fundamental learning processes (e.g., generalized imitation).

There are several potential advantages to this account. (1) It is arguably more parsimonious than Rachlin's because it is more firmly based on traditional principles of (extrinsic) reinforcement. (2) It raises the possibility of quantifying altruism in terms of how rapidly positive outcomes are discounted as a function of social distance. Highly altruistic individuals are those for whom positive outcomes for complete strangers are nonetheless highly reinforcing. Selfish persons are those for whom only outcomes that benefit themselves are reinforcing. (3) The analogy with self-control is sharpened because social discounting parallels the way in which reinforcing efficacy decreases with delay. The study of intertemporal choice attempts to quantify an individual's impulsivity in terms of how reinforcing outcomes are discounted as a function of time (e.g., Chapman 1998; Kirby 1997). The relationship between altruism and self-control may be understood as similar discounting processes for extrinsic reinforcement, social and temporal. (4) It is possible, at least in principle, to investigate empirically the process of social discounting. For example, social distance might be related to (and in the laboratory, measured by) the history of exchange of reinforcers and punishers between two parties. (5) Finally, it may be better to understand altruism as a series of case-by-case choices. One problem with Rachlin's account is that his notion of behavior pattern is vague. His example of how missing the last three minutes of a symphony can ruin the entire experience might seem to suggest that a single selfish act in old age might destroy what otherwise was an altruistic life. Surely someone who chooses to act altruistically 99% of the time is still altruistic.

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

Socially discounted reinforcement may seem similar in some respects to vicarious reinforcement, in that the outcome's effects are observed rather than directly experienced (e.g., Bandura et al. 1963; Deguchi 1984). Vicarious reinforcement has been proposed as a necessary component of observational learning (Bandura 1972). However, the crucial difference is that whereas vicarious reinforcement serves an instructional and motivational role, as understood here, socially discounted reinforcement acts directly on the altruist's behavior.

Overall, Rachlin's target article is an important and provocative contribution. However, we are curious to know what his account can provide beyond the one we have outlined. Perhaps some, if not all, altruistic behavior can be understood in terms of socially discounted extrinsic reinforcement. This framework might be more amenable to experimental investigation and more conducive in the long run to understanding this important phenomenon.

Cognitive control in altruism and self-control: A social cognitive neuroscience perspective

Jeremy R. Gray and Todd S. Braver

Department of Psychology, Washington University, St. Louis, MO 63130.
jeremy_gray@post.harvard.edu tbraver@artsci.wustl.edu
http://artsci.wustl.edu/~jgray http://iac.wustl.edu/~ccpweb/

Abstract: The primrose path and prisoner's dilemma paradigms may require cognitive (executive) control: The active maintenance of context representations in lateral prefrontal cortex to provide top-down support for specific behaviors in the face of short delays or stronger response tendencies. This perspective suggests further tests of whether altruism is a type of self-control, including brain imaging, induced affect, and dual-task studies.

The idea that altruistic behavior is a special case of self-controlled behavior is deeply intriguing. However, although Rachlin's argument is elegant and particularly strong on analysis, it is not as well grounded empirically. A social cognitive neuroscience perspective (Ochsner & Lieberman 2001) suggests multiple ways to test whether being altruistic in a prisoner's dilemma situation requires self-control. We first present a task analysis of the two decision-making paradigms.

Self-control in the primrose path paradigm might require cognitive control (Gray 1999): the control of thought and behavior by representations of context information actively maintained in lateral prefrontal cortex (PFC) (Braver & Cohen 2000; Gray 2001). Context is defined as any information that is represented internally in a form that is able to bias processing in the neural pathways responsible for task performance. Goals are a paradigmatic type of context information and must be active (rather than latent) to influence behavior. Context information can also include task instructions or a prior stimulus, or it can be the result of processing a sequence of prior stimuli and responses. Active representations of context can control behavior by biasing brain activity in structures that subserve task-specific processes (e.g., mapping stimuli to responses). Such top-down support is critical for bridging short delays or in the face of stronger behavioral tendencies (e.g., an overlearned or salient stimulus-response mapping that is usually adaptive but is contextually inappropriate). For participants to do well in the primrose path task, top-down support is useful and perhaps necessary to keep track of contingencies over time, and to resist choosing the option that is locally better but globally worse.

Cooperation in the prisoner's dilemma paradigm could also require cognitive control for similar reasons. Although the Braver and Cohen model has not explicitly incorporated social variables, lateral PFC mediates remarkably diverse control functions. Domain-general rather than domain-specific mechanisms are likely to be used for actively maintaining information, including information about other people. Patients with lesions to lateral PFC have gross impairments in both social and nonsocial behavior.

Finally, both paradigms appear to require not just bridging short delays and resisting a prepotent response, but also the integration of information (across time or individuals). Lateral PFC is critical for integration (Gray et al. 2001). Therefore, unless participants are responding by rote, decision making in both paradigms is likely to require cognitive control and lateral PFC function. This task analysis might seem to flatly contradict data showing that patients with lateral PFC damage were not impaired at decision making during the Iowa gambling task, whereas patients with medial PFC damage were impaired (Bechara et al. 1998). However, not all forms of decision making are identical. The Iowa gambling task assesses the ability to learn about discrete risks and rewards – which can be done associatively, that is, using stimulus-response learning, with no contextual dependence of the mappings and hence little need for cognitive control.

How might this task analysis be useful? Rachlin's argument makes a strong prediction: If a particular manipulation biases people to be altruistic, then it must also bias them to be self-controlled. Rachlin presents results showing that (1) manipulating reinforcement contingencies had similar effects on performance in both tasks, and (2) manipulating the context produced similar effects on both tasks. These results weakly support Rachlin's prediction: Many such experiments can be envisioned, and if even just one did not find identical influences of a given manipulation on both tasks, it would argue against a strong form of Rachlin's hypothesis; if they converge, it would further support it. To our knowledge, the following three methods have not been applied to investigate both the primrose path and prisoner's dilemma paradigms. We expect considerable but not perfect overlap.

First, functional brain imaging provides access to internal states that are critical for the control of behavior (e.g., as shown by lesion studies). In principle, two tasks can show identical behavioral effects of different manipulations and yet be very different in underlying mechanisms. If altruism requires self-control, then brain regions that contribute to cooperation in the prisoner's dilemma should also contribute to self-control in the primrose path. The paradigms are not identical in content, so different loci within lateral PFC could be activated. Activation should be sustained across trials but may not be event related.

Second, affective variables are important in many social and nonsocial forms of decision making. Pleasant moods increase the likelihood that people will spontaneously help others (Isen 1972). Pleasant mood enhances the perception that other people belong to one's social group (Dovidio et al. 1995). Thus pleasant emotion should promote cooperation on a prisoner's dilemma task, which it appears to do (Lawler et al. 2000). Does a similar effect of positive mood hold for self-control? Perhaps: Pleasant mood can help people delay gratification (Fry 1977). What about unpleasant mood? Stress and threat-related affect decrease self-control in the primrose path (Gray 1999), suggesting that unpleasant affect should increase selfish responding during a prisoner's dilemma.

Third, dual-task manipulations can be used to reveal whether a task requires cognitive control. If performance degrades when participants must perform another task concurrently, then the primary task requires control. Both Rachlin's hypothesis and the current task analysis predict that participants should be less self-controlled and less cooperative under dual-task conditions.

ACKNOWLEDGMENTS

Preparation of this commentary was supported by a grant from the National Science Foundation, BCS 0001908. We thank Deanna Barch, Len Green, and Yvette Sheline for their useful discussion of these issues.

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

The basic questions: What is reinforced? What is selected?

Patrick Grim

Group for Logic and Formal Semantics, Department of Philosophy, State University of New York at Stony Brook, Stony Brook, NY 11794-3750.
pgrim@notes.cc.sunyb.edu
http://www.sunysb.edu/philosophy/new/faculty/grim.html

Abstract: Any behavior belongs to innumerable overlapping types. Any adequate theory of emergence and retention of behavior, whether psychological or biological, must give us not only a general mechanism – reinforcement or selection, for example – but a reason why that mechanism applies to a particular behavior in terms of one of its types rather than others. Why is it as this type that the behavior is reinforced or selected?

A very basic question arises regarding virtually all attempts to explain behaviors, whether genetic or psychological – a very basic question regarding the individuation of behaviors.

Biologists attempt to explain emergence or retention of a behavior in terms of evolutionary pressure: that behavior appears or is retained because it is selected for (or, as in the case of genetic drift, is at least not selected against). Psychologists attempt to explain emergence or retention of a behavior by reinforcement: that is, behavior emerges or is sustained because it is positively reinforced (or at least not negatively reinforced).

But behaviors are not atomic items, nicely individuated by a universe cut cleanly at its joints. Any behavior, at any time, is part of larger patterns and larger types: a pattern of repeated instances of that behavior, a pattern of behavior of the same "kind" in different contexts or locations, a pattern of behavior over time, and a pattern of the "same" behavior shared among members of a wider community. When I butter this very piece of burnt toast at midnight in the pantry, of what single and particular *type* is my behavior? The answer, of course, is that there is no such single and particular type. My action is of many types: a scraping, a buttering, a nocturnal pantry bumbling, a pursuit of a surreptitious snack, a self-indulgence, a midnight lark. It is an instance of each of these types tightly indexed to me, and also an instance of each type spread across members of my community. If behavior is selected for by evolutionary pressures, precisely which of these behaviors is selected for? If my behavior is reinforced, precisely which of these behaviors is it that is reinforced?

No biological or psychological explanation of emergence or retention of a behavior can be complete unless it tells us what category is selected for or reinforced, and why it is that particular category that is important. This is one of the general lessons that Rachlin's more specific article points up: that within both evolutionary and psychological theories there are different options for the category taken to be selected for or reinforced, and the theory will give different results depending on the category chosen. Rachlin emphasizes plausible cases of self-control where it is a pattern of choices over time that must be taken as a significant unit for the individual, not just atomic choices at atomic times. He compares plausible cases of altruism where a pattern of choices over time or by members over a community must be taken as significant units, rather than just atomic individual choices at particular times. Rachlin favors a psychological explanation in each case because self-control and altruism can change over the course of a life in ways we would not expect of behaviors genetically programmed. But there is a very central point that applies to biological and psychological explanations for behavior, and that applies well beyond the specific case of self-control and altruism. Because atomic behaviors are always part of innumerable overlapping types, any theory of their emergence and retention must give us not only a general mechanism – reinforcement or selection, for example – but a reason why that mechanism applies to the behavior in terms of one of its categories rather than others. Why is it as this type that the behavior is reinforced or selected?

There is no reason to think that this must remain a question without an answer. In some cases there are reasons written in the molecular details explaining why selection operates on a particular category of heritage or behavior: Two traits stand or fall together under evolutionary pressure because they are genetically and inseparably coded together. In some cases there will be explanations in terms of reinforcement for why it is as *a particular* type that a behavior is reinforced: Some prior or more general pattern of reinforcement leads individuals to conceive of what is happening in certain terms, and that is why they respond to this particular pattern of reinforcement as they do. The important point is that we desperately need those more complete explanations for significant typing to get a more complete understanding of behavior, regardless of whether that understanding is biological, psychological, or a mixture of the two.

When one takes seriously the fact that behaviors come to us as members of overlapping types rather than as individual atoms, it also becomes clear that we may be wrong to look for the explanation of emergence or retention of a behavior in terms of just one of its types. Natural selection (or selections) may simultaneously act in different ways on a particular organ, an individual, and the group of which that individual is a part. It (or they) may also act simultaneously on a behavior in terms of its different types. For some reason we have conceived of natural selection as a single force and evolution as a single historical strand. It may be time to rethink them in the plural, as forming a braid of overlapping processes acting simultaneously on different categories at different levels. What holds for biological explanation will also hold for psychological: Here again we may be wrong to paint a picture of single behaviors with single explanations. A behavior may instantiate any of various types simultaneously, and patterns of reinforcement may operate differently on those different types.

It is the emphasis on reinforcement as acting on behaviors as a part of larger patterns or larger categories that I find most promising in Rachlin's piece. I consider his game-theoretic illustrations of the point compelling, and I have seen similar indications in my own simulation work (Grim 1995; 1996; Grim et al. 2002). I think much the same point can be pressed regarding natural selection in evolutionary explanations. But Rachlin is not always careful to distinguish this point from others. "Teleological behaviorism," as he outlines it, also carries a claim that the reinforcers for a behavior are somehow internal to the pattern: "[T]he reinforcers of real-life self-controlled and altruistic behavior are intrinsic in the patterns of those behaviors" (target article, sect. 7).That is a very different type of claim, of which I am not so convinced, and calls for a different kind of support.

So be good for goodness' sake

John Hartung

Department of Anesthesiology, State University of New York Downstate Medical Center, Brooklyn, NY 11203. **jhartung@downstate.edu**
http://hometown.aol.com/toexist/index.html

Abstract: Altruism is traditionally encouraged by promoting a goal, for example, going to heaven. In contrast, Rachlin argues that altruistic behavior can be sufficiently reinforced by the abstract intrinsic reward that comes from maintaining an unbroken pattern of altruistic behavior. In my experience, there are very few people for whom this is true. For fellow atheists and anti-theists, I suggest an alternative.

Why be good for goodness' sake?

Because he knows when you are sleeping, he knows when you're awake, he knows if you've been good or bad . . . and you risk getting a lump of coal instead of toys if your good-to-bad ratio is not good enough. Or, in some adult versions, because you will go to heaven instead of hell, or be reincarnated a rung up the ladder, or reach nirvana, and so on.

To his credit, Rachlin is having none of the above. Even more

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

to his credit, he is offering an alternative. He argues that being good for goodness' sake can be sufficiently reinforced by the abstract intrinsic reward that comes from maintaining an unbroken pattern of altruistic behavior – a reward that is analogous to, but profoundly greater than, the reward that some people get from doing crossword puzzles. This is real. It works for Rachlin and five people whom I have known personally – one in Ethiopia, two in graduate school, one where I work, and my brother in-law.

Therein lies the problem – only five habitual altruists. Mind you, those five people constitute a substantial portion of the best people I know, but as Rachlin asks, "Can the job be done . . ." this way? Can "real-life altruistic behavior" be developed in reprobates like myself who have been in Brooklyn too long to buy pie in the sky, and who do not appreciate the abstract intrinsic reward of being good for goodness' sake enough to be good for goodness' sake? Rachlin thinks it can, for two reasons. First, there is no genetic variance in propensity toward altruism; and, second, the latent sainthood that resides in all of us can be developed by applying established "real-life self-control techniques."

Unfortunately, altruism does appear to have heritability (e.g., Davis et al. 1994; Rushton et al. 1986), which suggests that some of us would need to be developed more strenuously than others. Still, great progress could be made if Rachlin's critically important dichotomy exhausted the real-world possibilities: "Habitual altruism is a happier mode of existence than habitual selfishness." Fortunately, this assumption is almost always true for humans above the age of two. Unfortunately, however, as Rachlin's own behavioral laboratory research demonstrates, a dangerous third rail powers many, if not most, personal engines. To wit: Be good when it is good to be good and be bad when you can almost certainly get away with it (i.e., when the probability of adverse consequences times the magnitude of those consequences is lower than the benefit of bad behavior).

The difference between habitual self-control and habitual altruistic behavior is that the extrinsic rewards of self-control follow from the exercise of self-control (having not had a drink since 1:30 AM, February 12, 1987, I know all about this), whereas the abstract intrinsic rewards of maintaining a pattern of altruistic behavior only follow from that behavior if the principled, habitually altruistic actor has formulated an array of intrinsic reward receptors that are sensitive to that behavior – a Catch-22 circumstance that reminds me of a line from an old Ray Charles tune, "If you gotta have something before you can get something, how you get your first is still a mystery to me" (Charles 1962).

What would cause a person to have a sufficiently sensitive set of intrinsic reward receptors for altruism-all-the-time to engage in such behavior? Rachlin has not addressed this issue, but common observation reveals that relief from fear of death has been the most reliable motivator – which brings us back to the problem that such relief has only been widely sought through appeasement of adult versions of Santa Claus and religious fantasies of afterlives administered by such gods . . . with bonus points for putting money in the offering plate and a world of trouble between differing Santa-ologies (Hartung 1995).

When asked to define a saint, Martin Luther reportedly replied, "A saint is a person who understands the egoism in his every motive."[1] Rachlin has argued, in distinction, that a pattern of socially beneficial behavior, even if consistently maintained across a lifetime, is not altruistic if it is egoistically motivated. Taking some of the wisdom that can be found in both of those understandings, we need an engine for good behavior that is fueled by long-range self-interest. We also need an objective against which to judge the goodness of behavior. After all, running into a burning building to save someone else's child could be socially irresponsible if the rescuer has four dependent children at home.

"These are the times that try men's souls" – more so than when Thomas Paine wrote those words (Paine 1776), because more and more people cannot take the prospect of eternal life quite seriously. The best substitute for eternal life that I have been able to take seriously is the prospect of having eternal consequence on future life (Hartung 1996). What if our collective good-to-bad ratio will determine whether we become a critical link in the evolution of descendants who will evolve forever?

What if the universe is not slated to contract into a black hole or expand into nothing? What if people who spend time mapping imagined areas of dark matter are providing modern examples of the extremes to which people will go to fabricate evidence in defense of cosmological dogma? Recent work by Fred Hoyle and colleagues opens the door for a quantum leap in the prospect of a permanently inhabitable universe (Hoyle et al. 2000). My guess is that a visionary cosmologist will come along within twenty years to do for the Quasi-Steady-State Universe what Dyson (1979) and Frautschi (1987) have done for the increasingly implausible Big Bang universe (Hartung 2001).

Goal-oriented moral behavior works on a mass basis as long as the motivating goal is perceived as worthy and attainable. Thus far, more people have faith in heaven than evolution, but these are philosophically trying times because reality is gaining respect at an accelerating rate. I offer eternal consequence as a basis for moral behavior on the off chance of adding a few goal-oriented altruists to the additional habitual altruists that Rachlin's contribution may foster.

NOTES

1. I heard this quote in a tape of a lecture delivered by Robert Oden, president of Kenyon College. Despite much searching, I have not been able to find the quote in Luther's writings, nor have I been able to obtain it from Oden. Try odenr@kenyon.edu.

Reinforcement stretched beyond its limit

Robert A. Hinde

St. John's College, University of Cambridge, Cambridge, CB2 1TP, United Kingdom. **rah15@hermes.cam.ac.uk**

Abstract: The concept of "intrinsic reinforcement" stretches the use of "reinforcement" beyond where it is valuable. The concept of the "self-system," though fuzzy at the edges, can cover experience as well as the behaviour of altruistic acts.

Professor Rachlin argues that altruistic behaviour can be accommodated within reinforcement theory. He supposes that, while individual altruistic acts may be of no or negative value, they are maintained by "intrinsic reinforcement" as parts of series of such acts. Drawing a parallel with group selection theorists who postulate more than one level of selection, Rachlin postulates more than one level of reinforcement – "reinforcement of particular acts and reinforcement of groups or patterns of acts" (target article, sect. 3).

The choice of a reinforcement model reflects the author's academic inclinations in the shape of a preference for a behavioural approach rather than a cognitive one – although he concedes that both should be pursued. But the choice requires him to depart from classic reinforcement theory and to subscribe to "teleological behaviourism." To most psychologists the postulation of an *intrinsic* reinforcing effect just because an act is repeated is little better than postulating an instinct for eating just because an organism eats. The old discussions on functional autonomy tried hard to avoid this morass (Bindra 1959).

The pursuit of an entirely behaviouristic approach leads Rachlin to adopt such concepts as "commitment," used descriptively, and "intrinsic reinforcement," used as an explanation. The latter term undermines all that is useful about the Skinnerian approach. And Rachlin's insistence that he is using a consistently behavioural approach inevitably lapses from time to time – such as when he explains the cooperative behaviour of laboratory subjects by saying, "They have decided to cooperate in life and continue to do so in the experiment" (sect. 5), or the behaviour of some soldiers under fire as "rule governed" (sect. 9). Accepting that humans can

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

learn rule-governed behaviour, does reinforcement provide a fully adequate explanation of all that is going on in the process?

If such language is used, why not use a cognitive model that makes such lapses unnecessary? The concept of "self-system" can cover the phenomena with which Rachlin is concerned and much else besides, and recent studies of the "self" (Baumeister 1999) are generating an approach which is as hard-headed as teleological behaviourism. Rachlin uses the "self" only as "existing contingently in a series of temporal intervals during which behavior occurs in patterns" (sect. 4), a model that he suggests would imply that people's selves evolve and change. But of course they do. It is well established that the "self" (in the sense of a self-description) changes with age and context (Harter 1998; McGuire & McGuire 1988). Rachlin seems unaware of recent developments in this field.

Self-descriptions include references to moral precepts ("I try to be honest"). It is a not unreasonable suggestion that what we describe as "conscience" involves comparison between past, present, or intended action and the moral code incorporated in the self-system (Hinde 2002). This is not incompatible with Rachlin's description of altruism – "What is highly valued is a temporarily extended pattern of acts into which the particular act fits" (sect. 1.2). Even Rachlin postulates a "coherent sense of self" for the maintenance of altruism. This is entirely in keeping with the results of studies of extraordinary individuals who served as exemplars to many. Such individuals had great certainty about the decisions they made, as though morality were completely integrated into the self-system and altruistic actions involved no conflict with other personal motivations (Colby & Damon 1995; Youniss & Yates 1999). Indeed, studies of personal relationships in the context of exchange theory suggest that being overbenefited as well as being underbenefited can provoke compensatory behaviour (e.g., Prins et al. 1993). If this is confirmed, it indicates that not only a moral code but a social contract is incorporated in the self-system.

How are moral issues incorporated? Young infants show a great deal of proto-prosocial behaviour – sharing, caregiving, showing sympathy, and so on (Rheingold & Hay 1980) – as well as selfish or egocentric behaviour. Rachlin rejects the idea of an inherited mechanism for altruism, but the evidence points to a predisposition to learn to please caregivers. Developmental psychologists have shown how prosocial tendencies are moulded through relationships, especially those within the family, and thus come to form part of self-descriptions (Turiel 1998). Yes, of course reinforcement plays a part, but broad moral precepts, like "You should protect others from danger," may be incorporated even if the opportunity to act on it has never occurred. There is no denying that reinforcement may play some role in the genesis and maintenance of (apparent) altruism. Minor altruistic acts often receive nods of approval and may contribute to the actor's status (see discussion of meat-sharing by hunter-gatherers, Hawkes et al. 2001). And reinforcement, if used in a strict Skinnerian sense, is a more hard-headed concept than the "self-system," which still has fuzzy borders. But reinforcement loses its edge if it is pushed beyond its limits. The use of the "self-system" as something more than an intervening variable but perhaps not quite a hypothetical construct (MacCorquodale & Meehl 1954) can embrace not only the behaviour but also the experience of individuals.

Toward a better understanding of prosocial behavior: The role of evolution and directed attention

Stephen Kaplan[a] and Raymond De Young[b]

[a]*Departments of Psychology and of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109;* [b]*School of Natural Resources and Environment, University of Michigan, Ann Arbor, MI 48109.*
skap@umich.edu **rdeyoung@umich.edu**
http://www.snre.umich.edu/~rdeyoung

Abstract: Rachlin's thought-provoking analysis could be strengthened by greater openness to evolutionary interpretation and the use of the directed attention concept as a component of self-control. His contribution to the understanding of prosocial behavior would also benefit from abandoning the traditional (and excessively restrictive) definition of altruism.

Discussions of altruism routinely exclude from consideration any behavior from which the actor receives pleasure or other benefit. That Rachlin adopts this traditional approach is understandable but unfortunate. In a perceptive and inadequately appreciated analysis, Wallach and Wallach (1983) point out that there are two distinct meanings of self-interest (or, in their terms, "egoism"). In their example, you can be motivated by helping someone because you expect something in return or because "the other person's relief from distress or the other person's happiness is itself what you want to achieve and what would make you happy" (p. 201). As they point out, the two situations are equally self-interested only in the most trivial sense. Yet it is this trivial sense that studies of altruism call upon when they use the traditional definition, that is, that one is acting against one's self-interest. The result is that the enormously important topic of what motivates prosocial behavior tends to be neglected in favor of a focus on special and atypical cases.

Even accepting this limitation, however, Rachlin's argument is flawed by his determination to eliminate the potential role of evolution as a component of altruism. This commitment harks back, unfortunately, to an earlier era, when a then-dominant behaviorism argued that the existence of a behaviorist explanation demonstrated that all other explanations were irrelevant. This notion that an explanation at one level usurps the possibility of a useful explanation at another level has been sufficiently pervasive to have received several colorful appellations, such as "nothingbutism" and MacKay's (1965) more elegant "fallacy of 'nothing buttery'." This way of thinking is no more acceptable now than it was then; if indeed there is a demonstrable role of habit in altruism, this in no way eliminates the possibility that there is a role for evolution as well.

A particularly interesting component of Rachlin's discussion is his use of intrinsic motivation. It is also, however, a topic where a bias against an evolutionary perspective is a serious handicap. His interpretation of intrinsic motivation as arising from a string of habits is less than convincing. The fascination with crossword and jigsaw puzzles seems far more likely to be an expression of the human inclination to solve problems, a tendency humans share with nonhuman primates (Harlow 1953). The very widespread character of this motive strongly suggests its evolutionary origins.

Closely related is Rachlin's argument that "most of us would indeed choose to be heroes rather than cowards" (sect. 9, last para.). His explanation for the origin of this motivation is not clear. A fairly straightforward explanation arises from Campbell's (1975) suggestion that humans innately have both self-interested and social motivations, and Goldschmidt's (1990) impressively documented argument that the inclination to work for the respect of one's fellows is a central component of human nature. In fact, much prosocial behavior may well be traceable to the way in which cultures use respect as a reward for such behavior. This also provides a nice example of the way in which an innate inclination could provide the leverage for a great deal of learning. Far from being in conflict with an explanation based on learning, the evolutionarily based motivation would be what makes the learning possible.

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

Rachlin is undoubtedly correct in his assertion that self-control is learned. However, his analysis would be strengthened by including the role of inhibition in managing our behavior. Inhibition is essential to self-control. Without the ability to inhibit the effect of the immediate environment, long-term goals cannot possibly affect present behavior. If people were unable to inhibit any stimuli, essentially forced to attend and respond to every next thing that the environment presented, then contemplation, recollection, and behavioral continuity, necessary for all of Rachlin's examples, would be unattainable.

Rachlin supposes that self-control is accomplished by an innate learning mechanism. Yet such a mechanism would be unable to inhibit immediate stimuli so as to allow a longer-term pattern to come into play. The mechanism of self-control involves more than just learning habits; there is also the need to direct one's attention. Directed attention (Kaplan 1995; Mesuluam 1985) is useful in dealing with just the sorts of short-term versus long-term behavioral choices that Rachlin sets up: inhibiting the power of the immediate environment so as to allow consideration of less salient but nonetheless valued patterns. Directed attention allows for a variety of prosocial behaviors (e.g., pursuit of an important social goal despite interesting competition in the immediate setting, helping others despite unmet personal needs, and resisting temptation to maintain devotion to a larger pattern).

A re-analysis of Rachlin's examples offers some insight on the role of inhibition. The example of a woman entering a burning building is ambiguous because many of the stimuli present (e.g., onlookers screaming that someone is trapped, a child's scream for help) are both involuntarily fascinating (James 1892) and conceivably capable of prompting a short-term pattern (e.g., entering the building) that is closely linked with the longer-term pattern of prosocial behavior. The behavior of a recovering alcoholic better demonstrates the enormous adaptive advantage offered by inhibition. Here the environmental stimuli conspire mightily against sobriety. Yet, the recovering alcoholic's self-control is only possible because of the ability to hold the immediate environment at bay and the insertion of cognition between stimulus and response.

The desire not to have to use self-control is a most interesting and useful contribution which fits well with the recent work showing that directed attention is a scarce and labile resource. When under continual demand, our ability to direct our inhibitory process tires, resulting in a condition called directed attention fatigue (DAF). This condition reduces mental effectiveness and makes consideration of abstract long-term goals difficult. A number of symptoms are commonly attributed to this fatigue: irritability and impulsivity that results in regrettable choices, impatience that has us making ill-formed decisions, and distractibility that allows the immediate environment to have a magnified affect on our behavioral choices. The symptoms of DAF can be summarized as a reduced ability to make and follow plans, and the inability to mentally restrain impulsive thought or action. In short, DAF makes prosocial behavior at any temporal scale less likely.

We would thus like to commend Rachlin for his fascinating treatment of the problem of long-term versus short-term interests, for his focus on self-control (and its limitations), and his linking all this to prosocial behavior. At the same time, we would encourage him to consider evolutionary perspectives less extreme than those he has apparently been reading, and to explore the possible role of directed attention as a useful tool in his further exploration of the self-control concept.

Is the prisoner's dilemma metaphor suitable for altruism? Distinguishing self-control and commitment from altruism

Elias L. Khalil

Behavioral Research Council, American Institute for Economic Research, Great Barrington, MA 01230. **elk@aier.org** **http://www.brc-aier.org**

Abstract: Rachlin basically marshals three reasons behind his unconventional claim that altruism is a subcategory of self-control and that, hence, the prisoner's dilemma is the appropriate metaphor of altruism. I do not find any of the three reasons convincing. Therefore, the prisoner's dilemma metaphor is unsuitable for explaining altruism.

Rachlin claims that altruism is a subcategory of self-control, known also as the precommitment or weakness-of-will problem (Elster 2000). I can surmise three separate reasons behind this unconventional claim.

1. For Rachlin, self-control and altruism share one element, namely, that a single action has no meaning. The action has to be part of a pattern, which may be reinforced, to provide the context. The context allows one to judge whether the action is self-control, as when an alcoholic prefers soft drink over scotch, and whether the action is altruism, as when we see a woman running into a building on fire and emerging with a child that is unrelated to her biologically. Rachlin calls his approach "teleological behaviorism" and relates it to Aristotle's concept of habit as motivated by final causes. He characterizes it as "complex ambivalence" to show the repetitiveness of the acts – in opposition to the work of Platt (1973) and Messick and McClelland (1983). Context matters: The alcoholic might have made his choice because his boss is watching; the woman might have gone back into the building to salvage her jewelry when she stumbled on the child.

I have three reasons to doubt the issue of pattern so emphasized by Rachlin. First, Rachlin himself in many places mentions that altruism is a single act that does not need to be reinforced – in the woman-running-into-burning-building example, it's possible she might not exit alive from the burning building. He states that "an altruistic act is defined as a choice of the *t*-length fraction of the longer activity over the brief activity under Conditions 1, 2, and 3" (target article, sect. 4). If so, it is a single act. Second, even if Rachlin consistently defines altruism as a pattern like self-control, why should the observation of a pattern of apparently series acts of altruism or series acts of self-control make us sure that what we are observing is altruism or self-control? It is possible that each time one observes the woman doing what appears to be an altruistic act, she has her own private reason; or each time one observes the alcoholic abstaining from scotch, he has his own reason. The consistency of the pattern does not make one more certain of the motive – the desire of the actor to please the observer might be the motive. So, if you are a behaviorist, the logical problem of deducing the motive behind a pattern of acts is no less problematic than deducing the motive in a single act. Therefore, why resort to the idea of a pattern, when you have to end up asking the person anyhow if she is an altruist or if he abstains from alcohol? To ask the agent is fraught with problems – which do not go away with the idea of a pattern. Third, even if the pattern idea is a crucial element for deciding on altruism and self-control, this hardly makes them similar enough to justify the use of the same prisoner's dilemma (PD) metaphor for both. Bats and birds both fly. This does not make the bat's forelimbs wings. The sharing of the pattern feature at best suggests a heterologous metaphor – not a unificational or even a homologous similarity to justify the use of the same conceptual machinery (Khalil 2000).

2. Although Rachlin emphasizes the pattern issue, it is not the only similarity he finds. He argues that the "particular components of an altruistic pattern, like those of a self-controlled pattern, are less valuable to the actor than are their immediate alternatives" (sect. 4). That is, the alcoholic finds that one instance of abstention from scotch is less valuable than its immediate alternative

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

(drinking scotch). Likewise, for Rachlin, one instance of altruism is less valuable than its alternative (selfishness). In the case of alcoholism, the agent benefits from defecting (drinking scotch) in a game he is playing with his future self. So, if the future agent abstains, the present agent who surrenders to a weak will would have acted optimally. The problem is that the present self does not know if the future self will defect also – the core structure of the PD problem. The agent in the *single* altruistic act, however, does not like to defect in the first place. It would not be optimal. If it costs the benefactor little trouble to give a charitable donation, he would be acting optimally if the beneficiary experiences enormous happiness. I agree with Rachlin that one does not need to appeal to the inclusive fitness hypothesis or its equivalent in economics (Becker's 1981 theory) to explain altruism (Khalil 2001). But the single act of altruism is not similar to the single act of cooperation in the self-control problem: While the agent maximizes a utility function in altruism, the agent maximizes utility function in defection.

3. To achieve his goal and make the single act of altruism look like the single act of cooperation in self-control, Rachlin invokes a third condition, social cooperation – what is known as the commitment or trustworthiness problem surrounding the PD game (Khalil, in press). Given that self-control resembles commitment substantially, Rachlin would achieve his goal if he shows that commitment resembles altruism. But commitment does not resemble altruism for three reasons.

First, altruism is not a game to start with. Who is the other player? Rachlin invokes the community, which explains his discussion of group selection. In altruism, the agent is weighing costs and benefit vis-à-vis someone else who is passive. In contrast, in PD, the agent is weighing costs and benefits vis-à-vis himself, because the other player is not passive. The payoff in altruism is predictable, while the payoff in PD is strategic. Second, in altruism we have redistribution of same resources, while in PD we have an absolute reduction of resources because defection is the dominant strategy. That is, as measured by the production of public good, the collective is not hurt if there is no altruism, but the collective is hurt if there is no cooperation. Third, altruism is a voluntary action, while commitment involves paying one's debt or (in the case of self-control) promise. One may or may not want to give money to the victims of floods in a far-away country. But one is obligated to pay for the local fire department (Khalil 1999; 2002; Fehr & Gächter 2000).

In light of the preceding arguments, the prisoner's dilemma metaphor, which might be suitable for the self-control or commitment problem, is unsuitable for the altruism problem.

Adaptive altruistic strategies

Dennis L. Krebs

Psychology Department, Simon Fraser University, Burnaby, BC, V5A-1S6, Canada. **krebs@sfu.ca**

Abstract: Biological, cognitive, and learning explanations of altruism, selfishness, and self-control can be integrated in terms of adaptive strategies. The key to understanding why humans and other animals sometimes resist temptation and sacrifice their immediate interests for the sake of others lies in mapping the design of the evolved mental mechanisms that give rise to the decisions in question.

Explaining altruism has been a perennial challenge in the biological and social sciences. Rachlin weighs in with the idea that although particular altruistic acts may be immediately costly to those who perform them, the patterns of behavior, or habits, of which they are a part may pay off better in the long run than more selfish patterns of behavior, or habits. An attractive feature of the principle implicit in this explanation is that it also accounts for self-control. In this commentary, I do not quarrel with the basic idea that Rachlin advances, but I do take exception to the contention that this idea is better housed in the explanatory framework of teleological behaviorism than in biological and cognitive explanatory frameworks.

I believe our understanding of altruism will be advanced best through models that integrate the contributions of different theoretical orientations. Let me take a stab at outlining one. The patterns of behavior, or habits, central to the author's explanation of altruism can be conceptualized as strategies. Assume that individuals (human and other animals) inherit mental mechanisms that give rise to selfish and altruistic strategies. Such strategies operate in terms of "if-then" decision-rules. Viewed in these terms, operant conditioning would involve a strategy such as "repeat acts that were followed by a reward and suppress acts that were followed by a punishment." Contexts are important because they contain the "if" conditions. Assume that different individuals inherit proclivities to adopt different strategies, and that developmental experiences calibrate strategies and shape them into habits.

Assume that the mental mechanisms and the strategies and decision-rules they contain were designed by natural selection. Strategies that most effectively propagated the genes that designed them evolved. Note that the payoffs in question are long-term ultimate payoffs of strategies, reckoned over the life span of individuals. Individuals who delay gratification and sacrifice their immediate interests for the sake of others may well come out ahead of more self-indulgent and selfish individuals in the long run. Note also that the payoffs do not necessarily equate either to survival or to happiness. Surviving or happy individuals who fail to propagate their genes count for nothing in evolution. Finally, note that the payoffs in question and the strategies they designed were selected in the distant past in environments that differed in significant ways from those of the modern world.

I believe this conception of evolved strategies is equipped to account for the evidence the author adduces in support of the teleological behaviorism model he prefers, and more. Evolved strategies account for the biological and behavioral compatibility of selfishness and altruism, but in a different way from the author's version. Consider the woman who runs into a burning building to save someone else's child. There are two ways in which the strategy underlying such an act could have been adaptive (i.e., reinforced biologically) in ancestral environments. First, most children in the relatively small groups formed by our hominid ancestors comprised kin. Although helping a child who shares one's genes may be altruistic on a behavioral level, it is selfish at a genetic level. Second, the act could have been reinforced through reciprocity. Note that in this line of thought, we would expect individuals to invoke the strategy only in certain "if" conditions. For example, the higher the probability the child in question is related to the woman, the less the chance of harm to the woman, and the greater the probability of fitness-enhancing rewards (to self or kin), the greater the probability of emitting the act.

The strategies in question could have evolved through group selection in essentially the same way they could have evolved through kin selection. There is no inconsistency between the principles of evolution and sacrificing one's interest for members of one's group with whom one shares genes. And there is no reason why group-selected strategies could not be shaped by experience over the life spans of those who inherit them. Indeed, it is quite clear that inherited strategies do change with experience. As one example, the strategies invoked by young children are quite different from the strategies invoked by aspiring mates and doting grandparents.

The prisoner's dilemma strategies described by the author have been found to be winning strategies in game theory models of evolution. Strategies like tit-for-tat are as fruitfully conceptualized in terms of decision rules such as "open with a cooperative response and copy the subsequent responses of your opponent," as they are in terms of reinforcement theory. The important lesson learned from game theory models of evolution is that although selfish strategies cannot be beaten on any one move, cooperative strate-

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

gies such as tit-for-tat and Pavlov that enable individuals to cut their losses in exchanges with selfish players and reap the benefits of cooperating with cooperators may pay off in iterated games in the long run. Note that these strategies are commonly considered cooperative, not altruistic.

With respect to commitment, the more one has invested in a strategy, the greater the potential costs of switching. This said, it is in individuals' interest to abandon losing strategies. Thus, for example, we might expect a reformed alcoholic to revert to drinking if sobriety fostered depression.

The author argues that "the crucial issue is whether or not altruism is a subcategory of self-control . . . there is no need to postulate an innate altruistic mechanism; the job can be done by . . . an innate learning mechanism" (sect. 8). I disagree for two reasons. First, an evolved strategy could give rise to both self-control and altruism as easily as an innate learning mechanism could. It doesn't really matter what you call the mechanism. Second, although it is possible to define altruism in a way that requires self-control, I am not sure what is gained by viewing it as a subcategory of the process. The key to the selection of altruism is not, in my view, self-control; the key is the adaptive benefits of the overriding strategy in question.

Why cooperate? Social projection as a cognitive mechanism that helps us do good

Joachim I. Krueger and Melissa Acevedo

Department of Psychology, Brown University, Providence, RI 02912.
Joachim_Krueger@brown.edu **Melissa_Acevedo@brown.edu**
http://www.brown.edu/Departments/Psychology/faculty/krueger.html

Abstract: The mother sacrificing herself while rescuing someone else's child is a red herring. Neither behaviorism nor cognitivism can explain it. Unlike behaviorism, however, the cognitive process of projection can explain cooperation in one-shot social dilemmas.

Making the case for teleological behaviorism as an explanatory framework for altruism and other forms of selfless cooperation, Rachlin "does not deny the existence of [other, cognitive] mechanisms," but he considers it "unnecessary to postulate the existence of such a general mechanism" (target article, sect. 1.1). This view is unremarkable unless one takes it to mean that teleological behaviorism is the better explanation because only its mechanisms offer a *necessary and sufficient* explanation of altruism. We think that the case for the sufficiency of teleological behaviorism has not yet been made, and we offer an example of a sufficient cognitive mechanism.

The Mother running into a Burning House (MBH) to save somebody else's child while risking her own life is the paradigm of altruism throughout the article. Any ambitious theory of altruism must attempt to explain such extraordinary behavior because everyday acts of altruism are readily explained away by some lurking self-interest. It is only fair to ask whether teleological behaviorism rises to the challenge. The explanatory tale is that some people have been collecting delayed or long-term rewards for altruism or other forms of self-controlled behavior. As a result, they have formed an enduring commitment, motive, or habit of extending this pattern of behavior into the future.

The case of the MBH poses a problem. One must assume that the individual differences in habit strength or commitment are highly reliable and transferable to new situations. Unfortunately, individual differences in personality, of the type assumed here, emerge as usable predictors only after massive aggregation across situations. Psychometricians consider predicting individual acts a near-hopeless enterprise. Darley and Batson's (1973) study of Good-Samaritanism is a classic example of how psychometrics failed to predict who would help. Beyond its rarity, the case of the MBH is complicated by its extremity. It is difficult to find a class of acts with which it can be categorized. What are the charitable behaviors that shaped the habit that is now being activated? Suppose the woman had a routine of taking the neighbors' kids to the bus stop. This habit may well have been shaped by mutually reciprocated cooperation over time, but can it now be considered the cause for the woman's self-sacrifice? To suggest that it can puts credulity to the test, especially when no theoretical, empirical, or quantitative lever is offered as a guide. While teleological behaviorists and the parents of the saved child may respectively see a good habit and saintliness at work, the woman's own family may feel rather differently. Indeed, the perspective of the woman's family would probably best predict how the woman herself would feel when confronted with the existential challenge of a burning house. By casting the self-sacrificial rescue as an act of self-control, teleological behaviorism must ask which base and self-defeating impulse is being kept at bay. It would appear to be fear of death, which begs the question of what kind of learning history prepares one to scoff at death. Perhaps there is none, and that's why women with little children are particularly hesitant to die for the children of others, whatever their altruistic commitments might be otherwise.

Cooperation in the prisoner's dilemma game (PD) is far more common. Although its prevalence makes it more tractable psychometrically, cooperation depends at least as much on the perceived personality of the opponent than on the player's own personality (de Bruin & van Lange 1999). Most disturbing is the finding that once players learn *that* their opponent has either cooperated or defected, almost all defect. They cooperate only as long as they do not know *whether* the opponent cooperates (Shafir & Tversky 1992). If habit and commitment were such strong forces, why should uncertainty matter?

One answer lies in the cognitive mechanism of projection, which Quattrone and Tversky (1984) first applied to the social dilemma of voting, and which Baker and Rachlin (2001) introduced to the PD. Projection is a generalized expectation that others will reciprocate whichever course of action one chooses. Thus, cooperation increases with the perceived probability of reciprocation. When projection is perfect, the PD devolves into a choice between the payoffs for mutual cooperation and the payoffs for mutual defection. The dilemma disappears, and the player can cooperate out of self-interest. Projection can be learned, but such learning is not necessary. The expectation that others will act as we do may well be an adaptive default handed down by evolution. If anything, gradual learning about how others actually behave reduces rather than enhances perceptions of similarity (Krueger 1998).

Neither teleological behaviorism nor projection can explain the MBH. Projection can, however, parsimoniously explain why many people cooperate even in the one-shot PD when they do not know what the opponent will do, but defect when they know what the opponent did. Teleological behaviorism would have to appeal to commitments that are conditional on uncertainty, in which case they would not be terribly sincere as commitments go.

Teleological behaviorism and altruism

Hugh Lacey

Department of Philosophy, Swarthmore College, Swarthmore, PA 19081.
hlacey1@swarthmore.edu

Abstract: Rachlin shows that experiments about social cooperation may fruitfully be grouped with experiments on self-control, and that this suggests interesting possibilities for practical behavioral controls. The concepts of *selfishness* and *altruism,* however, that inform his theorizing about these experiments, do not serve to provide understanding of the behavior that commonly is referred to, derogatorily, as selfish.

A core thesis of Rachlin's teleological behaviorism is that "mental terms" – these include common value and intentional terms – re-

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

fer to observable patterns of temporally extended overt acts of a person (Rachlin 1994; 1995a). I have argued elsewhere (Lacey 1995a; 1995b) that substantiating this thesis requires actually showing, in real-life as distinct from experimental situations, how such patterns may be identified (if they exist) without essentially deploying intentional terms to do so. Rachlin's current arguments about selfishness and altruism, although they involve a rich development of teleological behaviorism and suggest potentially fruitful applications for behavioral controls, do not avoid this criticism.

For Rachlin *selfishness* and *altruism* are compatible; indeed, conceived with the appropriate abstractness, *altruistic* acts are a species of *selfish* acts. The moral sting that commonly accrues to "selfish" is thus removed. A *selfish* act is simply one that has reinforcement value or that is part of a temporally extended pattern of acts that has reinforcement value. (I italicize Rachlin's uses of *selfish* and *altruistic* to keep them clearly separate from the morally laden uses discussed below.) All acts are *selfish,* and their variety and the principles that govern them are the object of investigation. Paraphrasing Rachlin, an act (A) of a person (X) is defined to be *altruistic* if and only if X chooses to do A, where A is part of a longer activity or a pattern of behavior (L), rather than A' (a brief activity of the same duration as A), and (1) X prefers L to *n* repetitions of A' (where the duration of L is *n* times as long as that of A'); (2) X prefers A' to A; and (3) a group (to which X belongs) benefits more when X chooses A rather than A'. He adds that A, *qua altruistic* act, is reinforced only insofar as L constitutes a pattern that is "intrinsically valuable" or (in accordance with Premack's theory of reinforcement) itself functions as a reinforcer; and also that, whereas all acts of choice are reinforced, an *altruistic* act may never be reinforced directly.

Because for Rachlin an act is *self-controlled if* it is chosen rather than another under Conditions 1 and 2, it follows that *altruistic* acts are a subset of *self-controlled* acts. He is thus able to bring the experimental studies on social cooperation that he discusses under the same umbrella as the experimental studies on *self-control* for which he is well known. This is a significant theoretical gain. It also justifies his claim that *altruism,* like *self-control,* can be brought under the control of contingencies of reinforcement and lends credence to his expectancy that "the same behavioral laboratory studies that have proved useful in developing real-life self-control techniques may be equally useful in developing real-life altruistic behavior" (sect. 7, last para.). These are important conclusions that stand regardless of the criticism that follows.

Rachlin is ever alert to draw extrapolations from experimental studies to real-life settings, not only about the possibility of identifying behavioral techniques of potential practical utility, but also about how to deploy his experimentally-derived principles to explain behavior in ordinary social settings (e.g., the woman and the burning building). With the latter – albeit as an "assumption of theory" or "method of procedure," not as an "empirical finding" – he intends to undermine the common moral connotations associated with selfish. If all acts are selfish it can hardly reflect a vice to act selfishly; that he admits different senses of selfish, does not change this. Rachlin proposes as an assumption of teleological behaviorism that, in principle, its categories can be successfully deployed to offer explanations of those acts in ordinary life that we commonly grasp with the aid of the morally laden categories. I concur that if the explanatory roles of selfishness and altruism were to be displaced (which they would be if there were good reason to hold that all acts of choice are *selfish*), then their moral sting would indeed have been removed. However, if they are not displaced, the compatibility of *selfishness* and *altruism* has no implications at all for common moral discourse and practice. I question whether Rachlin's extrapolations from experimental settings to explanations of behavior in ordinary social settings can be sustained.

In common moral discourse, X performs a selfish act when he acts for his own gain without paying due attention to, and perhaps undermining, the legitimate interests of others. I contrast selfish acts with "other-regarding" acts that, at their noblest (altruism), may manifest such values as love, compassion, and justice, even to the extent that these values may become thoroughly embodied in the trajectory of X's life, or come to the fore in times of crisis, so as to subordinate the value of personal survival itself. Altruism so understood is not opposed to the quest for enhanced personal well-being; the self and "others" are not opposed, for the self is conceived of in relationship to others, as participating in a common project with (perhaps selected) others. Acts that are altruistic typically are also acts that serve to enhance personal well-being; but they are not selfish (Lacey & Schwartz 1996). In this discourse altruism and selfishness are incompatible, and generally selfishness is morally reprehensible.

It is clear, I think, that acts of *self-control* and *altruism,* while by definition *selfish,* need not be selfish. Are altruistic acts *altruistic*? I have characterized altruistic acts as those that manifest to a high degree such values as love, compassion, and justice. Are they also parts of temporally extended patterns that are intrinsically reinforcing, where the patterns are complex but in principle observable (like the sequences of acts in social cooperation experiments)? What kind of pattern would the act of the woman entering the burning building fit into? I have no objection to attempts to further the project of teleological behaviorism. But, unless such patterns are identified in real-life situations, our common moral categories remain secure. The burden is on teleological behaviorism to identify the patterns. My hunch is that any patterns that can be identified can only be described with the use of common value and other intentional categories. This is not sufficient for Rachlin's categories to displace the morally laden uses of selfish.

An economist's perspective on altruism and selfishness

David K. Levine

Department of Economics, University of California at Los Angeles, Los Angeles, CA 90095. **dlevine@ucla.edu** **http://levine.sscnet.ucla.edu**

Abstract: Few disagree that altruism exists. The frequency and source of altruistic behavior remain mysterious, however.

Rachlin argues that altruism arises, as does self-control, as a kind of habit. He concludes with the observation that "Once we abandon case-by-case decisions, there will come times in choosing between selfishness and altruism when we will be altruistic even at the risk of death" (target article, sect. 10). Few people, even economists such as myself, would disagree with this observation. The key questions are how many people would do so, and why? Economists and biologists generally view preferences as instrumental, meaning that there must be some benefit to the individual. Rachlin takes as a basic example the "woman who runs into a burning building to save someone else's child." This is not a terribly good example from his point of view, because such an act involves a decent chance of survival followed by a substantial reward. A better example might be the "woman who throws herself on a hand grenade to protect her fellow soldiers," because such an act involves the protection of unrelated individuals, and a near certainty of death, and so no appreciable chance of any future reward. Of course, while there are recorded instances of individuals sacrificing their own lives to protect their comrades, there is little statistical data on how common this is, so we do not know if this is likely or unlikely to occur. It is strange also, in discussing hypothetical examples of this sort, not to acknowledge the widespread belief, present in most religions, in the afterlife. While it may be difficult to verify the presence of a reward after death, it is the belief in the reward that counts, not its verifiability. Since divine rewards have little social cost, one could argue that this is an efficient way of inducing altruism.

Regardless of frequency, altruistic behavior does exist and so demands explanation. The explanation offered by Rachlin is that such behavior is a part of a pattern of acts that is ordinarily re-

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

warded, but occasionally has a negative payoff. In other words, we are generally altruistic because we are rewarded for being so, and consequently develop the habit of altruism, which persists even in instances when it may not be a good idea. This is a plausible idea, if not a new one. The two central theoretical objections would be that it is not apparent why it is so difficult to develop a more complex habit – be altruistic most of the time, but not when it is obviously a very bad idea (such as the hand grenade example). This would seem to have much greater survival value than the simple habit. The second objection is that it is a theory of altruism through miscalculation. That is, a more perfect individual capable of carrying out more complex habits would not be so altruistic – our altruism simply reflects our imperfection as calculating machines. One might expect, based on this theory, that as people become more educated and have more leisure time to reflect on the value of different alternatives, they would behave less altruistically, that is, make fewer miscalculations. In particular, one would expect richer and more highly educated individuals to be less altruistic. Is there evidence that altruism is inversely related to wealth?

The second central point raised by Rachlin is the analogy between altruism toward others and toward one's future self. This is illustrated with the "alcoholic's dilemma." However, it seems a poor analogy because, whereas the genetic overlap in altruism toward others is generally low or nonexistent, the genetic overlap with the future self is 100%. Hence the mystery is not why we are altruistic toward our future selves, but rather why we are not more altruistic than we are, that is, the opposite of the mystery of altruism toward others. Notice that the issue of altruism toward one's future self is more common than the example of alcoholism might suggest: Every investment decision has this character, and in fact, most decisions involve a trade-off between current costs and future benefits.

The theoretical arguments in the article are buttressed with some experimental work done by the author. I am unsure why he focuses on the results of hypothetical experiments when he has done laboratory experiments under controlled conditions with monetary payoffs; the latter would seem more conclusive. In addition, the existence of altruism in the laboratory (as well as spite, which Rachlin does not discuss) has been well established for a long period of time. Modern debate is about the prevalence and not the existence of altruism.

Finally, I found Rachlin's discussion of group selection confusing. This is obviously a difficult area, the most difficult problem being to explain why an altruistic group will not be taken over by egoists. But this involves a long and complex literature, and perhaps it would be best just to say so. The issue of altruism in teams also needs to be treated with caution. In a team setting, it is possible for each individual to be decisive and bear the entire social cost or benefit: Mimicking this artificially is the heart of the Groves-Ledyard mechanism.

Altruism is never self-sacrifice

Michael Lewis

Institute for the Study of Child Development, Robert Wood Johnson Medical School, New Brunswick, NJ 08903. **lewis@umdnj.edu**
http://www2.umdnj.edu/iscdweb/

Abstract: Altruism by definition involves the self's evaluation of costs and benefits of an act of the self, which must include cost to the self and benefits to the other. Reinforcement value to the self of such acts is greater than the costs to the self. Without consideration of a self-system of evaluation, there is little meaning to altruistic acts.

The statement, "by definition, as an altruistic act, it is not reinforced," appears to be at the heart of Rachlin's argument. It is not reinforced because "this act forms part of a pattern of acts . . . a pattern that is valuable in itself, apart from the particular acts that compose it" (target article, sect. 1.2). But of what value and to whom? The answer from a self-psychology point of view is a value to the person who commits the act; it reinforces that person's sense of self and contributes to the sense of worth.

How can it be thought of then as a selfless act, which, of course, is at the heart of what we consider altruism? By giving up a seat on a bus to an old man, the young woman is said to commit an altruistic act. There are several questions to be raised in this example, which may be hidden when we use more extreme examples such as giving up one's life for another. For instance, is it a more altruistic act if the young woman is herself very tired than if she is not? I think that we might all agree that it is more altruistic if she is tired. Thus, from her perspective she is more altruistic the more it costs her something.

In a sense then, altruistic acts become more so, the more costly they are for the person performing them. In understanding altruistic acts, it is therefore necessary that we know (that is, take into consideration) the cost to the person performing the act. This requires, of course, that the person's selflessness, which can only be gained from knowledge of the person's self-perspective, be taken into account. Clearly, the act of charity of a poor man and a rich man must differ; $100 from each does not represent the *same* altruistic act.

But let us return to whether altruistic acts are or are not reinforced. It seems reasonable to believe that there is a reinforcement principle at work; indeed, the paper itself implies one when speaking of a person's life pattern. Adherence to or violation of such patterns are themselves internal reinforcements for the particular altruistic act. The tired woman who gives up her seat for the old man feels good about herself. This feeling good about oneself has value and is a reinforcement for the act itself. Not to give up her seat would make her feel bad. Thus, self-reinforcement is at the heart of the matter. Clearly, the value of this reinforcement has to be weighed against the cost of the act, but as we all recognize, feeling good (or bad) about oneself is an extremely important feature of our internal life. Not to feel good about the self (negative self-attributions that may be global and stable) can lead to depression or other serious illness (Lewis 1992).

The development of altruism then must follow from the child's development of self-structures, including the emergence of the self (Lewis 1979) and the ability to evaluate one's behavior in terms of standards and rules (Stipek et al. 1992). In fact, these require the reinforcement of the child's social world including parents, teachers, and friends, and seem to emerge only after three years of age (Lewis 1995). It requires that the child be capable of understanding not only his or her own needs, but be capable of empathy; that is, understanding the needs of others through imagining what it might feel like for the other (in our example, to be old and have to be standing on a moving bus). Thus, a theory of mind becomes an important part of an altruistic act. Although it is possible that an altruistic-like act can be totally impulsive and without thought, or somehow biologically pre-wired, we would have to agree that such acts are, in fact, less altruistic because they do not take the risks into account; in the same fashion that a $100 gift of a rich man is not as altruistic as the same gift from a poor man.

All of these features, risk consideration, empathy, and self-feelings require that any analysis of altruism be considered from a self-psychology point of view. Not to do so is to limit what we humans are capable of doing. Moreover, to argue that altruistic acts are by definition not reinforced is to deny that there exists a self who possesses a theory of mind, and who has a moral system as well as a system of values capable of weighing costs and rewards.

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

Altruism and Darwinian rationality

Howard Margolis

Harris School, University of Chicago, Chicago IL 60637.
hmarg@uchicago.edu

Abstract. Rachlin adds to the already long list of proposals for reducing what might be seen as social motivation to some roundabout form of self-interest. But his argument exhibits the usual limitations, and prompts questions about what drives this apparently unending quest.

Rachlin provides the latest in the long line of proposals for reducing human social motivation to self-interest. An anthology would have to provide generous sections for game theorists, philosophers, sociologists, anthropologists, psychologists, and of course economists. Biologists would constitute an especially distinguished category. But within evolutionary biology, the prevailing (though certainly not unanimous) distaste for group selection has rarely been focused on the one species in a million which happens to occupy full attention in this *BBS* exchange. A species that is off-scale good at planning, improvising, and communicating might reap benefits from cooperation across a group that is very large compared to a species not so generously endowed. Consequently, with respect to this one species out of a million, a presumption that within-group selection favoring self-interest overwhelmingly dominates between-group selection favoring social cooperation looks arbitrary.

That there is such a rich supply of proposals for reducing social behavior to self-interest suggests the possibility of something fundamentally unworkable here: perhaps the social science counterpart of the numerous proposals for perpetual motion machines or squaring the circle. Circle-squaring, perpetual motion machines, and proposals for reducing apparent social motivation to roundabout self-interest share the peculiar property that they seem to never convince an audience going much beyond their creator. And of course it would follow that if none of the numerous self-interest reductions really work, opportunities to continue the search would never fade, as opportunities never fade for devising the first successful perpetual motion machine or the first coherent squaring of the circle.

What prompts all this effort is that an account of behavior as *transparently* governed by self-interest leads to transparent difficulties. In numerous routine ways (bothering to vote, helping old ladies cross streets) and occasional utterly non-routine ways (Rachlin's baby-rescuer) people fail to do what transparently maximizes self-interest. Human beings may seem miserably inadequate relative to our higher aspirations. (But that we all know what higher aspirations are, provides a hint that we are not likely to be creatures attuned only to maximizing self-interest.) Yet, however far we fall short of our aspirations, we plainly do not fail completely to rise above self-interest (and that "above" is the usual usage here, is another indicator of the previous parenthetical point).

I write as a partisan (Margolis 2000). My NSNX ("neither selfish nor exploited") model works out the consequences of a pair of rules governing marginal use of resources. The rules reflect the competing effects of within- versus across-group Darwinian selection. Jointly, the pair of rules yield predictions of how far a chooser will compromise self-interest to promoted social values. So on this account, motivation includes a measure of what E. O. Wilson (1975) calls "hard-core" altruism, meaning group-oriented motivation that goes beyond what can be accounted for by reciprocal altruism, kin-altruism, and extensions thereof.

The alternative (somehow reducing apparent altruism to self-interest) always entails at least an implicit argument about human cognitive limitations. The most common version (variants abound) starts from the possibility that a person makes a social choice not because she really values contributing to social effort, but because she fears punishment or hopes for reward. But often the context is one where neither motive makes much sense (no one is looking, or if someone unexpectedly were looking the chooser could offer an excuse, or the discounted punishment – even if no excuse is possible – is unimpressive relative to the direct advantage to self-interest).

But then there are fallback positions. A choice that benefits others may also be straightforwardly self-interested. Acts that appear altruistic may turn out to serve self-interest indirectly when account is taken of side-effects (rewards and punishments), or delayed effects (reputation). But, while plainly all these situations occur, this still leaves far too much unexplained. However, further options remain. In fact, the door can be swung open as wide as you please by resort to mistaken self-interest, or archaic self-interest (the behavior would have been self-interested within a small hunter-gatherer community), or on-average self-interest. All are variants on the cognitive limitation theme.

But, as in all of this last group, sooner or later, the defense of self-interest always is driven to some argument turning on cognitive limits: poor judgment, entrenched habit, "fossil" self-interest left over from evolutionary history, or some variant makes the person really only pursuing self-interest serve a wider interest by mistake. In the case at hand, Rachlin argues that the mother rushes into the burning building out of a special sort of habit, where *on average* the habit serves self-interest even though on this occasion it does not. He suggest that this is parallel to the case of self-control, where it may be in a person's self-interest to be committed to a certain pattern of behavior, since otherwise weakness of will would too often lead to short-sighted choices. And, indeed, cases of that sort are worth attention (Ainslie 2001). The person, out of habit, acts in a way that is usually to her benefit, but not always, which turns out better than trusting to willpower or judgment case by case. Rachlin's fictional mother runs into the fiery building. Her motivation (Rachlin argues) is ultimately selfish, though on this occasion perhaps fatally mistaken.

But a sufficiently rational mother would have no occasion to behave that way. She would act from considered choice. She would not be constrained from doing that by mere slowness of wit. She would be smart enough not to need entrenched habits to get her to do what is overall best for herself. For she could do even better by dealing with each choice as it comes. In Rachlin's example, if the risk clearly exceeds the rewards she can plausibly expect, she will stay with her own child, not rush into the building trying to rescue someone else's child. But a human mother, like a human father, is not that smart and might indeed do best if bound by habit. That is Rachlin's argument.

What is wrong with that, since obviously we are indeed cognitively limited? Empirically, the very work Rachlin cites (Herrnstein 1991) finds that even highly sophisticated people reveal themselves to be remarkably vulnerable to the short run trap that Rachlin's Figure 1 illustrates. Even if we suppose that cooperation brings self-interested rewards that more than offset its cost, Herrnstein's experiments should make us doubt that Rachlin's fictional mother would learn this. And in fact there is really not much long-run offsetting reward to many sorts of social acts we commonly see (helping strangers, and so on), and often no serious risk of penalties for free-riding.

Overall, we can see far more cooperation than we could plausibly expect on Rachlin's argument. And as theory, Rachlin's argument does not look very parsimonious if you think about actually trying to use it. Even highly cooperative people on occasion free ride, and even highly selfish people occasionally do the social thing. An adequate theory needs to be able to say something effective about when social cooperation is sustainable and when it collapses, when people take the trouble to help a stranger and when they do not. For all these cases (and numerous other contrasting pairs) are common. The relative prevalence of social and cooperative choices (sometimes on a heroic scale) relative to their opposites is plainly crucial to the viability of human social arrangements.

Rachlin would need vastly more than he provides to deal with all that. As is common across the library of such reductions to self-interest, the *parsimony* of the reductions would turn out to be il-

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

lusory. This leaves a puzzle about the amount of effort that so many scholars have put into finding a way to reduce social motivation to a disguised pursuit of self-interest. Perhaps we are seeing, at bottom, the sense of human moral degradation that shows up in religious doctrines of original sin and its parallels in non-Christian traditions. But maybe, deep down, we are not so abjectly rotten.

What is an altruistic action?

Marco Perugini

Department of Psychology, University of Essex, Colchester CO4 3SQ, United Kingdom. **mperug@essex.ac.uk**
http://privatewww.essex.ac.uk/~mperug

Abstract: Rachlin's argument rests on his definition of an altruistic action. Three main features characterize this definition: An altruistic act (1) always has a negative value, (2) is a subset of self-controlled actions, and (3) is meaningful only in the context of repeated interactions. All three features are highly questionable.

Many readers would find Rachlin's definition of an altruistic action peculiar. In my view, it is also fundamentally flawed.

First, Rachlin defines an altruistic act as an action that is not reinforced; that is, it results in a negative value if we add up costs and benefits. This definition may appear obvious and uncontroversial at first, but I suspect it is so only if one embraces a teleological behaviorism view. The critical assumption here is to equate material and psychological costs and benefits. There is a huge literature in several fields, however, showing that they are not the same. Take, for example, interdependence theory (Kelley & Thibaut 1978). The authors argue that actors transform a given decision matrix, such as the one associated with the prisoner's dilemma game, according to their motives into the effective decision matrix. It is this subjectively transformed matrix that influences the actual actor's decision. For example, if one is competitively oriented, the key transformation will be such that the entries in each cell of the matrix will be the difference between one's and the other's payoffs. In this frame, altruistically oriented persons would transform the entries in the cell such that the other participant's payoff will be weighted more than one's own. A more defensible definition would therefore be that an altruistic action has a negative value in terms of material costs and benefits to self, leaving open whether it has always a negative value when accounting also for psychological costs and benefits. If we adopt this definition, it is easily conceivable that there are conditions (e.g., when the benefit to other is much higher than the cost to self) or persons (e.g., those who gain psychological benefits from performing an altruistic action) whereby an altruistic action does not have a subjectively perceived overall negative value.

The work of Batson (1990) provides an example of one mechanism, empathy, which evokes altruistic motivations and altruistic actions. Actors who act on the basis of empathy transform the situation into one where the benefit for the other becomes their main concern. The psychological benefits given by helping the needy other by their altruistic actions can therefore outweigh the negative value of the personal material costs and benefits, lending to an overall subjectively perceived positive value of the action. Similarly, the work of Frank (1988) focuses on other mechanisms, anticipated negative emotions such as shame and guilt, which should increase the psychological costs associated with not being altruistic. Again, it is easy to think of persons and situations whereby the overall value of an altruistic action is positive, because of the balance between the psychological costs of not helping and the material costs of helping.

Second, Rachlin assumes that all altruistic actions are self-controlled. The very first problem with this assumption is that to be meaningful at all one should also accept the proposition that habitual altruism is a happier mode of existence than habitual selfishness. Why should altruism lead to being happier than selfishness? According to what mechanism? Why should all people feel happier by being altruistic? Rather than being a reasonable proposition, this is an assumption that washes away most of the theoretical power of the concept of self-control. Rachlin should explain with his theory why people are happier by being altruistic instead of assuming it. For the sake of the argument, suppose for a moment that Rachlin's proposition is tenable, which it is not. It would still be unclear why all altruistic actions are self-controlled. Why would people necessarily need to focus on the long-term pattern of acts to be altruistic? Aren't there countless real-life examples and experimental results showing that people can behave altruistically even without self-control? Moreover, if we differentiate between psychological and material costs and benefits, it follows that at least some altruistic actions do not carry an overall negative value. This alone implies that conceptually there is no need to invoke self-control as an exclusive explanation of altruism.

Even more puzzling is how Rachlin tries to accommodate his theory with some experimental findings, including his own experiments, showing that the performance of altruistic actions is highly sensitive to features such as the other's behavior (reciprocation) and the context (framing). In Rachlin's theory, there are only two elements that can explain changing rates of altruistic actions: self-control and the payoffs in the interaction matrix. How then can features such as reciprocation and framing influence the rate of altruism? One possibility is that self-control changes as a function of framing effects and other people's reciprocal behavior. It is extremely hard for me to understand how self-control can change in this way – perhaps it could in a lifetime as a consequence of a history of reinforcements and punishments, but not during a single experimental session. The other possibility is that the payoffs change depending on features such as framing effects and reciprocal behavior. But this would imply that there is a difference between given (material) payoffs and transformed (psychological) payoffs. This is compatible with other accounts of altruistic actions, such as those in terms of framing (Larrick & Blount 1997), of social value orientations (Van Lange 1999), and in terms of reciprocity (Perugini & Gallucci 2001). However, this type of explanation is explicitly excluded by Rachlin as a consequence of his basic assumption of altruistic actions always carrying negative values.

Finally, Rachlin argues that altruistic actions can be understood only in the context of repeated choices. Of course, this follows directly from, and is consistent with, his definition of an altruistic action. However, it is inconsistent with experimental and real-life evidence. For example, it has been shown that people show increased altruistic behavior provided that they first have empathic feelings for the other (e.g., Batson et al. 1995) or provided that the other previously has behaved kindly (Gallucci & Perugini 2000). These altruistic actions occur also in contexts of anonymous exchanges with no expectations of future interactions. Moreover, they occur not only for people with some kind of personal disposition towards being altruistic, but also for people whose action is either elicited by situationally evoked empathic feelings or by the willingness to reciprocate someone else's action in the ongoing interaction. Of course, Rachlin could argue in reply that this is irrelevant, although obviously it is not: What matters is whether the same person acts in an altruistic way across different situations. One of the consequences of Rachlin's analysis, in fact, is that given a certain level of self-control such that one would act altruistically, that person will act altruistically also on other occasions to preserve his or her cherished pattern of acts (i.e., being altruistic) against a specific now-convenient action (i.e., selfish choice). This implies a remarkably high level of predictability of altruistic actions based on the previous actions, therefore supporting the notion of altruistic habit or altruistic personality as a main explanation of altruistic actions. But, even in this case, the empirical evidence is thin. Although there is evidence for the existence of an altruistic disposition or personality, it can only explain a portion of all altruistic actions (e.g., Carlo et al. 1991).

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

A theory must also be judged by considering aspects such as its capability to explain and predict the phenomenon of interest. Rachlin's theory does not rank high in this respect. Although it is likely that some altruistic actions can be explained in terms of self-control, most of them cannot. A range of alternative explanations, including empathy, reciprocity, and framing, can do a better job. I suspect that the only way in which Rachlin can (partly) explain altruistic actions is by adopting a definition that is idiosyncratic and far away from what most other people and researchers would use. But, then, what is an altruistic action?

ACKNOWLEDGMENT

Preparation of this commentary was supported by RPF grant DGPC40 from the University of Essex.

Altruism: Brand management or uncontrollable urge?

Daniel Read

Department of Operational Research, London School of Economics and Political Science, London, WC2A 2AE, United Kingdom. **d.read@lse.ac.uk**

Abstract: The act-pattern model of altruism is primarily a brand-equity model, which holds that being altruistic can be traded for social benefits. This is a variant of the "selfish" altruism that Rachlin decries, with altruism being dictated by cold calculations. Moreover, personal and social "self-control" may not be as similar as Rachlin suggests – although we have good (biological) reasons to sacrifice the interests of our current selves in favour of our future selves, we have no such reason to sacrifice ourselves for our neighbours. When we do sacrifice ourselves – giving in to true altruism – we will be repaid with extinction.

An altruistic act is usually defined as one that will harm the actor but benefit others. A woman running into a burning building to save someone else's child would be an altruist if she was harmed by the action while the child was helped. At first glance, altruism seems impossible. The job of any organism is to maximise its inclusive fitness, or the representation of its genes in subsequent generations, and so natural selection should quickly eliminate any altruism genes. Theoretical biologists, however, have shown that something that looks a lot like altruism can arise from a subtle form of self-interest. One way is illustrated by Haldane's (perhaps apocryphal) claim that he would lay down his life for two brothers or four cousins (Medawar & Medawar 1983). Altruism can increase our fitness when we are judiciously helpful to our relatives, because we share lots of genes with those relatives. Hamilton's law says that we will help others if the cost of our helping does not exceed the benefit to the recipients multiplied by their relatedness to us (Hamilton 1964). A second kind of selfish altruism is reciprocal altruism, whereby we are nice to others because they are nice to us, or at least because we are counting on them to be so (Trivers 1971). For reasons related to both kinship and reciprocal altruism, reproductive potential can also be influenced by what others think of us. For example, if we (males especially) are altruistic to people at large, it may be evident to others (females especially) that we will probably be particularly generous toward them and their offspring – if their offspring also happen to be ours (Gintis et al. 2001; Gurven et al. 2000). This is a kind of brand-equity or signalling model of altruism, in which being altruistic is a valuable brand. Brand-equity altruism works because it enables the altruist to have more offspring, and also because it will make others willing to help the altruist because they can count on reciprocation. Rachlin's account of altruism appears to be of the brand equity variety. Altruism is better than nonaltruism because being an altruist is a valuable commodity. Altruistic acts are part of the process of brand building.

One aspect of the act or pattern model that I do not accept is the idea, introduced early in the target article but perhaps abandoned by the end, that there are some acts that would not be done if rational calculations were done over them, yet are nonetheless beneficial because they form part of the pattern of altruism. I do not think this even conforms to the "act versus pattern" distinction, which Rachlin has described before (Rachlin 1995a). As I see it, individual acts in a pattern *always* pay off more than acts that are not part of the pattern. The payoff may not come immediately, but this is no different than any other choice with delayed consequences. To illustrate, in the primrose path experiment in Figure 1 of the target article, the individual will be better off choosing Y once even if he never chooses it again. The next 10 choices between X and Y, regardless of which one is chosen or how often, will yield a higher payoff than they would have without the initial Y choice. So this is a first-order return for the choice (or *act*) of Y without any reference to the pattern of which it is a part. It might be true that if the person thinks of many consequences in combination, they will better *recognise* that contingency (cf. Read et al. 1999), but that is not the same as claiming that the contingency is not there for the individual choice. If altruism pays off in the long run, then an altruistic act cannot be distinguished from the act of the bank manager who loans us money with the expectation of repayment with interest. We can more realistically think of the altruist as being like a fireman armed only with sophisticated fire fighting equipment and wearing fireproof clothing. Perhaps he has to overcome some fear before rescuing the child, but when he comes out after a successful rescue he gets the applause of the crowds, the adulation of the womenfolk, a fat bonus in his paycheque, and a job for life. Any rational analysis would tell him to pick rescue over no-rescue. It may not even be logically coherent to suggest that some acts *reduce* the expected payoffs to the actor, yet at the same time increase those payoffs.

Rachlin's argument also turns on a connection between individual self-control and altruism. There are certainly interesting parallels between the two domains, and I have written about them myself (Read 2001), but can we conclude that, because the two domains are *analogous,* they are also *homologous*? Homology – common underlying structures and processes – is necessary for us to freely move back and forth between domains, applying theoretical concepts from one to the other. The interesting direction (as Plato recognised) is going from the social to the individual. Because we do not fully empathise with our future selves, we can think of ourselves as standing in relation to them as we do with other people. The *analogy* can, however, easily be stretched too far. A major difference between intrapersonal and interpersonal altruism is that we share 100% of our genotype with our future selves; hence we should rationally want to be extremely altruistic to them. But we share no more than a random sampling of genes with a stranger. This means that acts of self-control that will benefit our future selves at the expense of current ones can be the product of rational biological choice – we want to avoid those impulsive acts that reduce our future reproductive potential. Although we impulsively reach for a cigarette, reason forces us to abstain because we believe, rightly or wrongly, that we (taken as a whole) will be better off not smoking. Rachlin's altruism, however, often seems more akin to taking the cigarette rather than abstaining. This is explicit with the burning building example, which is clearly given as an act that would *not* be taken if reason prevailed. Viewed from this perspective, social altruism and self-control corresponds to personal selfishness and impulsivity – we are nice when we don't stop to think of how foolish we are being.

In summary, I see two kinds of altruism being discussed in Rachlin's paper. Brand-equity altruism is a currency that can be exchanged for cooperation and reproduction. The second kind of altruism occurs when the givers forget their biological interests and impulsively sacrifice themselves for others. From my point of view, it would certainly be nice to meet more of the second kind of altruist. From their point of view, however, it would be better if they did not meet anybody. I suspect, unfortunately, that they are an endangered species.

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

Defying evolutionary psychology

Phil Reed

Department of Psychology, University College London, London, WC1E 6BT, United Kingdom. **p.reed@ucl.ac.uk**

Abstract: That altruism and selfishness can be explained behaviorally without resort to Darwinian psychology has wide implications. Although currently fashionable to avow an evolutionary approach, such a posture is unnecessary. The links between Darwinian approaches and behavior analysis are weak. This weakness is to the benefit of behavior analysis, as evidence for Darwinian psychology has been weak for over a century.

As with most interesting and important issues, the ones arising from this article are rather troubling. The major of these issues concerns the relationship of behavior analysis to Darwinian psychology (e.g., evolutionary psychology, sociobiology, etc.). That there may be tension between these two approaches is implicit in the target article, and deserves comment as it is of considerable interest to behavior analysts.

A behavioral explanation of altruism and selfishness appears eminently self-evident, at least to a behaviorist. Yet, there are some quibbles with the specifics of argument advanced by Rachlin. The use of the notion of "intrinsic reinforcement," which has never been resolved satisfactorily, is one concern. The identification of extended behavioral units is a further concern, as is the relationship of this latter concept, which stems from a molecular approach to the analysis of behavior, to the molar approach advocated by Rachlin. However, the important questions arising from this target paper are: "Why is it necessary to write such an article?" and "Why would it attract controversy?" The answers to both of these questions revolve around the observation that "altruism" and "selfishness" are usually addressed by approaches couched in Darwinian terms, and their "just so" theories have been accepted with little objection. This situation needs to be challenged. Rachlin's challenge is important and could be amplified.

Rachlin suggests that the main reason to doubt the traditional accounts of altruism and selfishness are their adherence structural modes of explanation: an explanation that uses *internal* explanations as causes of the behavior. This criticism is partly correct and partly incorrect.

Darwinian psychology (e.g., evolutionary psychology, sociobiology, etc.) tends to leach off the work of geneticists in order to give some scientific credibility to its speculations. However, an appeal to genetics that involves description of the operation of transcriptors, protein manufacture, and so on, can be viewed either as orthogonal to a behavioral explanation, or as part of the setting conditions for behavior.

Darwinian psychologists' explanations often involve appeal to selection and fitness. These terms could be viewed as descriptions of external pressures limiting the types of behavior that could emerge within a population. Yet, often they are taken to be pressures that influence cognitive structures that, in turn, control and limit behavior. These structures are taken to be coded for, and selected for, in the genetic make up of the organism. This form of explanation, as Rachlin points out, is not compatible with a behavioral analysis.

Given the incompatibility, it is surprising that many within behavior analysis argue to create links with Darwinian approaches to behavior. These reasons should be outlined. One obvious and nonscientific reason for the attraction is that Darwinian psychology is a scientific approach to biological science that must be protected against nonscientific alternatives. This appeal may carry some weight in countries with a large religious fundamentalist minority, but this political argument does not carry such weight in countries lacking such a problem.

Other authors have drawn links between the historical development of Darwinism and behavior analysis. Especially important in making this argument is the manner in which Darwinism was dismissed for much of the early part of the last century. This dismissal was on the grounds that the evolutionary approach displayed no hard evidence. Its acceptance came only after the new synthesis with genetics, and not for reasons that Darwinism offered such evidence itself. Adherents to Skinner should know that arguments through historical analogy are not strong. Objections to behavioral psychology are not based on lack of evidence on its own terms. Rather, they are made because these terms are not accepted. In this sense, it is the fruit of Mendelian genetics that holds a stronger relationship to behavioral psychology than the fruit of Darwinian evolution.

Darwinism was never a strong historical precursor of behavioral psychology. Unfortunately, Skinner's own analogies with evolution after 1945, especially references to "selection through consequences," do not help to maintain a separation between these disciplines. In fact, physiology was the real precursor of behavioral psychology. The links are clear between physiology and early associationism, Pavlov and contemporary associative psychology, and Crozier's general organism physiology and Skinner.

A stronger reason for behavior analysts' claiming a link to evolution is the notion that humans and nonhumans can be assumed to be governed by similar processes. Yet, this is a view that most Darwinian psychologists deny. For example, the attack on General Learning Theory, and the suggestion of niche specific intelligences, both strike at this notion of an easy extrapolation between human and nonhuman. However, the extrapolation being attacked by Darwinian psychologists is an abstract philosophical notion of equivalence between species and it is countered by abstract theoretical arguments. In contrast, behavioral psychology promotes the notion of extrapolation as an empirical concept. Behavioral regularities are observed in the data across species placed in similar situations, and are not assumed to exist on an abstract theoretical or philosophical level.

I offer an explanation of two concepts often intrinsically linked to courses taught about Darwinian psychology. With some reservations, it presents a cogent argument that such behaviors can be explained without reference to evolution. This is an important break, and not just for the area of direct concern to target article. Darwin is a historical figure, as are Skinner and Newton. They all helped to shape their fields, and contemporary workers owe them a debt. However, there is no reason to let these figures hold back the development of the field and certainly no need to argue for a link, not only to the past, but to a past that fails to deliver a strong empirically supported present theory.

Altruism is a social behavior

Richard Schuster

Department of Psychology, University of Haifa, Haifa, 31905, Israel.
Richard.Schuster@psy.haifa.ac.il; schuster@asu.edu.

Abstract: Altruism and cooperation are explained as learned behaviors arising from a pattern of repeated acts whose acquired value outweighs the short-term gains following single acts. But animals and young children, tempted by immediate gains, have difficulty learning behaviors of self-control. An alternative source of reinforcement, shared by animals and humans, arises from social interaction that normally accompanies cooperation and altruism in nature.

Rachlin offers an ingenious and provocative extension of reinforcement theory to address the dilemma that also challenges evolutionary biology: What mechanisms permit a behavior to occur when it is immediately costly to the performer? This problem arises by definition for altruism (Krebs & Davies 1993) and also for cooperation, either when outcomes are not allocated equitably or individuals can do better by operating alone (e.g., Boesch & Boesch 1989; Caro 1994; Packer et al. 1990; Scheel & Packer 1991). Rachlin's solution is to find a way for such behaviors to pay off eventually for the performer. The central claim is that altruism, like self-

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

control, is a learned behavior reinforced by long-term gains from repeated acts – organized into a pattern and perhaps identified by an abstract rule – that outweigh whatever short-term gains follow from single instrumental acts. The focus is also on reinforcement for individual behaviors (Skinner 1953) without considering whether additional reinforcement arises from the social contexts within which altruism and cooperation typically occur. This commentary will argue that social influences cannot be ignored when explaining why and how altruism and cooperation occur.

For Rachlin, it is sufficient to explain altruism and cooperation from processes also responsible for learning arbitrary behaviors of self-control. But this kind of behavior can be difficult to generate, especially in animals and young children, because small immediate rewards are regularly preferred over larger delayed rewards (Logue 1988). This also explains why animals have difficulty in learning to cooperate when participating in iterated prisoner's dilemma games, preferring instead the larger short-term gains obtainable from defecting (e.g., Clements & Stephens 1995; Green et al. 1995). So the target article leaves unexplained how altruism and cooperation can be acquired by animals living under free-ranging conditions. Rachlin sidesteps the problem by focusing on our own species, with examples like rescuing strangers from burning buildings, leaving tips in restaurants never to be visited again, and cooperating more than expected in prisoner's dilemma games. To explain how such behaviors can be learned, the suggestion is that they are "jump-started" by acquiring "value" from cultural and/or religious principles transmitted by the likes of parents, teachers, and therapists. So, for humans, social contexts are important for learning altruism.

Social contexts may also underlie the widespread expression of cooperation and altruism in animals (Dugatkin 1997; Krebs & Davies 1993), but based on less sophisticated learning processes. One strong possibility is that acts of altruism and cooperation evoke intrinsic and immediate reinforcement arising from social interactions between familiar individuals that usually precede, accompany, and/or follow such acts (Schuster 2001; in press). Similar processes may also exist in humans (Frank 1988; Schuster, in press; Sober & Wilson 1998). In the brain, opioid systems have been linked with social interaction and social rewards, the latter hypothesized to provide reinforcement from mechanisms that are behaviorally and physiologically distinct from other reward mechanisms (Panksepp et al. 1997).

Behavioral evidence that social interaction contributes significantly to how and why animals cooperate is emerging from experiments in which pairs of laboratory rats are reinforced with saccharine-flavored solution for coordinating back-and-forth shuttling within a shared rectangular chamber (Schuster 2001; in press; Schuster et al. 1982; 1993). No cues are available to facilitate coordination beyond the presence and behaviors of the animals. Under these conditions, subjects readily learn to work together in ways that resemble cooperation and altruism under free-ranging conditions. Behaviors are highly social acts in which familiar participants enjoy unrestricted interaction and the freedom to develop idiosyncratic ways of working together. Under such conditions, pairs exhibit roles, dominance, and levels of coordination influenced by social interaction and sensitive to strain, sex, and kinship (Schuster 2001; in press).

This kind of model contrasts with most laboratory models in which unfamiliar subjects are physically isolated in separate chambers and reinforced when both perform individual acts such as key pecking or bar pressing (Clements & Stephens 1995; Hake & Vukelich 1972; Skinner 1953). Such models are the laboratory expression of a "scientific paradigm" that regards all behaviors, whether social or individual, as individual acts that are ultimately selfish, benefiting individuals or their genes (Dawkins 1976/1989; Dugatkin 1997; Krebs & Davies 1993; Skinner 1953). The irrelevance of social interaction is shown by "nonsocial" models that isolate subjects from any kind of interaction by means of opaque partitions or separate cubicles (Hake & Vukelich 1972). Cooperation and altruism have, in effect, been transformed into models of individual action. Despite these conditions, human subjects in prisoner's dilemma games still tend to cooperate more than expected (Brann & Foddy 1988; Palameta & Brown 1999). Perhaps this is a reflection of a sociocultural norm, although some players seem more concerned about their anonymity and the potential embarrassment from meeting other players whose outcomes were adversely affected by defection (e.g., Forsythe et al. 1994). But if animals play such games under the same conditions of isolation, behavior is consistently dominated by the temptation of larger immediate outcomes (Clements & Stephens 1995; Green et al. 1995).

When cooperation incorporates unrestricted social interaction, as in the model of coordinated shuttling, social influences on the motivation to cooperate can be demonstrated (Schuster 2001; in press). In one set of experiments, subjects chose in a T-maze between coordinated shuttling and individual shuttling with no difference in the frequency and likelihood of obtained reinforcements. Unlike prisoner's dilemma games, the coordinated shuttling was a genuine social alternative. If outcomes alone influence choice, there should have been no preference for either option. But the majority of subjects in two experiments (39 of 50) preferred to cooperate in eight choice trials, and overall, the majority of all 400 choices (74 percent) were for the cooperative option. Moreover, preference was positively associated with how well pairs were coordinated when cooperating, suggesting that the relationship between cooperators was influencing whether or not they preferred to cooperate. Preference, however, was not predicted from relative rates of reinforcement from the two options.

The second kind of evidence suggests that the incentive value of reinforcements is affected by how they are obtained. The consumption of the reinforcing solution was measured immediately following sessions in which animals were reinforced under conditions that varied in the level of social interaction: individual shuttling, cooperative shuttling while separated by a partition of vertical bars, and cooperative shuttling with unrestricted interaction. Postsession consumption was markedly increased only by the last condition, suggesting elevation of either the need for the reinforcing solution, its hedonic affect, or both (Berridge 2000). The same finding has just been replicated across groups of subjects (Tsoory & Schuster, unpublished).

The above results resonate with Rachlin's suggestion that altruism cannot be explained entirely by "extrinsic" reinforcement (e.g., money or food pellets) without considering "intrinsic" reinforcement arising from the performance of altruistic or cooperative behaviors. One source may indeed be the acquired value of temporally extended patterns of behavior, but this is probably confined to humans and perhaps also to cognitively advanced animal species like higher primates. It is significant that macaque monkeys and humans are also better at choosing larger delayed reinforcements in self-control experiments (Logue et al. 1990; Tobin et al. 1996). But the rat data cited previously suggest that an important source of intrinsic reinforcement arises from nothing more complex than social interaction and socially mediated coordination when behaving altruistically or cooperatively. Moreover, such processes would be well within the capability of most if not all species that demonstrate cooperation and altruism. Consistent with Rachlin's perspective, reinforcement from social interaction would also not be completely innate and unvarying, but affected by the social relationships that emerge from repeated acts by the same individuals (Schuster 2001). Positive reinforcement is expected when behaviors are affiliative and well coordinated, and negative reinforcement is expected when behaviors are aggressive, poorly coordinated, and highly competitive over access to shared outcomes. Corroborating evidence comes from the variety of species, including fruit flies (*Drosophila*), in which females choose mates based on coordination of displays (Maynard Smith 1978). And in the context of aggression, violence is often constrained by engaging instead in highly coordinated displays of movements or songs that have earned the sobriquet "dear enemy" (Beecher et al. 2000; Krebs 1982).

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

Affective states also accompany social acts in our own species, providing a homologous mechanism for explaining cooperation and altruism not dependent on the value of patterned acts or sociocultural rules. Do many of us not feel good when donating time and resources to aid another? Or experience pleasure not only from drinking a fine wine, seeing a movie, or gazing on a spectacular view, but even more from doing these things with a good friend? Are we affected not only by dancing and love making, but by how well the behaviors mesh together? More cognitively, do we not feel rapport when discovering that another person shares our likes and dislikes? Moreover, this "click" from interpersonal relationships may be even more powerful in large groups. Highly orchestrated, ritualized, and coordinated ceremonies seem to evoke enjoyment, cohesion, and a sense of belonging in contexts as varied as religion, politics, sports, military parades, music, and dance (McNeill 1995). Even without deliberately orchestrated coordination, there is a tendency for people, regardless of age, to "behaviorally match" the movements and postures of others in ways that affect feelings and attitudes about them (e.g., Chartrand & Bargh 1999; Meltzoff & Moore 1977). These are all examples of affective states that point to fundamental processes, or perhaps only one unified process, designed to create and maintain social relationships from cohesion in social interactions. To identify all the processes underlying behaviors like cooperation and altruism, perhaps it is time for modern behaviorism to become social.

ACKNOWLEDGMENTS

This commentary benefited from conversations with Peter R. Killen. The research described herein was supported by grant no. 96–293 from the U.S.–Israel Binational Science Foundation to the author and Peter R. Killeen.

Internal mechanisms that implicate the self enlighten the egoism-altruism debate

Constantine Sedikides and Aiden P. Gregg

Department of Psychology, University of Southampton, Highfield Campus, Southampton SO17 1BJ, England, United Kingdom. **cs2@soton.ac.uk**

Abstract: Internal mechanisms, especially those implicating the self, are crucial for the egoism-altruism debate. Self-liking is extended to close others and can be extended, through socialization and reinforcement experiences, to non-close others: Altruistic responses are directed toward others who are included in the self. The process of self-extension can account for cross-situational variability, contextual variability, and individual differences in altruistic behavior.

Rachlin discounts the role of internal mechanisms in guiding altruistic behavior, while augmenting the role of learning. He argues that altruism "is not motivated . . . by the state of an internal mechanism; it is rather a particular component that fits into an overall pattern of behavior" (target article, sect. 1.2) and that altruism "may be learned over an individual's lifetime . . . by forming particular acts into coherent patterns of acts" (sect. 1.1). Rachlin goes on to ask the question, "What are the patterns of behavior that the altruistic acts fit into?" (sect. 1.2). His answer to his own question is that an altruistic act "forms part of a pattern of acts . . . , a pattern that is valuable in itself, apart from the particular acts that compose it" (sect. 1.2).

However, Rachlin hardly seems to provide compelling grounds for discounting a role for internal mechanisms in the accounting for altruism. Indeed, the egoism-altruism debate could hardly get off the ground without reference to internal motivations: The debate is over such motivations (Batson et al. 1997). To argue, then, as Rachlin does, that all there is to altruism or egoism is an objective pattern of learned behavior, is to implicitly reject the debate in its essential form, not to contribute to it. Taking such a provocative and forthright stance can be, of course, dialectically aerobic. But adopting such a stance also runs the risk of prematurely dismissing potentially fruitful avenues of investigation. We think it would be unfortunate if stipulatively defining egoism or altruism in behavioristic terms were to hamstring scientific research into its mentalistic aspects.

Did we just say "mentalistic"? The word will surely send a cold shiver down the spine of any self-respecting behaviorist! It conjures up hated notions of subjectivity, teleology, and wishy-washy folk psychology. How could anything so infuriatingly nebulous possibly elucidate egoism or altruism? Better to sidestep all that humbug and focus on the concrete deeds themselves. However, the study of cognition, emotion, and motivation has come a long way since the heavy-handed introspectionism of Wundt. In fact, we would argue that social psychologists, in using mentalistic constructs like attitude, evaluation, self-concept, mood, and stereotype, behave very much like the theoretical physicists whom they are supposed to emulate. Social psychologists collect experimental data in an objective and replicable way using specialized instruments, and the data they collect then serve to support or refute sufficiently well-defined empirical theories. For example, a substantial amount of statistical and methodological expertise goes into designing and validating questionnaires to assess traits and attitudes hypothesized to have particular antecedents, correlates, and consequences, and social psychologists are increasingly moving toward the use of indirect measures that rely on reaction time and physiological responses (Reis & Judd 2000).

Social psychological theories do contain mentalistic constructs that are sometimes problematic for one reason or another; for example, the optimal way to operationalize them may be debated, because any particular operationalization fails to exhaust the original broader meaning of the construct. However, just because mentalistic constructs are problematic does not mean they are useless. In contrast, theoretical physicists are content to use mathematical abstractions that have the rather serious problem that they utterly defy anyone's attempts to intuitively grasp them. Their justification for continuing to employ them, nonetheless, is that those mathematical constructs are an essential component of powerful, interesting, and testable theories about the nature of physical matter. Similarly, social psychologists who manage come up with powerful, interesting, and testable theories about the nature of the human mind are justified in using mentalistic constructs that are just a tad more slippery than objectively defined lever pressing. Readers seeking a gentle introduction to the methodological logic of key social psychological experiments could do worse than to consult the forthcoming volume by one of the authors (Gregg) and his colleagues (Abelson et al., in press).

Getting back to the specifics of Rachlin's article, we note that, to our social psychological eyes, his thesis leaves several key questions unanswered: To whom are altruistic responses directed? How are altruistic responses socially learned? What is the function of altruistic responses? We argue that answers to these questions will be enlightened by taking internal psychological mechanisms into serious consideration.

We would like to focus on internal mechanisms that implicate the self. There is compelling evidence for explicit self-liking (Sedikides & Strube 1997). For example, people seek out positive (as opposed to negative) information even when it is nondiagnostic of their traits and abilities (Sedikides 1993), remember positive but not negative self-referent information (Sedikides & Green 2000), strategically exaggerate their strengths and downplay their weaknesses (Dunning 1993), and harbor illusions of control while expressing undue optimism for the future (Taylor & Brown 1988). There is also compelling evidence for implicit self-liking (Sedikides & Gregg, in press). People display an enduringly positive evaluation toward letters that appear in their own names or numbers that appear in their own birthdays (Kitayama & Karasawa 1997; Nuttin 1987), and toward consumer items (e.g., keychains, pens) on which they have only recently claimed ownership (Beggan 1992; Kahneman et al. 1990).

This overwhelmingly positive affective and evaluative orienta-

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

tion toward the self can and does generalize to others. To begin with, self-liking is extended to persons with whom one forms strong relational bonds. Such persons can be relatives (e.g., one's mother; Aron et al. 1991), relationship partners or close friends (Aron et al. 1992), and important groups to which one belongs (i.e., ingroups; see Smith & Henry 1996). Through internalization processes, these persons or groups are included into one's self-concept: They become an integral part of one's self-definition. Importantly, the self-inclusion process can generalize to distant relatives and friends, acquaintances or strangers, and even to benign (i.e., non-antagonistic) outgroups via socialization and reinforcement experiences such as the ones to which Rachlin refers or alludes. The inclusion of strangers in the self is a rather protracted process, but we believe it does reflect a cultural universal: People in all cultures find social interactions intrinsically rewarding, invest effort in discovering vital areas (e.g., hobbies) of overlap between self and other, explore these areas with curiosity, and engage pleasurably in the formation of relationships or alliances. This cultural universal is a manifestation of the evolutionary-based need to belong (Baumeister & Leary 1995; Leary & Baumeister 2000).

So far, we have attempted to describe a subset of internal mechanisms that regulate the ways in which the individual connects with her or his social environment. Importantly, these internal mechanisms are directly relevant to the egoism-altruism debate. In the following paragraphs, we will highlight several spheres of relevance.

First, the reinstatement of internal mechanisms reaffirms the role of evolution in altruism. In fact, these mechanisms provide a psychological explanation for why altruistic behavior is (generally) more likely to be directed to kin or kin-like others (e.g., romantic partners, close friends) as opposed to total strangers. Kin or kin-like others are included in the self, whereas others are not. Hence, the affection or positive evaluation for the self is also directed toward kin or kin-like others. The individual is nurturing and caring not only toward the self but also toward persons who are perceived as an extension of the self (Cialdini et al. 1997; Neuberg et al. 1997; though see Batson et al. 1997, for a contrary view).

Furthermore, internal mechanisms can account for cross-situational or cross-target variability in altruistic responses. On the face of it, it is somewhat puzzling to observe the same individual treating some persons or groups altruistically, while treating other persons or groups indifferently. Such seemingly paradoxical behavior is efficiently explained in terms of self-inclusion: Those who are included in the self are beneficiaries of altruistic behavior, whereas those excluded from the self are not.

Additionally, internal mechanisms can explain contextual variability in altruistic responding. It is bewildering to observe the same individual treating certain persons or groups altruistically on one occasion, but indifferently on another. Again, this seemingly paradoxical behavior can be accounted for in terms of self-inclusion. Altruistic behavior toward a given target is likely to be enacted when the construct of "target inclusion into one's self" is cognitively accessible (Sedikides & Skowronski 1991). Alternatively, altruistic behavior will likely not be enacted when this construct is cognitively inaccessible.

Moreover, internal mechanisms can account for individual differences in altruistic behavior. People differ in terms of the expandability of their self-concept boundaries (Duckitt 1992; Phillips & Ziller 1997). Some have easily expanded boundaries (i.e., are prone toward social integration and social tolerance), whereas others have rather sclerotic boundaries (i.e., are prone toward social differentiation and social intolerance). The former will generally behave more altruistically than the latter.

The reestablishment of internal mechanisms enlightens answers to critical questions surrounding the egoism-altruism debate. We posed three questions in the beginning paragraphs of this commentary. We are now able to provide answers to these questions. Altruistic responses are directed toward persons or groups that are included in the self. Altruistic responses are learned through socialization and reinforcement experiences. Lastly, the function of altruistic responses is to promote the individual's welfare, which frequently presupposes or relies on the welfare of others (persons or groups).

In summary, the reestablishment of internal mechanisms challenges the concept of altruism as simply a pattern of behavioral acts. These mechanisms make a credible case for altruism to be conceptualized as deeply rooted in egoistic (i.e., motivational) psychological hardware.

Putting altruism in context

Joel Sobel

Department of Economics, University of California at San Diego, La Jolla, CA 92093. **jsobel@ucsd.edu** **http://weber.ucsd.edu/~jsobel/**

Abstract: I argue that Rachlin's notion of self-control is imprecise and not well suited to the discussion of altruism. Rachlin's broader agenda, to improve collective welfare by identifying behavioral mechanisms that increase altruism, neglects the fact that altruism is neither necessary nor sufficient for desirable social outcomes.

Evolutionary biology provides powerful ways of understanding unselfish behavior to closely related individuals (Hamilton 1964) or to reciprocity in long-term relationships (Trivers 1971). Rachlin is concerned with instances of human altruism that are hard to explain using these theories. Evolutionary mechanisms that rely on group selection (Boehm 1997; Sober & Wilson 1998) or community-enforced morality (Alexander 1987) provide explanations for unselfish behavior in human communities. These approaches teach us that to understand altruistic actions, we should examine the individual in connection with the composition and norms of the society in which he lives.

Economic theory assumes that agents select only an action that maximizes utility from their available choices. Unselfish behavior does not occur. Economics adapts its methodology in the face of apparently contradictory evidence by broadening the definition of self-interest, for example, by assuming that individuals obtain utility from the act of giving or through the consumption of others, or by recognizing that economic relationships are dynamic and that an individual's long-run selfish best interest is best served by doing things that are not consistent with short-term selfish goals.[1]

The approaches of both biology and economics illustrate that the context of actions is important; and both require careful attention to the definition of altruism. Rachlin argues that one must consider altruism in the context of patterns of behavior and provides a definition that makes the mechanism supporting altruism a special case of the mechanism that determines self-control. Rachlin should be commended for pointing out the importance of patterns of behavior. Altruism generated by a preference for establishing a pattern of good behavior is internally motivated. It does not rely on generating reciprocal responses from others that are vital to the dynamic arguments in biology and economics. Rachlin's emphasis on patterns should motivate behavioral researchers to widen the context of their experiments. Choice models, at least as they are used in applications, may need to be broadened to permit the study of consumption patterns rather than instantaneous flows of consumption.

Rachlin's argument has three weaknesses, however. First, he takes as an axiom the most puzzling aspect of altruism. Second, he fails to provide complete and coherent definitions of his terms, making it difficult to evaluate the implications of his analysis. Third, his focus on the relationship between altruism and patterned behavior is artificial. An argument for paying more attention to patterns would be much more powerful in another setting. The remainder of this commentary elaborates on these criticisms.

An essential assumption for Rachlin's approach is that individuals value a lifetime of altruistic behavior more than a lifetime of

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

selfishness. This assumption begs the central question of evolutionary biology because it does not explain why such behavior should have a fitness advantage over individual selfish behavior. Rachlin can shift the discussion of altruism to another category of behavior, but he must still provide a reason why natural selection favors individuals who have a preference for maintaining altruistic patterns. Rachlin's goal is to identify a behavioral mechanism by which altruism can be developed over a lifetime. This task is an ambitious one, so let us grant him the premise and see how well he does with it.

Consider his definition of self-control. The first part requires an individual to prefer a long activity to n repetitions of a brief activity. The second part requires him to prefer the brief activity to a t-length fraction of the longer activity. This definition is incomplete for two reasons. First, the definition is incomplete because it does not explain how to divide up the long activity. Rachlin's own examples demonstrate that decompositions are problematic. Consider a few more examples. Suppose that the short activity is touching your nose with your hand. Assume that this action for 30 seconds is preferred to listening to a 30-second segment of a symphony. Is a preference for listening to the entire symphony over touching your nose for an hour evidence of self-control? You (or your next of kin) would be extremely unhappy if your six-hour flight from New York to Paris ended after four and a half hours. For purposes of the definition of self-control, what is a fraction of a transatlantic plane flight? Does one exercise self-control by staying on the plane for the entire six hours? These examples, and those provided by Rachlin, warn that careful definitions of the objects of choice must come before a discussion of self-control.

The second, related, sense in which Rachlin's definition is incomplete is that it does not define the domain of preferences. This weakness seriously interferes with an understanding of Rachlin's hypothesis. To satisfy the first condition of Rachlin's definition of self-control, an individual must be able to rank a pair of activities performed over a T-period horizon. The second condition compares one-period activities. To say whether one brief activity is preferred to another, however, one must take into account the entire interval of length T. An individual may prefer to have a drink in the first five minutes of a party followed by abstinence than to have no drink at all, whereas the same individual may prefer to completely abstain from drinking to having eight drinks in the evening. Does this person prefer the brief activity of having a drink to a short interval of abstinence? It depends on whether the person expects the brief interval to be followed by more drinking. To talk about the individual's preferences over the first five minutes, one must be explicit about what the individual will do for the rest of the evening. Rachlin does not do this, and consequently, one is left with several different explanations for the self-control problem.

One explanation is simple impatience. Much (but not all) of Rachlin's discussion is consistent with the notion that self-control is the result of placing low weight on future utility. Another explanation is based on time inconsistency. An individual could enter the party with the idea that optimal behavior is to have a drink in the first five minutes and then abstain, but the individual may be aware that she'll feel differently after she has the first drink. This individual may try to postpone – or avoid – her first drink to exercise control over her "future selves." This idea is similar to Ainslie's (1992), whereas Rachlin plainly is after something else.[2] It is impossible to support or refute Rachlin's hypothesis until he defines his choice environment more carefully.

The third problem with Rachlin's approach is that altruistic patterns of behavior are abstract, whereas the desire to maintain patterns is stronger when patterns are concrete. Rachlin's article (and several of the essays in Schelling 1984) provide examples of intuitive ways in which people follow simple routines to obtain desirable long-term outcomes. Although the ideal may be to have one or two drinks at a party, with the amount of drinking determined by context (who is at the party, the quality of the liquor, what is planned for the subsequent day, etc.), it is easy to identify a pattern of complete abstinence. The external mechanisms that we use to control ourselves, for example diets, automatic savings plans, and religious rituals, often demand rigid adherence to clearly patterned behavior. This suggests that when following a pattern of behavior is a goal in itself, the pattern should be transparent.

Altruism is different. There are too many opportunities to give to others for us to follow a uniform pattern of goodness. We all are part Shen Te (Brecht's *Good Woman of Setzuan*) and part Shui Ta (her selfish alter ego). Rachlin's altruistic woman may be willing to die to maintain a pattern of good behavior, but she quickly forgets that she did not place a dollar into a homeless person's outstretched hand. A major challenge to Rachlin's experimental research agenda is to understand better what can and cannot become a pattern.[3]

Rachlin's article has a hopeful subtheme: His behavioral view of altruism suggests techniques for increasing altruistic behavior, which would then lead to good collective outcomes. This position requires closer examination. Selfless actions aggregated across individuals need not lead to good collective outcomes.[4] Selfish actions taken in the context of well designed institutions may lead to good collective outcomes.[5] One can be skeptical about Rachlin's argument or even about the importance of human altruism,[6] and still believe strongly that humans can identify and construct stable institutions that lead to good outcomes even in the face of self-interested behavior.

ACKNOWLEDGMENTS

I thank Philip Babcock, Vincent Crawford, Jim Moore, and Joel Watson for useful comments and the National Science Foundation for financial support.

NOTES

1. Sobel (2001) provides an overview.

2. Gul and Pesendorfer (2001) provide an elegant formulation of self-control in terms of preferences over sets of choices. Although essentially static, their framework is a coherent alternative to that offered by Rachlin.

3. There is scope for both behavioral and evolutionary explanations of pattern formation. Nesse (2001) points out how emotions can serve as commitment devices and that psychiatric conditions like depression may have selective advantages in some environments. Obsessive-compulsive disorders provide examples in which the need to create and follow patterns is excessive. These phenomena may be exaggerated versions of the mechanisms that give us pleasure from establishing and following patterns of behavior.

4. Some would argue that both characters in O'Henry's "Gift of the Magi" would have been better off if at least one of them had acted selfishly.

5. Economic theory's fundamental theorems of welfare economics provide conditions under which a consequence of rational self-interested behavior is economic efficiency.

6. In the same way, Smuts (1999) criticizes Sober and Wilson (1998) for "their apparent assumption that a more benevolent view of human nature depends on the existence of altruism" (p. 323).

The role of social and cognitive factors in the production of altruism

Arthur A. Stukas, Jr., Michael J. Platow, and Margaret Foddy

School of Psychological Science, La Trobe University, Melbourne, 3086, Australia. **{A.Stukas; M.Platow; M.Foddy}@latrobe.edu.au**
http://www.latrobe.edu.au/psy/staff/stukasa.html
http://www.latrobe.edu.au/psy/staff/platowm.html
http://www.latrobe.edu.au/psy/staff/foddym.html

Abstract: We agree with Rachlin's aim to account for altruism within existing theory. However, his argument is implicitly dependent on social and cognitive constructs that are explicitly identified in other social-psychological theories. The account does not advance theory beyond available constructs (e.g., self-categorizations, motives, values, role-identities, and social structure), and Rachlin's implicit use of these strains the behaviorist account.

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

As social psychologists interested in prosocial behavior, we applaud Rachlin's view that altruistic behavior and cooperation should not be treated as anomalies, but rather as phenomena that can be explained within existing theory. However, we question whether Rachlin's account is sufficient. Specifically, his "neo-behaviorist" perspective seems to adopt implicit cognitive assumptions, including: (1) recognition of self, (2) recognition of other (i.e., not-self), (3) expectations of the future (needed to plan repetitions of specific, ultimately altruistic, behaviors), and (4) memory of the past (needed to gain utility from behavior repetitions). Social psychological accounts of altruism make these cognitions explicit.

Within a behaviorist framework, Rachlin argues that altruism emerges and persists because it is a "pattern" that is reinforced in society. He makes an analogy between altruism and self-control of addictions, suggesting that both require recognition of a temporal dilemma – short term gain is discarded for long-term benefit when a person recognizes that the net gain is higher for the latter. However, this pairing of self-control with altruism is a mixed blessing. Whereas there are advantages to recognizing that, like self-control, altruism may have a temporal aspect, the social dilemma inherent in altruism makes it different from the problem of self-control. This extra social dimension requires us to determine to whom the pattern of altruistic behavior is valuable (the individual or society, both or neither), and why it is highly rewarded. Nevertheless, social psychological theorizing may provide a solution to this conundrum by aligning self-interest and other-interest in flexible cognitive representations of self.

There are several examples of theories that build others' interests into an individual's calculation of self-interest. An early attempt was Social Value Theory (McClintock 1972), which assumed that individuals factor others' gain into their calculation of utility when assessing whether to cooperate or help. There are relatively stable individual differences in the weights people apply to their own and others' gain; cooperators value others' welfare as much as their own, and altruists value the latter more (Platow 1993). Although this approach had some predictive success (Foddy & Veronese 1996), it did not explain how and why people might come to value the outcomes of others.

Another approach, Self-Categorization Theory (SCT; see Turner et al. 1987), assumes a fluid boundary between individual and collective interests. Cognitive representations of the self are theorized to vary in levels of abstraction, from the unique individual (akin to personal identity) to a categorization of self with others within a shared group boundary (akin to social identity). Altruistic behavior under SCT is assumed to obtain more from social, rather than personal, self-categorizations. Behaving altruistically provides utility to self simply because self and others are cognitively interchangeable; helping others is helping self (Platow et al. 1999). As noted above, when individuals identify or categorize the self with others, altruism is not an anomaly, it is simply self/group maximization, because the two are the same. Outgroup helping will be most likely when the salient self-categorization shifts to include the outgroup.

Despite the benefits from altruism that obtain for self when recategorized to include others, there are other social-psychological mechanisms that suggest individuals can be socialized to help strangers or outgroup members, even at the personal, non-group level of categorization. Rachlin's suggestion that long-term patterns of altruism can become intrinsically motivating resonates with recent social-psychological studies. However, rather than requiring each individual to rediscover the "long-term" value of altruism, these accounts of prosocial behavior often explicitly position the incentive structure in cultural and institutional practices, which does not deny that it can ultimately become internalised and intrinsically reinforcing.

Current social-psychological approaches also more explicitly define the explanatory terms employed by Rachlin, such as "pattern" and "intrinsic motivation" (which, when placed in a behaviorist account, raise more questions than are answered). This research has tried to understand sustained patterns of helpfulness in terms of such constructs as functions or motives and role-identities. To the extent that individuals choose actions in line with ongoing motives, roles, or goals, they will maintain a sustained pattern of helpfulness. In other words, these mental representations can serve as frameworks in which to interpret action and to allow for the selection of actions that maintain ongoing patterns (Vallacher & Wegner 1987). Thus, Snyder et al. (2000) have identified functions that volunteerism serves for individuals and they have demonstrated that volunteers intend to continue helping as long as these functions are met through their helpfulness. Piliavin and Callero (1991) have shown that repeated blood donation results in a role-person merger (that is, a sustained role-identity as a blood donor) that facilitates and promotes continued donation, which is not far, conceptually, from Rachlin's discussion of "habit" as preceding motive. Such mental representations have the capacity to guide and direct behavior into the sustained "altruistic" patterns that Rachlin discusses.

This is not to say that the attempt to provide a "bottom up" or emergent explanation of altruism should not be pursued. There is a strong tradition of "bottom up" solutions to various problems of cooperative human behavior (e.g. Axelrod 1984; Ostrom 1998); however, it is also recognised that these must be accompanied by "top down" imposition of incentive structures (e.g., Lichbach 1996) that result from deliberative action. That is, people learn through experience that certain incentive structures are effective in producing collective good, and build them into the social and cultural institutions that influence individuals to adopt particular prosocial behaviors in a sustained and committed way (Foddy et al. 1999).

Therefore, although our views are consistent with Rachlin's, we believe his perspective is lacking because it fails to acknowledge explicitly the importance of social structure and the role of cognitions in determining altruistic behavior. We have indicated that altruism may be promoted by cognitively recategorizing the self and other, or by cognitively transforming self and others' interests to be aligned. We have also observed that temporal patterns have been defined more clearly in social-psychological research and that such patterns may be retained cognitively to allow for ongoing behavioral choices (either altruistic or self-controlled).

Dissolving the elusiveness of altruism

Wim J. van der Steen

Department of Biology, Vrije Universiteit, 1081 HV Amsterdam, The Netherlands. **wvds@bio.vu.nl**

Abstract: Rachlin provides an impressive integrative view of altruism and selfishness that helps us correct older views. He presents a highly general theory, even though he is aware of context-dependence of key notions, including altruism. The context-dependence should extend much farther than Rachlin allows it to go. We had better replace theoretical notions of altruism and selfishness by common sense.

Seldom have I read a general view of altruism and selfishness as impressive as Rachlin's. He unites theories and data from history of philosophy, cognitive psychology, game theory, evolutionary biology, and many more disciplines. Rachlin is aware that all sorts of conceptual pitfalls plague theoreticians concerned with the subject, and he shows how to avoid them. Is altruism *possible*? To the extent that it does appear to exist, should it be regarded as a covert form of egoism? These are old questions in new garbs, with the result that we are now facing a virtually inextricable conglomerate of interdisciplinary theories with context-dependent key notions for confusingly different concepts. Rachlin's analytical lessons help us untangle much of this.

By and large, my own approach to the theme has been in the same spirit (for sources, analyses, and methodological commen-

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

taries, see Van der Steen & Ho 2001). But I opt for a methodological rigor that does not aim to solve any problems with altruism, as Rachlin claims to do. I opt for dissolutions, not solutions. Rachlin does present a general view. At the same time, he uncovers much context dependence. That generates a tension that I would like him to alleviate, but I fail to see how he could manage that.

On the one hand, Rachlin dismisses, rightly so I would suggest, much common wisdom from recent thinking about genetic and evolutionary aspects of altruism. His underlying point boils down to the thesis that categorizations used to explain altruism are incomplete. For example, evolutionary biology combined with cognitive psychology cannot accommodate learning in adequate ways. On the other hand, Rachlin cannot avoid introducing categorizations of his own. He notes that we end up with intractable theories if we focus on altruism as a category in its own right. Instead, he argues, we need to see acts of altruism as a subset of a broader category, self-control. That facilitates the accommodation of learning, as altruism shares with other forms of self-control an essential background of learning. All sorts of evolutionary categories (cf. common conceptualizations referring to survival value) are thus dismissed by Rachlin as overly limited. But should not the same fate ultimately strike a category such as self-control?

Here the issue becomes tricky. Rachlin makes the point that his own conceptualization of altruism inherits context-dependence from his explication of self-control. Acts can be altruistic in one context, and nonaltruistic in a different context. More precisely, acts of altruism should be seen as belonging to broader patterns to which we can attribute value or disvalue, depending on the kind of pattern we are considering.

I am in sympathy with this emphasis on context-dependence, but I would extend it much farther than Rachlin does through examples. Let me provide an example of my own. I may have to decide whether I will help my neighbor who is ill, by preparing a meal for her. If I help her, the old (and odd) theoretical question would be whether my act is altruistic in the sense of enhancing my inclusive fitness. Rachlin would urge us to regard my act as part of a broader pattern including the way I learned to help persons, for example. Then we get an entirely new view of altruism. It is here that I begin to feel uneasy. It is true, in a way, that I am acting on a decision to help the neighbor. But in describing in this way what I am doing, I am taking one particular element from a rich situation that includes an endless array of other things. In walking over to the neighbor's house, I decide to listen to a singing bird or to ignore it. I decide to leave my coat at home or to put it on. And so on, and so forth.

I am afraid that theoretical trouble pits regarding altruism have depths beyond Rachlin's field of vision. The fact that debates about the issue in modern, analytical traditions have haunted us for a long time indicates that we have been engaged with the wrong problem. That is also Rachlin's view, because he intends to replace current inadequate categorizations by adequate categorizations. My objection is that we are using categorizations (in a particular set of theoretical domains) at all. Notions of altruism are so diverse that adding new notions can but enhance the confusion existing in science and philosophy.

Rachlin uses many categorizations to explicate his new key notions. Let me take just one example. From game theory, he borrows results concerning the prisoner's dilemma, which he extends in a creative way. In addition to this, he comments on tit-for-tat, another well-known game. Here we have a categorization of games that is not exhaustive. Many more theoretical games exist. Nonexhaustive categorizations are applied by Rachlin to concrete situations involving altruism. As I indicated through my example, the description of the concrete situations carries with it categorizations of its own. We thus get all sorts of categorizations at different levels, which combine to form an integrative theory that can but apply to highly restricted domains.

I do hope that we will manage to stop all this. Science and allied philosophy can help us improve on common sense, but they can also become a stumbling block for common sense. That, in my view, has been the case with altruism for a long time. "Altruism" is like garbage, containing gems amidst the thrash. Let us sort out the gems and provide them with names, different ones for different gems, like rubies, emeralds, and so forth. In ordinary conversations, I do not have much trouble with understanding what other persons mean when they are talking about things denoted by scientists through labels of altruism, selfishness, and all that. The other persons, likewise, do not appear to have such troubles. We all use a rich variety of words and concepts for interactions among human beings, which in domains of abstract science and philosophy are amalgamated into an impoverished terminology. Long live common sense.

Altruism, self-control, and justice: What Aristotle really said

Graham F. Wagstaff

Department of Psychology, University of Liverpool, Liverpool L69 7ZA, United Kingdom. **GWF@Liverpool.ac.uk**

Abstract: As support for his position, Rachlin refers to the writings of Aristotle. However, Aristotle, like many social psychological theorists, would dispute the assumptions that altruism always involves self-control, and that altruism is confined to acts that have group benefits. Indeed, for Aristotle, as for equity theory and sociobiology, justice exists partly to curb the unrestrained actions of those altruists who are a social liability.

Most definitions of altruism centre on the idea that altruism is acting in the interests of others, even if costs are involved. For example, as Rachlin notes, most sociobiologists argue that, as acting in the interests of others invariably results in some cost to the individual, altruism can be defined simply as benefiting another at some cost to the benefactor. Most introductory textbooks on social psychology also focus on the general theme that altruism is behaviour motivated by the desire to help someone else, even if at a cost to oneself (see, e.g., Sabini 1995). However, none of these definitions assumes a necessary connection between altruism and self-control. In fact, the idea that norm-related helping behaviours can be acquired through learning is already well established in the psychological literature (see, e.g., Cialdini et al. 1981), but there is no assumption within such approaches that altruism must always involve self-control. In contrast, Rachlin actually defines altruism as a subcategory of self-control. This, in itself, is problematic, as the hypothesis that "altruism always involves self-control" then becomes irrefutable (if a man is defined as "a male with two legs," there is no point asking if men have two legs).

In support of his definition, Rachlin argues that his ideas are not unusual; in fact, he says that the link between self-control and altruism is to be found in the writings of Plato and Aristotle. However, the writings of Aristotle, in particular, illustrate well some of the problems with Rachlin's position. According to Aristotle (1984), nature has determined that man has a purpose, or *telos,* namely, *eudaimonia,* or loosely, happiness. However, this purpose can only be achieved by practising virtue and overcoming our untutored passions. At first overcoming our passions will be difficult; but, in the end, by practising virtuous acts we will not only learn the appropriate ways to act, but our dispositions will change and fall in line. When this is achieved, what we ought to do is what we are able to do and automatically feel like doing; hence the result is, of course, happiness. Moreover, we are "hard-wired" by the laws of the cosmos to be capable of achieving this end. Thus, fundamental to Aristotle's view is that the truly good man (one who has achieved his telos) no longer needs to exercise self-control to be virtuous, as his actions fall in line with his dispositions. In Aristotle's scheme, therefore, there is no reason why a woman who runs into a burning building to save another's child, at great risk to herself, should necessarily be exercising self-control. Far from it,

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

she might need to exercise considerable self-control to stop herself from doing this, because not to act in this way would make her more unhappy.

Aristotle (and Socrates and Plato, for that matter), would thus challenge Rachlin's cynical premise that particular selfless acts are always experienced as less pleasurable than particular selfish acts; and, in support of his case, a modern Aristotle might wish to refer to the numerous empirical findings of social psychologists that show that helping others, even in particular cases, improves mood and can be more pleasurable (or at least less painful) than not helping others (see, e.g., Sabini 1995). Indeed, the learning account of these findings put forward by Cialdini et al. (1981) and Daumann et al. (1981) bears a far more striking resemblance to Aristotle's ethical theory than that of Rachlin.

Another feature of Rachlin's argument is that, by definition, altruism involves acting in a (self-controlled) way that benefits the group. But, again, this runs contrary to our everyday use of the term "altruism." For example, certainly it is plausible that a woman might run into a burning building to save a child because she finds a pattern of similar acts reinforcing, but it is not at all clear why the reinforcing "highly valued" pattern of acts should have anything to do with group benefits, or why this should be necessary to qualify the act as one of altruism. If large groups of mothers all died in well-meant but hopeless attempts to save each others' children out of highly valued patterns relating to "reducing the pain of empathy" (Batson 1987), their acts might be detrimental to the group (which would end up with few mothers); but why should their acts be classed as less altruistic than those of women who save children because, in the past, such patterns of acts have resulted in extrinsic rewards?

It was perhaps in recognition of this problem that Aristotle did not include altruism among his list of virtues. In Aristotle's scheme, the good man acts in the right way, at the right time, for the right reason; the man who impulsively sacrifices himself to help anyone, including gangsters and tyrants, may be acting selflessly, but he is also a social liability. This is why Socrates, Plato, Aristotle, Aquinas, Kant, J. S. Mill, and numerous others stressed that the highest ideal is not altruism, it is justice; meaning to give to each according to his due or desert; to reward the good and only the good, and punish the bad, and only the bad, in due proportion. The ubiquitous justice principles of proportional positive and negative reciprocity are to be found not only in virtually every moral code throughout history, but in much of the psychological literature on justice, such as equity and just world theory, and in sociobiological theory (Wagstaff 2001). Note, for example, how Dawkins's (1976/1989) "grudger" seeks to discourage the excesses of the "sucker" who indiscriminately, and without restraint, sacrifices himself to everyone, including "cheats." If Rachlin is to argue that altruism always involves self-control, he has first to deal with the notion of justice, for justice requires that, at times, we may have to exercise self-control to prevent our impulsive acts of altruism.

In summary, although Rachlin provides a plausible and innovative account of how certain classes of helping behaviours might be learned, his challenge is to go beyond a definition of altruism that begs the question of self-control and show that his analysis can be applied to a much more comprehensive range of behaviours we normally connect with the term "altruism," and why it is to be preferred to existing psychological as well as sociobiological accounts of altruism.

Valuable reputation gained by altruistic behavioral patterns

Claus Wedekind

Institute of Cell, Animal and Population Biology, University of Edinburgh, Edinburgh, EH9 3JT, Scotland, United Kingdom. **c.wedekind@ed.ac.uk**

Abstract: On a proximate level, altruism may well be a temporally extended pattern of behavior that often seems to be maintained without extrinsic rewards (we may find it just valuable to be an altruistic person). However, recent theory and experiments have uncovered significant and often nonobvious extrinsic rewards for "altruistic" behavioral patterns. Ultimately, these patterns may mostly lead to a net benefit.

Evolutionary theory predicts that "altruistic" behavior can only evolve when it normally leads to a net fitness benefit. In order to achieve this net benefit (i.e., on a proximate level), we as human beings may normally consider being an altruistic person as highly valuable. On an ultimate level, however, this altruism may mostly be self-interested. We do not need to be aware of this, that is, our altruistic patterns of behavior may not be calculated, but there are often a number of non-obvious external rewards that may ultimately explain the existence of most altruistic patterns. To translate this into Rachlin's example with the four soldiers (target article, sect. 9): Soldiers 1 and 3 weight immediate costs and benefits, that is, their behavior is calculated, whereas soldiers 2 and 4 follow behavioral patterns that are likely to have evolved because they normally lead to net fitness benefits. These soldiers may or may not be aware of the usual net benefits that come with their behavioral pattern, so it is not obvious whether their behavior is calculated. Their behavior does not need to be immediately calculating in order to evolve.

In the following, I briefly discuss some of the major game-theoretical contexts and the sorts of altruistic patterns that have been found to be, or are at least very likely to be, evolutionarily stable and economically rewarding. Humans are typically not very altruistic in situations where theory cannot find any long-term net benefits.

Direct reciprocity. The best behavioral strategies in direct reciprocity games like the famous "repeated two-player Prisoner's Dilemma" (PD) are normally cooperative ones (Axelrod 1984; Nowak et al. 1995; Posch 1999). Therefore, it is easy to see that a single altruistic act (a cooperative move) that forms part of such a strategy is typically rewarded, either immediately or in the long run. Human cooperative strategies in PD games are quite sly. In experiments (Wedekind & Milinski 1996), they adopt cooperative strategies that usually beat "generous Tit-for-Tat" or "Pavlov," that is, the winners of extensive computer simulations (Nowak & Sigmund 1992; 1993). If human memory capacity is experimentally reduced, however, the strategy becomes more and more similar to that of Tit-for-Tat (Milinski & Wedekind 1998). Hence, our temporally extended pattern of altruistic behavior in the PD is normally strategic and depends on memory capacity.

Multiplayer direct reciprocity games like, for example, repeated Public Goods games (Hardin 1968) are predicted by game theory to end in reduced cooperation (Hauert & Schuster 1997). And indeed, cooperation in Public Goods experiments normally breaks down rapidly (Milinski et al. 2002; Wedekind & Milinski, unpublished results). This predictability suggests that altruistic behavior in direct reciprocity is not performed because it is valuable by itself, but because it fits into evolutionary stable behavioral strategies.

Indirect reciprocity. The idea of indirect reciprocity is that helping someone, or refusing to do so, has an impact on one's reputation within a group. A reputation of being generous increases the chances of receiving help by third parties sometime later in the future when help may become needed (Alexander 1987; Zahavi 1995). Nowak & Sigmund (1998a; 1998b) proved analytically and in computer simulations that generosity could indeed evolve in such scenarios. Indirect reciprocity therefore provides an evo-

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

lutionary and economical explanation of, for example, generosity to beggars.

Experimental evidence that humans have evolved generous behavior in indirect reciprocity is accumulating (Milinski et al. 2001; Seinen & Schramm 2001; Wedekind & Braithwaite 2002; Wedekind & Milinski 2000), but I do not know of any example where comparable generosity exists in other species. Maybe, only humans have the necessary mental capacities to assess and reassess other people's reputation, and only humans possess the effective language that may be necessary to maintain indirect reciprocity. Without gossip, it seems difficult to keep track of the decisions of other group members, that is, to obtain a useful idea about their reputation.

A recent experimental study by Milinski et al. (2002) demonstrated that reputation can even help to solve the "tragedy of the commons" (Hardin 1968); that is, it maintains contribution to the public good at an otherwise nonexplainable high level. Such contributions look very altruistic, but they help to build up a reputation that may be useful in other contexts. *Altruistic* behavioral patterns in indirect reciprocity or Public Goods games can therefore be seen as a kind of payment into a social insurance policy that promises help should help become necessary somewhere in the future.

Punishment, rewards, and ultimatum games. Cooperation and *fair* behavior (e.g., in an ultimatum game; see Nowak et al. 2000) may often arise because of punishment and reward (Gintis et al. 2001; Sigmund et al. 2001). Obviously, punishment may directly teach a social partner not to defect again. It may also teach all future social partners who observe or hear about the punisher's reaction. Punishment then contributes to another kind of reputation that may serve as a warning and thereby prevent later defection (Sigmund et al. 2001). Moreover, punishing a free-rider is a kind of generous move in a Public Goods game (Gintis et al. 2001), which is likely to be rewarded in later direct or indirect reciprocity (Gintis et al. 2001; Milinski et al. 2002). All these long-term benefits that become likely with a reputation of being vindictive may eventually compensate for the costs of punishment and may lead to a net benefit.

Conclusions. Recent theoretical and experimental studies suggest that high levels of altruism can be evolutionarily stable. Although they are costly in the short term, they can eventually lead to net benefits in the long term. Rachlin's temporally extended pattern of altruism may be manifested in simple strategic rules or in abstract ideas about a person's reputation. Some kinds of altruistic patterns that we can observe in humans, but probably not in other creatures, seem to depend on our high mental capacity for assessing and reassessing different kinds of reputation.

ACKNOWLEDGMENTS

I thank Victoria Braithwaite and Mirjam Walker for comments on the manuscript, and the Swiss National Science Foundation (823A-064723) for financial support.

Decisions to follow a rule

Paul Weirich

Philosophy Department, University of Missouri-Columbia, Columbia, MO 65211. **weirichp@missouri.edu**
http://web.missouri.edu/~philwww/people/weirich.html

Abstract: Rachlin favors following patterns over making decisions case by case. However, his accounts of self-control and altruism do not establish the rationality of making decisions according to patterns. The best arguments for using patterns as a standard of evaluation appeal to savings in cognitive costs and compensation for irrational dispositions. What the arguments show depends on how they are elaborated and refined.

Rachlin raises many issues worthy of lengthy discussion. I will attend only to his views about the role of patterns in decision making. As I read him, he claims that an agent may rationally choose an act that fails to maximize utility if the act belongs to a pattern that maximizes utility. For example, an agent may decline a drink despite its appeal because declining fits an advantageous pattern of moderation. Similarly, a mother may enter a burning building to save another's child because such heroism is part of an altruistic life that she finds rewarding. Rachlin contends that many acts of self-control and altruism are nonmaximizing but belong to a maximizing pattern that justifies them.

Rachlin's view is similar to views expressed by philosophers such as Nozick (1993) and McClennen (1998). Its counterpart in ethics is rule-utilitarianism.

Altruism and self-control, as Rachlin portrays them, do not make a strong argument for deciding according to patterns instead of according to cases. He takes both phenomena to stem from enlightened self-interest. Habits of self-control and helpfulness to others, he says, make for a happy life. An agent's act of self-control or altruism promotes habits beneficial to her. Thus, taking account of all the act's consequences, including promotion of those habits, the act maximizes utility after all. Its rationality does not depend on a novel method of deciding according to patterns. The phenomena just show that maximizing among acts must take account of each act's total consequences, including its fitting into a valued pattern. In Rachlin's examples, the person who declines a second martini out of moderation does not fail to maximize utility. Declining promotes self-interest because it is a means of achieving moderation. Similarly, the mother who saves another's child from a fire does not fail to maximize utility. Her heroism promotes self-interest because she finds helping others rewarding.

Suppose we adopt an account of altruism more traditional than Rachlin's so that a genuinely altruistic act does not serve the agent's own interest, not even his enlightened self-interest. Even so, an altruistic act may still maximize utility for the agent. Self-interest and utility for an agent are distinct. Utility for an agent depends on his goals, which may include helping others even at his own expense. Thus, an agent's altruistic act may maximize utility for him despite not promoting his own interests. (See Weirich, forthcoming.)

Given that Rachlin's cases of self-control and altruism do not establish the rationality of deciding according to patterns, where can one turn to for an argument? Rationality requires maximization of some kind, but does it require maximization among acts or among patterns of acts? The received view is that it requires maximization among acts. But what can be said for the rival view?

How about an appeal to limits on powers of discrimination? Suppose a difference in consumption of one drink never makes a noticeable difference in sobriety. So one always has a reason to enjoy another drink. Nonetheless, having five drinks does make a noticeable difference in sobriety. As a result, there's a slippery slope to inebriation. May one resist a slide by deciding according to patterns? No, the appropriate way of resisting is recognition of consequences besides noticeable differences in sobriety, consequences such as raising the level of alcohol in one's bloodstream.

McClennen (1998) recommends deciding according to rules that maximize. He thinks the benefits here are greater than from selecting acts that maximize. For example, suppose I get \$5 now if I will reject \$1 later. An agent following a maximizing rule rejects \$1 later to receive \$5 now. An agent maximizing among acts moment by moment takes \$1 later, and so does not get \$5 now. Such examples do not support the rationality of deciding by patterns, however. In the example, a future nonmaximizing act is rewarded now. Its being rewarded does not make it rational. Because I would not lose \$5 already pocketed if I were to take an additional \$1 later, not taking it later is irrational.

Consider a similar example. Rawls (1971) observes that loving someone is a valuable experience, although it carries a risk of self-sacrifice. Genuine love calls for self-sacrifice should the occasion for it arise. Despite being able to resist self-sacrifice in the sense relevant for choice, someone in love cannot resist self-sacrifice in a way compatible with her psychological state. The emotion cre-

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

ates an insensitivity to self-interest, as an addiction or weakness of the will might. However, cases of love-generated self-sacrifice do not establish the rationality of deciding according to patterns. Because of love's value, one may maximize utility for oneself by letting oneself fall in love. Still, if love prompts an act that fails to maximize utility (not just puts aside one's own interest), that act is irrational.

Human cognitive limits ground a better argument for deciding according to patterns. Making decisions case by case carries a heavy cognitive cost. One may cut costs by adopting policies that apply routinely. Occasionally, following a maximizing policy may yield a nonmaximizing act. The policy's reduction in cognitive costs may compensate for deviations from case by case maximization.

Although this line of argument for deciding by patterns has promise, it does not undercut the reasons for maximizing case by case. Instead, it shows that case-by-case maximization should apply to decisions rather than to the acts they yield. Because of lower decision costs, decisions that follow policies are maximizing decisions even if they do not always yield maximizing acts. (See Weirich 2001.)

Here's another way of arguing for deciding according to patterns. Suppose that someone may maximize utility by having a glass of wine with dinner and then forgoing a drink at a party later. However, she knows that she will have a drink at the party. Although wine with dinner maximizes utility, forgoing it is rational, given her lack of resolve. Not maximizing now compensates for her disposition to drink later. Although decision by patterns is not rational for ideal agents, it may be rational for non-ideal agents.

This line of argument for deciding by patterns raises some issues. It condones patterns that do not maximize. In the example, the maximizing pattern is still wine with dinner and abstinence later. At best, forgoing wine with dinner fits a pattern that maximizes among "realistic" patterns. Whether such hedged maximization is rational remains an open question.

Altruism, evolutionary psychology, and learning

David Sloan Wilson[a] and Ralph R. Miller[b]

[a]*Departments of Biology and Anthropology, Binghamton University, Binghamton, NY 13902-6000;* [b]*Department of Psychology, Binghamton University, Binghamton, NY 13902-6000.* **Dwilson@binghamton.edu Rmiller@binghamton.edu**

Abstract: Rachlin's substantive points about the relationship between altruism and self-control are obscured by simplistic and outdated portrayals of evolutionary psychology in relation to learning theory.

Rachlin points out an interesting similarity between altruism and self-control. In both cases, actions are chosen that are aversive to the individual in the short run and beneficial only in the long run, if at all. Despite this interesting similarity and the mechanistic connections that might exist between altruism and self-control, we find problems with the target article. Rachlin's several useful points are obscured by false dichotomies and straw-man portrayals of complex issues. His largely additive view of the interaction between learning and inherited mechanisms reminds one of the weary nature-nurture distinction that has become a standard target for criticism. Nurture (learning) and nature (genetic predispositions) are no more additive than is the area of a table determined more by its length than its width.

Rachlin states that "we inherit good eyesight or poor eyesight," when in fact we inherit a predisposition toward good or poor eyesight, given conventional developmental experience. He sets up a straw man in claiming that biological compatibility posits a "general mechanism for altruism itself" (sect. 1.1). He argues that in humans altruism is learned rather than fueled by an innate self-sacrificing mechanism. There may well be a learned component (which plays out on genetic predispositions), but there may also be a component that results from genetic predispositions interacting with experience other than experience resulting in learning.

Although Rachlin's description of how altruism can evolve by group selection is accurate, in other places he writes as if selfishness is the only possible product of evolution (e.g., "From a biological viewpoint selfishness translates into survival value"; sect. 1.1), thereby concluding that altruism can only arise from the consequences of reinforcement. He also overlooks the possibility of psychological altruism, which involves valuing the welfare of others for its own sake rather than as a means to personal ends (Batson 1991; Sober &Wilson 1998). When psychological altruists help others, they experience the same kind of reinforcement as when they help themselves, which makes learning to help others as easy to learn as helping oneself over the short term. Genuinely other-oriented learning mechanisms solve the paradox that Rachlin sets up in a way that he does not anticipate. In general, Sober and Wilson's analysis of psychological altruism from an evolutionary perspective is as relevant to the target article as their analysis of the behavioral manifestations of altruism.

Although evolutionary psychologists such as Tooby and Cosmides (1992) make strong claims about the existence of specialized cognitive mechanisms (as noted by Rachlin), their argument is based primarily on functional considerations. There are so many adaptive problems to be solved, each one requiring attention to different aspects of the environment and different ways of processing the information, that all of these presumably cannot be accomplished by a single-domain general learning process. However, all learning, by its nature, requires a feedback process based on the commission and detection of errors. When errors are costly or difficult to detect, predispositions toward adaptive behaviors are apt to come to be built into the organism, and, if learning takes place at all, it must be in the form of triggering an already predisposed response. The claim is not that all specialized cognitive adaptations are devoid of learning, but that many are subject to their own specialized form of learning.

Applying these functional considerations to altruism and self-control, we might reason that self-control (i.e., preference for a larger reward later) is adaptive in some situations and not in others. The importance of current costs and benefits relative to future costs and benefits depends critically on the probability of surviving into the future. There is every reason to expect a predisposition toward innate cognitive mechanisms that evaluate risk and discount future costs and benefits accordingly. Wilson and Daly (1997) found that violent risk taking in men and teenage pregnancy in women, both of which are regarded as behaviors lacking in self-control, correlate strongly with life expectancy. Similarly, Chisholm (1999) interprets the seemingly maladaptive insecure attachment styles first identified by Bolby (1969), as an adaptive response to insecure social environments. The general functional considerations emphasized by evolutionary psychologists certainly apply to self-control. This does not mean that learning is irrelevant, but it does mean that learning needs to be studied in conjunction with these functional considerations, not as a separate add-on.

The adaptive expression of altruism has its own set of functional considerations, which only partially overlap with self-control. It is easy to imagine situations that favor altruism but not self-control, and vice versa. Self-control is largely a temporal problem with long-term consequences weighed against short-term consequences. In contrast, altruism need not be extended in time, although it can be. Apparently, Rachlin's view of altruism is heavily dependent on the iterative Prisoner's Dilemma game, which is only one of several potential models of altruism. Often altruistic behavior involves a single immediate action rather than a series of repetitions of a (distasteful) brief activity, which falls outside Rachlin's paradigm. His own example of a woman running into a burning building provides little scope for the learning of sequences.

Altruism has a spatial component in addition to its highly variable temporal component; all altruists fare worse than selfish in-

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

dividuals in their immediate vicinity, but the average fitness of altruists can still exceed the average fitness of selfish individuals on a larger spatial scale if the altruists preferentially interact with each other. The fact that the advantage of selfishness is local and therefore easy to perceive, whereas the advantage of altruism is more global and therefore difficult to perceive, presents interesting problems for learning altruism that do not apply to self-control. In addition to recognizing similarities between altruism and self-control from the standpoint of learning theory, we also need to recognize their differences.

A potentially testable mechanism to account for altruistic behavior

Thomas R. Zentall

Department of Psychology, University of Kentucky, Lexington, KY 40506-0044. **Zentall@pop.uky.edu**
http://www.uky.edu/AS/Psychology/faculty/tzentall.html

Abstract: It is assumed that self-control always has a higher value. What if it does not? Furthermore, although there are clearly intrinsic reinforcers, their measurement is problematic, especially for a behavioral analyst. Finally, is it more parsimonious to postulate that these behaviors are acquired rather than genetically based?

In the web of psychological research, Rachlin has made a (re)connection between two seemingly disparate fields, self-control and altruism, that have been studied in relative isolation. His extensive research on self-control gives him a solid platform from which to venture into the abstract area of altruism. The idea that a pattern of altruism may be intrinsically reinforcing allows us relief from the paradox that altruism is typically defined in terms of benefit to others and loss to the altruist. Intrinsic reinforcement, albeit somewhat abstract in its definition and assessment, potentially gives us the opportunity to study altruism within the context of reinforcement theory. The benefit of such a view is that it may demystify altruism, a behavior so often linked to morality. Several thorny issues remain because altruism is a somewhat abstract concept; however, the application to altruism of findings from the self-control literature, an area that has been extensively researched, has the potential to be quite enlightening.

One concept in the area of self-control that has concerned me prior to the proposal of its role in altruism is the assumed parallels in the human and animal literature. In the case of alcoholism, it is clear that the long-term goal of health and better relationships should have greater value than the short-term goal of an immediate drink. Less clear is the assumption that an animal should inherently prefer a large, delayed reward over a small immediate reward. Certainly, humans do not always do so, and when they don't, their behavior may not be seen as irrational.

Consider the home buyer who is willing to spend several times the selling price of a house (a mortgage paid off over time) for the immediacy of living in the house. Now consider the animal's options. In nature, delay to reinforcement may mean more than the postponement of that reinforcer because, unlike in artificial laboratory conditions in which delayed reinforcers can be delivered with certainty, in nature, delay to reinforcement is often correlated with a decrease in the probability of reinforcement on account of competition from other animals or the unpredictability of climatic conditions. The fact that, unlike nature, we experimenters can be trusted to provide the animals with the large, delayed reward as promised every time that the larger alternative is selected, does not alter the animal's predisposition not to trust its environment (us). Furthermore, for some species (e.g., small birds) a sufficient delay of reinforcement, regardless of the promised reward magnitude, can threaten the survival of the animal. In other words, animals that show self-control may not survive.

Assuming, however, that we had some way to evaluate the conflict between small, immediate rewards and large, delayed rewards as we presumably do with humans (e.g., "I really want to stop drinking but I find it too difficult to refuse the drink"), we are still faced with several problems. First, the patterns of behavior that are intrinsically rewarding are difficult to define without some circularity. Is it possible to know which soldier will behave heroically? One often hears of soldiers who wonder if they will act with honor when going into battle. Are the conditions under which altruism will appear, predictable? I suspect that they are, but if they are not, the value of the theory is greatly diminished.

Second, the concept of intrinsic reinforcement is needed to explain the variety of behavior that has no extrinsic material or social reward, such as crossword puzzle solving. But intrinsic reinforcers are difficult to assess. They are what is left once you have ruled out extrinsic reinforcers, and in the case of humans, typically we assess them by means of verbal behavior (e.g., "I just like doing it"). But this sort of definition can easily become circular, especially when we are talking about behavioral patterns that are themselves not clearly defined. One can hypothesize that extrinsic reinforcers become internalized, but that does not explain, it only describes. And why is altruistic behavior rewarded by society? Is it because societies are made up of selfish individuals who want to encourage altruism in others because it benefits themselves?

Finally, American psychology seems particularly averse to considering genetic bases of behavior. This antipathy may come from the feeling that genetic causation is invoked when we don't understand what is producing the behavior. But we know that at a very basic level, certain kinds of altruism, such as parental behavior, must be genetically based. Although we do not yet have access to specific genes for altruism, we are able to make predictions about the relation between perceived family relatedness and degree of altruism. Furthermore, it is interesting that Rachlin has chosen as a model of altruism the Prisoner's Dilemma (and studies on this) because the same problem has been used to argue for a genetic basis for altruism. When various strategies were played against each other, the tit-for-tat strategy fared very well (Axelrod & Hamilton 1981). But more important to the genetic argument, in other research (Axelrod 1984), following a round of competition the various successful strategies were reproductively "rewarded" by allowing an additional copy to be present in the next round (generation). After many rounds and through the course of several repetitions of the game, the strategy pool was dominated by tit-for-tat and other, similar strategies. Having a gene for a cooperative (or altruistic) strategy may not be needed to account for these results, but neither does it require learning.

The points raised in this commentary notwithstanding, Rachlin is to be commended for presenting a novel perspective (at least in modern times) on the relation between self-control and altruism. His views are likely to generate research that should clarify both areas, as well as the relation between them.

From reinforcement of acts to reinforcement of social preferences

Daniel John Zizzo

Department of Economics, University of Oxford, Oxford, OX1 3UQ, United Kingdom. **daniel.zizzo@economics.ox.ac.uk**
http://www.economics.ox.ac.uk/Research/Ree/ZizzoWebpage.htm

Abstract: Rachlin rightly highlights behavioural reinforcement, conditional cooperation, and framing. However, genes may explain part of the variance in altruistic behaviour. Framing cannot be used to support his theory of altruism. Reinforcement of acts is not identical to reinforcement of patterns of acts. Further, many patterns of acts could be reinforced, and Rachlin's altruism is not the most likely candidate.

After a few decades of sociobiological revival in various guises (e.g., Alexander 1987; Bergstrom 1995; Wilson 1975), it is re-

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

freshing to read an article unabashedly stressing the role of the environment in prosocial behaviour. That *in principle* one can imagine behavioural reinforcement as explaining virtually all of the variance in altruistic behaviour, is certainly possible: In the limit, we could have the case where evolution has been evolution for a purely environmentally plastic brain (e.g., Quartz & Sejnowski 1997). One can make a powerful case for some endogeneity of prosocial behavior (Zizzo, in press). Yet, although one can make an argument for the *partial* endogeneity of interdependent preferences, it is not obvious how the evidence is incompatible with a partial role of genetic inheritance in explaining behavioral variance, nor why this should not be considered as the most natural interpretation (e.g., Rushton et al. 1986).

Rachlin's evidence suggests that (1) conditional cooperation is important, and (2) it is subject to framing effects (i.e., whether you believe you are playing against a computer or otherwise). Concerning point 2, Rachlin is right to suggest that framing effects are pervasive (e.g., Cookson 2000). He claims that they are due to different frequencies of reinforcement; this may very well be true, but it is no more than a conjecture, and so it is unclear how it can be used as a proof of Rachlin's theory of altruism relative to other theories. Concerning point 1, reinforcement over an act is not identical to reinforcement over a pattern of acts, and to prove the latter, Rachlin would really need to discuss the evidence on knowledge transfer from one game to a different game, to see whether reinforcement in one situation translates into reinforcement in another situation.

In favour of Rachlin's thesis, there are contexts where this is the case, at least in the short run implied by the laboratory settings (e.g., Guth et al. 1998). In some current experimental work, I have subjects first play a set of games that change to the degree in which the subjects are cooperative or close to zero-sum, and then they play a set of new, never-before-encountered games (with different players, eliminating repeated game effects). When the first set of games is more cooperative, behaviour in the second set of games is also more cooperative. While not all the evidence can be reconciled with a simple reinforcement learning account, it is what Rachlin's theory needs.

A deeper problem is whether the pattern of acts that is reinforced is what Rachlin claims it to be ("altruism") or something entirely different. There are many possible preferences that would be able to explain why cooperation in the finitely repeated Prisoner's Dilemma (PD) is conditional on an expectation of cooperation from the other player. Preferences, as we economists use them, are a behavioral concept: They are preferences as revealed in behaviour and so are closely related to Rachlin's patterns of acts. They include utility functions with two elements, one based on material gain and the other on a payoff transformation component implying inequality aversion (Fehr & Schmidt 1999), reciprocity (Falk & Fischbacher 1998), trust responsiveness (Bacharach et al. 2001), pure or impure altruism (Palfrey & Prisbrey 1997), or perceived fairness (Konow 2000). They will all lead to different predictions depending on whether subjects believe they are playing against a computer, or if they believe they are playing with a human being, because you will be fairness-sensitive towards a human being, but not against a computer. Even if the agents are, and remain, purely self-interested, they may find it optimal to cooperate because of the repeated nature of the game, whether with humans (Kreps et al. 1982) or (in different ways) with computers (because subjects can try to "crack the system" of how to make the most money). Otherwise, for a *wide* range of payoff transformations, with a modicum of rationality, the PD becomes a different game where mutual cooperation is a possible equilibrium, and the greater the expectation is of cooperation from the co-player, the greater will be the expected payoff for cooperating and hence the likelihood of cooperation. Therefore, the interpretation of cooperation in the finitely repeated PD is likely to be difficult. This matters, because the preferences that subjects have or acquire may make very different quantitative predictions in many different game settings (e.g., for other trust games; see Bacharach et al. 2001). This is why experimental economists have been focusing on a variety of different games to assess what preferences subjects have (e.g., Charness & Rabin 2000; Zizzo 2000a): The PD paradigm is simply not discriminative enough.

Rachlin's section 4 definition of altruism appears based on the intrinsic value of an act that is beneficial to a group: This would correspond to what economists would label "impure altruism" or "warm glow," albeit further specified with relation to a group. Unfortunately, there is no specific reason to believe that this *is* the pattern of acts that gets reinforced rather than, say, others with greater predictive power such as inequality aversion (e.g., Fehr & Schmidt 1999; Zizzo 2000b). Perhaps his theory can be rescued by making it more general, but this may be at the cost of virtual unfalsifiability. If Rachlin wants to convince nonbehavioural psychologists, he might need to show how his theory is better than alternative theories that make precise quantitative predictions, and how it can then receive unequivocal support or falsification in the laboratory. Nevertheless, he is right in stressing the role of behavioural reinforcement, and behavioural psychologists like him can bring useful new perspectives to our understanding of prosocial behavior. In particular, framing effects are real, and none of the models I mentioned can really explain them except in specific cases or with auxiliary or unmodelled hypotheses. Zizzo (2000b) tried to fill the modelling gap among reinforcement, framing effects, and preferences using neural network agents learning to play "altruistically" or "enviously" in new games, but this work is very preliminary and tentative.

The importance of social learning in the evolution of cooperation and communication

Willem Zuidema

Language Evolution and Computation Research Unit, Theoretical and Applied Linguistics, University of Edinburgh, Edinburgh EH8 9LL, United Kingdom. **jelle@ling.ed.ac.uk** **http://www.ling.ed.ac.uk/~jelle**

Abstract: The new emphasis that Rachlin gives to social learning is welcome, because its role in the emergence of altruism and communication is often underestimated. However, Rachlin's account is underspecified and therefore not satisfactory. I argue that recent computational models of the evolution of language show an alternative approach and present an appealing perspective on the evolution and acquisition of a complex, altruistic behavior like syntactic language.

Rachlin calls attention to the role of social learning in the emergence of altruistic behavior in humans. This shift of emphasis in thinking about altruism has intriguing consequences. Acknowledging the important role of learning leads one to ask at least three new and challenging questions: (1) about the exact mechanisms by which altruistic behavior emerges in learning and development; (2) about the ways in which the existence of learning mechanisms has changed the evolutionary process; and, vice versa, (3) about the ways in which evolution has shaped the learning mechanisms that lead to altruism. We can no longer – as is common in traditional game-theory – ignore the intricate mapping between genotypes (the genes) and phenotypes (the behaviors) and the strong dependence of this mapping on the individual's (cultural) environment.

Rachlin's article is a welcome effort to underline this point, but I think his explanation for the emergence of altruistic behavior in humans suffers from *underspecification:* Some crucial concepts are too loosely defined to make it possible to really agree or disagree with his analysis. I will discuss Rachlin's answers to the previous questions from this perspective and then try to show that some recent computational models in the related field of the evolution of communication offer a more precise account of the evolution of altruistic behavior.

Rachlin's answer to the first question is a mechanism similar to

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

self-control. Humans discover that choosing for a whole pattern of altruistic activities is in the end more rewarding than repeating alternative, selfish activities, even though the latter offer more short-term benefits. The problem with this account is that it is unclear what constitutes a "pattern." Without a theory on how individuals represent and acquire this knowledge, we can never identify the different strategies that individuals can choose from.

A related problem arises for Rachlin's implicit answer to the second question. Rachlin gives the example of a woman who puts her life in danger to rescue someone else's child. His explanation of her brave behavior rests on the crucial assumption that the woman at some point in her development had to choose between life-long altruism or life-long selfishness. If there are only these two choices, and if the first choice is indeed more profitable in the long run, natural selection of course favors the tendency to choose it. However, Rachlin gives no arguments why the choice would be so constrained. I find it difficult to accept that with all the subtle influences that genes have on our behavior, selectively avoiding life-threatening situations was not a possibility.

Rachlin's implicit answer to the third question is no solution to that objection. Essentially, he explains the evolution of altruistic behavior by claiming that it is not really altruistic after all. Altruistic – at least in its traditional sense in evolutionary game theory (Maynard Smith 1982) – are those behavioral strategies that benefit others, but harm the individual that employs them *even though less harmful strategies are available.* A game-theoretic analysis of the evolution of alarm calls in certain bird species (Maynard Smith 1982) therefore emphasizes evidence that the calls really are harmful and that other strategies are really available. In contrast, the altruistic strategy in Rachlin's scenario is in the long run advantageous, and better alternatives are not available; it is thus not *really* altruistic in the traditional sense.

Rachlin acknowledges this, but he does not mention that the analogy between his explanation and group selection therefore breaks down. Group selection, like kin selection, is a mechanism that is capable of explaining real altruism. The decrease in the fitness for the *individual* is explained by assuming a higher or lower level of selection, that is, that of the *group* or that of the *gene.* Therefore, the fitness of a worker bee that does not produce any offspring really is low (it is zero by definition), but the fitness of the whole colony or the fitness of the genes that cause her sterility is high. The empirical validity of these explanations remains controversial, although their explanatory power is appealing.

Researchers in the related field of language evolution have already explored many aspects of the interactions between learning and evolution. Language is a complex behavior that is, at least in some cases, used for altruistic purposes (of course, sometimes selfish motives like intimidation, manipulation, and encryption can also play a role). The *population* as a whole benefits from the altruistic use of language, as it does from other altruistic behaviors. In particular, the population benefits from using syntactic language (Pinker & Bloom 1990), but it is not trivial to explain how an *individual* that uses syntax can be successful in a nonsyntactic population.

By using a methodology of computational modeling that avoids the underspecification of Rachlin's arguments, researchers in this field have shed some new light on how this behavior has emerged (Hurford 2002; Steels 1999). For example, these models have shown that when individuals learn language from each other with rather generic learning mechanisms, a rudimentary syntax can emerge without any genetic change (Batali 1998; Kirby 2000). The learning algorithms, for example, the recurrent neural network model in Batali (1998), provide – although far from finally – a fully *specified* candidate answer to the first question we posed previously.

Similarily, in recent work I have explored some provisional answers to the second and third questions. In Zuidema (2003, forthcoming) I explore the consequences of the fact that language itself can, in the process of learners learning from learners, adapt to be more learnable (Kirby 2000). As it turns out, this *cultural* process facilitates the *evolutionary process.* Evolutionary optimization becomes possible, because the cultural learning process fulfills the preconditions for a coherent language in the population. Moreover, the model also shows that much less of the "knowledge of language" needs to be innately specified than is sometimes assumed. Cultural learning thus lifts some of the burden of genetic evolution to explain characteristics of language. Alternatively, Zuidema and Hogeweg (2000) present results of a spatial model of language evolution. These results show that syntax can be selected for through a combined effect of kin selection and group selection.

These answers are far from final, but I believe that such well-defined models present an appealing perspective on *how* cultural learning can lead to the successful acquisition and creation of a complex, altruistic behavior like syntactic language, and *why* the learning mechanisms operate the way they do.

ACKNOWLEDGMENT

The work was funded by a Concerted Research Action fund (COA) of the Flemish Government and the Vrije Universiteit Brussel.

Author's Response

Altruism is a form of self-control

Howard Rachlin

Psychology Department, State University of New York, Stony Brook, NY 11794-2500. **howard.rachlin@sunysb.edu**

Abstract: Some commentators have argued that all particular altruistic acts are directly caused by or reinforced by an internal emotional state. Others argue that rewards obtained by one person might reinforce another person's altruistic act. Yet others argue that all altruistic acts are reinforced by social reciprocation. There are logical and empirical problems with all of these conceptions. The best explanation of altruistic acts is that – though they are themselves not reinforced (either immediately, or delayed, or conditionally, or internally) – they are, like self-controlled acts, part of a pattern of overt behavior that is either extrinsically reinforced or intrinsically reinforcing.

The commentaries demonstrate the enormous variety of approaches that may be taken to explain altruism. Though these approaches each afford a different perspective on the target article, I have attempted to classify them under a few general and overlapping headings. I will discuss each heading in turn, referring to specific commentaries as I go. Although all of the commentaries are thoughtful and deserve thorough discussion, it is not possible in this limited space to answer each commentator in detail. Instead, I have tried to highlight crucial points and respond to common criticisms.

R1. Teleological behaviorism, cognition, and neuroscience

Gray & Braver draw implications from the behavioral correspondence of self-control and altruism for both cognition and neuroscience. Their suggested empirical tests are certainly important and worth doing. But I do not believe that you can crucially test a behavioral model, or even a purely cognitive model, with neurophysiological measurements.

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

The inputs and outputs of a cognitive system are the manipulations and observations of cognitive experiments. If a cognitive model can predict future behavior and explain current behavior, the model is considered successful. However, an important issue within cognitive neuroscience is the extent to which a successful cognitive model says anything about how the underlying neural hardware is organized.

In other words, is the purpose of cognitive psychology only to predict behavior, or is it also to provide a structure that may be filled in by neurophysiological observation? It is certainly possible that a single cognitive model could serve both purposes (behavior prediction and physiological instantiation), but there is no inherent reason why it should.

A behavioral model makes the same kinds of observations as the cognitive model does but explains them in terms of contingencies in the environment between signals and consequences, or behavior and consequences, or between signals (discriminative stimuli) and behavior-consequence contingencies. Just as cognitive psychology is not obligated to specify internal neural mechanisms if it predicts and explains behavior, so too, teleological behaviorism is not obligated to specify internal cognitive mechanisms, if it predicts and explains behavior. A cognitive or physiological explanation for why I am writing these words now, for example, would take into account the state of my internal cognitive or physiological mechanisms as they were affected by reading, understanding (or not understanding) the commentaries, storing the information therein, processing it, and so forth, and then producing these written words. A teleological behavioral explanation would first analyze my habits over a long time-period to establish my goals (to influence the thinking of people outside, as well as within my field of research, for example), and then show how my current behavior fits into that pattern. The goals thus observed may or may not fit into my wider goals. If they do, those wider goals may be seen as reinforcing the narrower ones. If my habits do not fit into still wider habits, then they are valuable in themselves – intrinsically valuable.

Hartung believes that teleological behaviorism necessarily implies that there are "intrinsic reward receptors" in our brains, but, as I said, teleological behaviorism implies no particular physiological model. It would seem unlikely, moreover, that any single brain center could account for reward as such. We know from a series of experiments by Premack (1971) that a single act (wheel-running by rats, for example) may be either a reinforcer (of lever pressing) or a punisher (of eating), or itself be reinforced or punished, depending on contingencies. You would have to say that the very same response (wheel-running) stimulated the reward center when it was contingent on lever pressing, but stimulated the punishment center when it was contingent on eating.

Krebs's cognitive strategy for operant conditioning: to "repeat acts that were followed by a reward and suppress acts that were followed by a punishment," would have to be modified to take Premack's (1971) findings into account. Perhaps: "increase the rate of acts followed by a higher valued act . . . ," and so on, might be better. But Krebs's so-called strategy is an English sentence. How this sentence might get translated into an internal mechanism that explains a person's choices between rewards involving self-control – for example, between abstract patterns of acts, like social cooperation, and particular acts, like drinking another glass of scotch – Krebs does not say. What Krebs is doing with his "strategy" is taking a set of behavioral observations, characterizing it as best he can, and then placing his own description into the behaving organism's head. This kind of internalization of behavioral observation has never worked. I believe that when all is said and done, the most useful general characterization we have of our behavioral observations is: "behave so as to maximize utility." This "strategy," however, is not encoded within the observed organism but rather, is a *method* used by the observer – a successful method for making predictions and explaining behavior.

Teleological behaviorism puts a person's mental and emotional life strictly into the hands of the observer of the person's overt behavior. **Lacey** does not believe that you can give a behavioral account of a term such as "love." I do not believe that any other account is possible. If a man who has abused his wife and children throughout his life declares on his deathbed that he really loves them, then the state of his current cognitive and physiological systems may well correspond to some cognitive/physiological definition of love. However, his wife and children will not believe him. He may not be lying, but he is certainly wrong. He does not love them (according to teleological behaviorism) – regardless of the current state of his internal mechanisms, regardless of his intrinsically "private reasons" (**Khalil**) or "psychological costs and benefits" (**Perugini**) or "mentalistic aspects" (**Sedikides & Gregg**). Love, according to teleological behaviorism, is a classification of behavioral patterns made by observers (i.e., society) for their own benefit, rather than an internal state or condition that might be measured by an MRI machine or heart rate monitor or reported by introspection. It is useful for us to divide other people into those who love us and those who do not. And this division depends on the patterns we observe in our interactions with them – verbal, as well as nonverbal. According to teleological behaviorism, verbal reports of cognitive or emotional states, such as saying "I love you," are truthful to the extent that they describe and predict the speaker's own past and future verbal and non-verbal behavior; conformance of a verbal report to the state of an internal mechanism or a physiological or hormonal state would be suggestive but not strictly relevant to the behavioral model. According to the teleological behaviorist, "memory," "perception," "decision" – like "love" – are fundamentally names for behavioral patterns, not the internal mechanisms underlying these patterns. (Thus, contrary to **Hinde**'s assumption, the use of such terms by a behaviorist is not necessarily a "lapse," and, contrary to **Sedikides & Gregg,** it need not send shivers down a behaviorist's spine.)

This is not to say that the organism is empty. There are internal mechanisms behind our behavior (just as valves, pistons, and spark plugs are internal mechanisms in a car's engine). But, for a teleological behaviorist, our thoughts and emotions themselves (like a car's acceleration) *are* our behavior patterns regardless of the mechanisms behind them. Thoughts and feelings are not inherently private events going on inside the skin. There are lots of private events, which are efficient *causes* of thoughts and feelings, but thoughts and feelings themselves are not events of this kind. Likewise, altruism and selfishness are not private events but judgments made by an outsider – judgments of whether or not a given act falls into a given pattern. In other words, they are not a fixed property of an act, still less a cognitive or physiological state.

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

Broude believes that a behaviorist must see a cost-benefit analysis as "morally repugnant." But my analysis of self-control and altruism *is* fundamentally a cost-benefit analysis – or at least an economic analysis. However, I determine the costs and benefits from the person's overt choices (what economists call "revealed preference"), rather than by my own introspection or by a phone call to the biology department. The issue between Broude and myself, therefore, is not whether a cost-benefit analysis is valid. The issue is who is doing the analysis. For Broude and other cognitive theorists it is the behaving organism; for me it is the observer.

Stukas, Platow & Foddy highlight the difference between behavioral and cognitive approaches regarding the use of mental terms. They believe that such terms can only stand for internal mechanisms; a behaviorist who uses them must therefore be retreating from behaviorism. For these commentators, even social constructs have no place in a psychological theory unless they are "internalized." They agree that self-control and altruism may be related – but only in terms of "flexible cognitive [i.e., internal] representations of self." **Carlo & Bevins** and **Perugini** go even further. They seem to feel that only by postulating numerous emotions (internal empathy, sympathy, personal distress, uncomfortable feelings, and many others), which are vaguely connected in an internal system and observable only by introspection, can you generate testable hypotheses about behavior. For them, you know you're in love by just knowing it. For some, you know you're in love when you feel your heart go pitter-patter. For me, you know you're in love when you're actually kissing the girl, buying her flowers, being kind to her, and overtly behaving in a certain pattern; in other words, you know you're in love in the very same way that she knows you're in love.

I agree with **Lacey** that no explanation of altruism in the laboratory, however complete, will capture all there is to say about love, compassion, and justice in the real world. But still, a teleological behavioristic approach to these topics – based on overt patterns of behavior and their social consequences – maps better onto real-world behavior than an approach based on introspection or the action of cognitive or physiological mechanisms. I disagree with Lacey and with **Zentall** that I am obligated to precisely specify the real-world pattern into which a given bit of altruistic behavior fits – any more than the physicist has to specify the pattern of a leaf as it falls from a tree. It suffices to show that the pattern conforms to conditions 1, 2, and 3 of my definition – which Lacey states much more elegantly than I could do. For a teleological behaviorist the focus of all mental terms – "raw feels" as well as intentional terms – is in our overt behavior over time (verbal and non-verbal), not somewhere in our heads. This goes for the scientist's and the philosopher's perceptions, as well as those of subjects in a psychology experiment.

Baron claims that both self-control and altruism arise from illusions. **Krueger & Acevedo** attribute self-control and altruism to miscalculation. **Weirich** speculates that they may be attributed to "limits on powers of discrimination." There is no doubt that illusions do occur (for a behaviorist these would appear as inconsistencies among behavioral patterns or between verbal behavior and non-verbal behavior). The reason I used the woman running into the burning building as an example of an altruistic act is because it would seem crystal clear that any extrinsic reinforcers of this particular act would be dwarfed by its costs. Yet people do perform such acts, do not in general claim to regret performing them and, more importantly, sometimes repeat them. How can you be happy and satisfied with the outcome of a mistake or an illusion? If the woman does not stop to calculate the costs and benefits of this particular act, it is not the same as miscalculating, being under an illusion, or failing to discriminate. She is not fooling herself. She is behaving so as to maximize utility in the long run – exactly as she should.

Krueger & Acevedo and **Levine** claim that only in extreme examples like this one, or in the even more extreme example of a woman throwing herself on a hand grenade, will we find altruism without any possibility of reciprocation. However, consider the members of my audiences (about half) who anonymously chose Y in the prisoner's dilemma game diagramed in Figure 1. Those people would have lost $200 if the money were real, regardless of the choices of the other participants. There was no conceivable "expectation that others will reciprocate" (Krueger & Acevedo's cognitive mechanism of projection) since no one knew who the Y-choosers were. Why choose Y? Because, as opposed to choosing X, this act was part of a valuable pattern (perhaps consisting of voting, tipping, not littering, and so forth) and not an isolated case. (I would bet that if the money were real, even more people would choose Y; the real money would put the game even more firmly into the context of real-life decisions.) Altruistic behavior may initially be a mistake or illusion; my act may turn out to help my community when I thought it would help only me. But why repeat such acts if their pattern is not valuable in itself? Does a person who helps to build a house for a homeless family really believe that they will build a house for him in return? And if they do not reciprocate, does he regret helping them? I doubt whether there are many people – even atheists – who, on their deathbeds, express regret that they have been too altruistic through the course of their lives; or who believe that, by being altruistic, they made a mistake, or miscalculated, or succumbed to an illusion.

Weirich suggests that if an individual act fits into a highly valued pattern, we ought to treat this very fact as a separate reinforcer along with other reinforcers of the act – putting altruism and self-control on the same plane as selfishness and impulsiveness. This, I believe, would be a category mistake (Ryle 1949). Before Gestalt psychology came on the scene, psychologists of the Wurzburg school claimed that a sensation – a tone, for instance – had a quality called a *form quality* that stood beside the loudness and pitch of the tone as a property of the tone itself derived from the other tones around it. This was generally recognized to be a category mistake. Form quality is a quality of the pattern of the tones, not of any individual tone. Similarly, the value of self-control or altruism is a quality of a pattern of acts, not of individual acts.

Margolis claims that there are only two ways to explain altruism: (1) the notion of altruism as an illusion or mistake, which he believes that I must believe but which I do not believe; and (2) the notion of a "hard core" altruism that cannot in any way be reduced to selfishness, which he believes. The choice of Y by people playing the prisoner's dilemma game diagramed in Figure 1 would be an example of hard-core altruism by Margolis's criterion – no reinforcement, no way, no how. Margolis says that altruistic behavior is a product of an inherited mechanism, NSNX, that balances two rules: one that says, don't be too selfish (NS), the other that

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

says, don't be exploited (NX). (This is a typical cognitive theory of altruism. One has only to read the commentaries to see that it is just one of dozens. By Margolis's own standards, this very fact would disqualify it as a valid theory.) I believe that the people who chose Y did so because the present act was part of a highly valuable pattern of acts. I don't see why Margolis and I can't both be right. I like my theory better because it points to correspondences between altruism and self-control, and can be tested in various ways, whereas Margolis, in trying to discern the workings of a hypothetical and vaguely specified internal mechanism on the basis of its inputs and outputs, is very much like someone trying to figure out the plumbing of a submarine by walking around outside with a thermometer. Margolis, like **Broude,** just cannot see a wholly behavioral theory as a theory at all. For them, the satisfaction of postulating an immediate internal cause for each individual act (in the form of an internal mechanism) makes up for deficiencies in prediction, control, and meaningful explanation. Such satisfactions are inevitably ephemeral.

Wedekind, on the other hand, believes that all particular acts that may seem altruistic can be individually traced to "net benefits in the long term." But I never claim that human altruism lies in ignorance of long-term costs and benefits of particular acts. The difference between **Schuster**'s and Wedekind's views, and my own, is subtle but important. They think that all social reinforcement is extrinsic to action, whereas I think that patterns of social activity may be intrinsically rewarding – that is, done for their own sake. Their view directs behavioral research towards a search for individual reinforcers for each social act, including altruistic acts, whereas mine directs behavioral research towards a search for patterns of actions into which the present act fits.

Similarly, **Read** believes that "individual [altruistic] acts in a pattern *always* pay off more than acts that are not part of the pattern." Schuster would probably agree but add that the payoff is always social. It is true that in the primrose path experiment I described, individual self-controlled acts do eventually pay off (although the discounted value of the payoff may be less than the present cost of the act). However, in the prisoner's dilemma game of Figure 1, individual choices were made anonymously, and individual choices of Y never paid off – either socially (in terms of reputation) or non-socially. And, in real life self-control situations, individual isolated choices to reject an immediate reinforcer (having 39 instead of 40 cigarettes today, for example) may also never pay off in terms of increased health or social acceptance, for all the present anguish they cause.

As **Stukas et al.** and **Schuster** say (and **Wedekind** implies), social reinforcement (obtained through being a frequent cooperator, for example) maintains social cooperation; it also maintains self-control. Relative addiction theory (Rachlin 1997) relies heavily on social reinforcement as an economic substitute for addiction. Nevertheless, contrary to Schuster's assertion, even nonhuman animals can learn to cooperate in prisoner's dilemma games versus a computer – that is, without social reinforcement (Baker & Rachlin 2002). Moreover, where social reinforcement in everyday human life may sometimes be strong and immediate, it is often only loosely correlated with individual acts (as Wedekind notes). Members of happily married couples, for example, do not generally reinforce one another's every specific act. It may well be, as Schuster says, that whatever brain mechanism gives patterns of social interaction their extremely high value, differs from the mechanism that gives good health its high value (just as brain mechanisms for food and water reinforcement may differ). Still, in behavioral terms, social and non-social patterns of reinforcement act in the same way – they reinforce their own components.

Buck argues that complex behavioral patterning depends on linguistic structure. It is certainly true that language is an effective discriminative stimulus for complex behavioral patterning. It is also true, as Buck asserts, that rational behavior is tempered by emotions. But neither language nor emotion acts like a "ghost in the machine." Neither is a gratuitous expression of an internal state thrown out onto the world for no reason. Neither occurs in a vacuum. Both language and emotion have functions in the world. Choice behavior may be emotional, but if so, then it is for a purpose. For, emotions, like all mental states, are fundamentally patterns of choices over time. If, in a social situation, I tell you I am feeling happy at this moment, I am telling you about my past and future behavior, not about my inner state. If I tell you I am feeling happy but act sad, I am not lying or mistaken about my internal state; I am lying or mistaken about my past and future behavior. In the face of such a lie or mistake, the behaviorist asks not what went wrong in the internal connection between emotion and language, but what reinforced the verbal behavior that was emitted – not *how* the verbal behavior was emitted but *why* it was emitted. I agree with Buck, however, that both questions need to be pursued.

In a more cognitive version of **Buck**'s argument, **Fantino & Stolarz-Fantino** see altruism as a case of rule-governed behavior. I agree with this characterization; but (contrary to **Weirich**) a behavioral rule is a regularity observed in overt behavior, not an internal edict that the actor is trying, perhaps failing, to obey. Let readers ask themselves which rule truly governs a person's behavior in cases where she says she's obeying one rule but her behavior actually conforms to another. Discrepancies between verbal behavior and cognitions, like the discrepancies between verbal behavior and emotions discussed in the previous paragraph, are best explained not in terms of bad connections in an internal system, but in terms of reinforcement for the verbal behavior. The usual reinforcement for verbal descriptions of our own cognitions and emotions (as behavioral patterns) lies in the coordination of these patterns with those of other people. But sometimes our verbal behavior itself is reinforced, independently of its accuracy, and that is when mistakes and lies occur. In other words, these instances too are problems of self-control.

If **Zuidema** is correct that syntax may emerge from an exactly specified internal learning mechanism as individuals in a population learn language from each other, this is a fascinating fact. It does not follow from this fact, however, that altruism cannot be understood without an exact specification of the internal mechanism underlying it ("how individuals represent and acquire this knowledge"). I certainly did not mean to imply that in order to behave altruistically a person has to choose "life-long altruism" over "life-long selfishness" at some particular point of his or her life. As section 9 of the target article indicates, such a choice is like the choice many of us make to lead healthy lives. We find ourselves behaving in a certain pattern, perhaps through extrinsic reinforcement of particular acts, or through imitation, or because this rule bears certain simi-

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

larities to previously reinforced rules. Once we behave that way, the pattern is maintained by its own high value. This does not mean that we never slip or even that we can state what the pattern is. People who perform highly altruistic acts often cannot give good reasons for them (as opposed to those who act selfishly who often can give reasons). This irony has often been noticed by novelists, from Dickens (see Khalil 2001) to Highsmith. I doubt whether the people who choose Y in the prisoner's dilemma game of Figure 1 could explain their choice as well as those who chose X could.

Sobel takes me to task for not defining self-control carefully. His counterexample to my definition is: touching your nose for 30 seconds (short activity) versus listening to a symphony (long activity). According to my definition, not touching your nose would constitute self-control only if you could not touch your nose while listening to the symphony (you can't have your cake and eat it), and if touching your nose for 30 seconds were actually preferred to 30 seconds of listening to the symphony. The latter condition might apply if, say, your nose itched. Imagine then that your nose itched so much that only a 30-second scratch could alleviate the itch, but if you touched your nose, someone would turn off the symphony you were listening to. Under these conditions not touching your nose would constitute self-control. The problem is not with my definition but with Sobel's not taking it seriously.

Wagstaff disputes my interpretation of Aristotle as linking altruism and self-control. Of course, there have been many interpretations of Aristotle, some diametrically opposed to others. For Wagstaff, self-control and altruism are separate inner forces (dispositions) that might or might not be expressed in overt behavior; he sees reinforcement as immediate (and internal) pleasure. My own interpretation of Aristotle's view of altruism comes from Apostle's (1984) translation of the *Nicomachean Ethics.* Apostle does not translate any particular virtue analyzed by Aristotle as "altruism." But this does not mean that Aristotle did not consider altruism to be a virtue. According to the OED, "altruism" was coined by Comte in the nineteenth century and did not appear in English until the second half of that century. The Aristotelian virtues, "generosity," "high-mindedness," and "bravery" consist of overt habits (*hexes*), not internal dispositions. These habits would, it seems to me, include such acts as choosing Y in my prisoner's dilemma game and a woman's running into a burning building to save someone else's child – that is, altruistic acts.

R2. Reinforcement

I agree with **Zentall** that choice of a smaller, earlier reinforcer over a larger, later one is not always, or even not usually, irrational. No rational person would prefer $10 next year to $9 right now. As **Weirich** points out, the issue of rationality and its relation to self-control and altruism is complex. The only thing I have to say about it here is that, for a behaviorist, labeling an act "rational" or "irrational" is not to label the internal mechanism that generated the act, but to classify the pattern into which the act fits. The criteria for that classification would, in turn, depend on the purposes of the labeler rather than the actor.

Alexander and **Zentall** do not see how you could measure a preference for pattern A over pattern B and also measure a preference for a component of pattern B over a component of pattern A. This is how: Concurrent schedules of reinforcement provide a graded measure of choice among individual acts; concurrent-chain schedules provide a graded measure of choice among patterns of acts. These schedules have been developed to measure relative value. For example, from concurrent-chain tests, we know that a hungry rat strongly prefers the pattern: *lever-pressing plus eating* to the pattern: *not lever-pressing plus not eating* when these patterns as wholes are the objects of choice. At the same time, from simple concurrent choice tests, with only pressing or not pressing the lever as objects of choice, we know that the rat prefers a component of the dispreferred pattern (*not lever-pressing*) to its alternative (*lever-pressing*). Pressing the lever satisfies conditions 1 and 2 in my definition of self-control. Thus, when a rat presses a lever for food (despite preferring not pressing to pressing as such) it is in a primitive sense exhibiting self-control.

Contrary to **Hinde**'s assertion, there is nothing fuzzy about these measurements; we can make them very precisely in the laboratory on a narrow scale. Similarly, in the laboratory, we can test the effect of patterned choices in prisoner's dilemma situations with human and nonhuman subjects (Rachlin 1995b). The real human world is, of course, infinitely more complex than a Skinner box. A person has only one life. We cannot ask a person to choose between forced commitments to alcoholism, teetotalism, or social drinking for the rest of his life, and then make predictions about his behavior in a second life. However, we can and do make predictions on a smaller scale. We see patterns in a person's behavior and make predictions based on those patterns. The woman's act of running into a burning building is not unique in her life. (If it were unique its apparent altruism would be accidental.) Her act fits into one of several behavioral patterns (including verbal behavior) in her life; this is its *only* meaning. I agree with **Krueger & Acevedo,** who pose more complex scenarios and feel that, because real life is complex, "predicting individual acts [is] a near-hopeless enterprise." This is true even in the laboratory. But we don't have to predict individual acts (each lever press, for instance) to make predictions about patterns of acts (response rates). The fact that we cannot actually establish a giant concurrent-chain schedule and measure a person's preferences before predicting his or her behavior does not mean that we cannot or should not make predictions in the real world. As I said, the fact that a physicist may not be able to predict the path a leaf will take when it falls from a tree doesn't mean that physics has no application in the real world.

On the other hand, explanations of altruism in terms of an internal "self-system" (**Lewis**) or a "process of self-extension" (**Sedikides & Gregg**) that **Hinde** feels are generally hard-headed and only fuzzy at the borders, depend fundamentally on introspection. In stating the three "universal concepts" of his self-system, Baumeister (2000, p. 9) places "*reflexive consciousness,* through which knowledge of self develops through awareness" in the number-one spot. The most fundamental concept of self-system theory is thus removed from objective measurement. This theory is not just fuzzy around its edges but at its very core. The purpose of the concept of reinforcement is to serve as a method of prediction and explanation, not to point to an internal mechanism. Teleological behaviorism is not a failure because it does not develop hypotheses about the existence

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

and operation of one or another internal mechanism – cognitive (software) or physiological (hardware). It makes as little sense to expect to reduce behavioral theories to physiology, as it does to reduce economics to physiology.

Wilson & Miller say that a difference between self-control and altruism is that "self-control is largely a temporal problem," whereas "altruism need not be extended in time, although it can be." This is a crucial issue. The main point of the target article is that altruism is *always* extended in time. In apparently "one-shot" prisoner's dilemma games such as the one I play with my lecture audiences, the people who chose Y (the "altruistic" ones) did so, I claim, precisely because this act was part of a sequence of acts extended in time both before and after my lecture. Those who chose X (the "selfish" ones), on the other hand, did so because for them this act was isolated from real-life choices (not necessarily because they are generally selfish). The people who chose X were indeed playing a one-shot prisoner's dilemma game; those who chose Y were playing an iterated game even though, at the moment, all choosers were playing the very same game. Wilson & Miller would claim that all players were playing a one-shot game. Therefore, to explain the players' behavior, Wilson & Miller must find an immediate or delayed or conditional reinforcer for Y-choices in this very game to offset the extra (hypothetical) $200 Y-choosers would have gained had they chosen X. In Wilson & Miller's view, this reinforcer is "the welfare of others" acting on each Y-chooser's own reinforcement system. The people who chose Y are experiencing "the same kind of reinforcement" as if they had chosen X. They get this reinforcement by vicarious satisfaction – from what? Choosing Y increased N by 1, thereby giving those who chose X $100 more than they would have gotten otherwise. By virtue of reinforcers "of the same kind" they valued a hypothetical $100 extra for each of the other players, including those who chose X, more than their own hypothetical $200 loss. This unlikely implication is the result of a relentless search for a reinforcer for each individual altruistic act. (As I said previously, I would bet that the use of real money would produce more, not fewer Y-choices, because it would set the game more firmly in real life.)

Several other commentators (**Ainslie & Haslam, Fantino & Stolarz-Fantino, Grace, McLean & Bragason, Kaplan & De Young**), some of whom are behaviorists, attempt to account for altruistic behavior in the same way as **Wilson & Miller** do – by imagining that Person A's act may be reinforced by Person B's consummatory behavior. That is, my act of buying the apple may be reinforced by your act of eating the apple. Other less behaviorally oriented commentators (**Buck, Carlo & Bevins, Gintis, Hinde, Lewis, Perugini**) treat empathy as an emotional mediator between altruistic behavior and its reinforcement. First you perform an altruistic act, then another person is rewarded, and then you feel good due to empathy; this good feeling self-reinforces the particular altruistic act you just did – just as if you were rewarded yourself. The most fundamental problem with these explanations is that the other person's reward need have nothing to do with your act. Person A puts a dollar into the Coke machine and gets a Coke. We can then say that drinking the Coke reinforced inserting the dollar. But if Person A puts the dollar in the machine and Person B (a stranger) gets the Coke (and neither shares it with A nor thanks him but simply walks away with his soda), A is not likely to put more dollars into the machine – not even if B got two Cokes or ten Cokes for that matter. Altruistic acts, as individual acts, are never reinforced; otherwise they wouldn't be altruistic.

Self-control and altruism are qualities of patterns of acts, not of individual acts. **Ainslie & Haslam** call this an unnecessary assumption. But it is a necessary assumption if you want to explain altruistic behavior in behavioral terms. Otherwise you end up, as they do, by postulating some vague process of internal self-reinforcement ("the primary rewardingness of vicarious experience") and the existence of internal bargaining processes. I am an admirer of Ainslie's brilliant insights into the nature of self-control. But I cannot follow him in his abandonment of behaviorism and his undefined conception of reinforcement. The issue between Ainslie and me comes down to the nature of the conflict inherent in self-control. For Ainslie, it's a conflict between now and later. For me, it's a conflict between wide and narrow temporal extent. The former conception forces behaviorists into a bad exchange. To retain the idea that each particular act must have its own particular reinforcer (immediate, delayed, or conditional), they postulate the existence of inner responses, inner stimuli, inner reinforcement, or inner bargaining between past present and future self-representations. Like **Hinde,** Ainslie & Haslam create a mirror image of environmental contingencies inside the organism. For this they give up the great advantage of behavioral analysis: the ability to observe, measure, predict, and control.

R3. Emotions

Teleological behaviorism sees emotions such as empathy, shame, guilt, and regret as themselves patterns of overt behavior. But let us assume, for the moment, that they are internal states. Can emotions, as internal states, cause overt behavior? There are two ways in which they conceivably might do so. One is, directly. You have the emotion and the behavior comes out like steam escaping from a pot. This is what **Gintis** means when he refers to punishing a defector at a cost to oneself as "venting anger." But then we would have to ask why anger should be vented in the particular (and costly) direction of punishing the defector, rather than by screaming or pounding on the table or any of a multitude of less costly directions. Explaining altruism as "venting anger" or venting empathy, for that matter, is just another way of saying, "I don't know why they do it." We still need some explanation as to why the behavior takes the form it does.

The previous section discussed problems with the concept of empathy as a *positive* emotion that could reinforce particular altruistic acts. The concept of reduction of *negative* emotions such as guilt and shame (conceived as internal states) as immediate reinforcers of particular altruistic acts (**Gintis**) is the other side of the emotional coin. Here, altruism is conceived as an avoidance response; it avoids shame and guilt. None of the commentators specified how this avoidance mechanism is supposed to act, but, in the study of avoidance, there have been serious problems with emotional states as explanations. To take just one line of evidence, Rescorla and Solomon (1967) showed long ago that the latency of an overt act of avoidance was generally shorter than the latency of the negative emotional response. At best, both the overt avoidance response and the internal

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

emotional response occurred simultaneously. Thus, an overt avoidance response (the altruistic act in this case) cannot be reinforced or punished by reduction of a covert emotional response (empathy, guilt, or shame). The emotion – the supposed motive – does not even *appear* until after the act. The event most immediately following the overt act is an increase rather than a decrease of the aversive emotion.

A one-factor explanation of altruism as avoidance of shame, guilt, or regret would not rely on the reduction of an internal aversive emotion after each altruistic act. You could just say, as **Fantino & Stolarz-Fantino** do, that altruistic acts avoid an unpleasant emotion: regret. This one-factor explanation relies on a negative correlation over time between altruistic acts and regret; it provides no reinforcer for individual acts of altruism. Thus, it relies on the response patterns and teleology that Fantino & Stolarz-Fantino want to do without. Further, it has the disadvantage of using the very vaguely defined concepts of regret, guilt, and shame, as explanations of altruistic behavior; it takes these emotions as given, as something we just know. A teleological behavioral approach, on the other hand, would treat regret itself as an overt behavioral pattern to be explained: What reinforces regret? What is the advantage, in our social system, of regret as overt behavior? Once these questions are put, regret may be studied and understood rather than simply postulated on the basis of introspection.

R4. Context

Broude believes that self-control cannot be context-specific, and that, therefore, anyone who says that altruism is a form of self-control must believe that altruism cannot be context-specific. I don't know where Broude got this idea. Of course self-control, hence altruism, is context-specific. Not every heroin addict is also a cocaine addict, a gambler, a cigarette smoker, and an alcoholic. A person cannot just learn "any old thing" about self-control or anything else. Self-control, like altruism, is specific to its context. **Zizzo** cites several excellent and penetrating demonstrations of the effect of context in social choice: "inequality aversion," "reciprocity," "trust responsiveness," "pure and impure altruism," "perceived fairness." He concludes that "the interpretation of cooperation in the finitely repeated PD is likely to be difficult." This is certainly true. But if each such demonstration is thought to uncover a different internal process, one that might be opposed by another internal process, measurement by revealed preference techniques would be impossible. These processes need to be defined in behavioral terms so that they can be manipulated and measured. I am justly criticized for writing the target article in the absence of such behavioral analysis. But I do not believe that our understanding of altruism in general, or in the prisoner's dilemma in particular, is furthered by simply adding terms representing internal events to a utility function. I am no expert on cognitive mechanisms, but if each instance of context-specificity demanded a separate mechanism, we would need one altruism mechanism for sharing toys and another for sharing food; and why not one for each kind of food? This is the same barren path that led need-reduction reinforcement theorists to posit a different need for each individual instance of reinforcement (see **Sedikides & Gregg** for "need to belong").

Alexander believes that it makes no sense for a given act to be seen as altruistic or not on the basis of the context of the act. **Van der Steen** makes a similar point. This criticism goes to the heart of what "altruism" means. Whether an act is altruistic or not is a social judgment and therefore depends on the purposes of the group doing the judging. If "altruism" stood for a fixed internal process, a given act would be altruistic or not on the basis of the process by which it was generated. But if (as **Grim** and I claim) the meaning of an act depends on the pattern in which it is embedded, a given act could be part of more than one pattern – just as a given note could be part of more than one melody in a symphony. A person's charitable gift, for example, may be tax exempt. This does not necessarily prevent it from being altruistic. From the point of view of the nation making laws that reinforce such gifts, the act would be one among a series of acts reinforced by the tax system. But from a more local point of view, the act may be one among a series of acts, many of which are not tax exempt; therefore it would be altruistic. **Grim** sees reinforcement (or function) as selecting from among several overlapping patterns; according to Grim, this would "lead the individual to conceive of what is happening in certain terms." I agree, except I see the person's own conception and his behavior, classified by the observer, as one and the same thing. There would be no point in asking the person for his "private reasons" for including the act in one pattern or the other, as **Khalil** suggests, because the pattern into which we classify the act depends on our purposes, not his.

I agree with **Danielson** that reciprocity of cooperation is extremely important. Baker and I (Baker & Rachlin 2001) found that cooperation in the prisoner's dilemma varied monotonically with probability of reciprocation. Yet people still perform altruistic acts with no possibility of reciprocation. To take a more mundane example than the woman rushing into the burning building, and a more realistic example than cooperating in a prisoner's dilemma game: people at least occasionally walk for blocks through filthy streets holding a candy wrapper to deposit in the wastebasket. Another example is voting: in what way is voting in a national election individually reciprocated? To understand these acts we have to think beyond the idea of reciprocation as an individual's reward for a particular act of cooperation. For instance, even in prisoner's dilemma games some people *consistently* cooperate without reciprocation. They do so, I believe, because – contrary to the experimenter's instructions – they treat these games in the laboratory as part of the patterns of their lives rather than as isolated events.

R5. Nature versus nurture

Sobel says that I must provide a reason why natural selection favors individuals who have a preference for maintaining altruistic patterns. I disagree (although I do explain how group selection might work to favor such individuals). Economists see no need to provide an evolutionary story behind every utility function. Why should psychologists have such an obligation?

I agree with **Reed** that biological explanation is orthogonal to behavioral explanation; and in the target article, I tried my best to avoid the nature-versus-nurture argument. I evidently did not try hard enough. Of course, all behavior is explicable in both terms. If I show that a given behavior may be learned, you can always argue that the ability to

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

learn it is genetic; if you show that a certain behavioral trait is genetically determined I can always show that the expression of that trait depends on the environment. Whether the properties of a ripe red apple in my left hand are genetically or environmentally determined depends on what I am holding in my right hand. If it's an orange, then the apple's properties will seem genetic; if it's another apple in my right hand, say, a green, unripe one from the shady side of the tree, then the apple's properties will seem environmentally determined. To say that altruism may be learned is not to say that "there is no genetic variance in propensity toward altruism" (**Hartung, Zizzo**), or "to eliminate the potential role of evolution" (**Kaplan & De Young),** or to be "averse to considering the genetic basis of behavior" (**Zentall**), any more than the fact that calculus may be learned means that there is no genetic variance in the ability to learn calculus or that the role of evolution in our ability to learn calculus has been eliminated. The fact that, in free choice tests, a hungry rat spends more time in eating than in wheel-running, and more time in wheel-running than in lever-pressing, has two kinds of implications: one, the behavioral implication that certain contingent relations among these three activities will be found (e.g., wheel-running will reinforce lever-pressing but punish eating); another, that some internal mechanism or mechanisms that are partly innate, partly themselves learned, underlie the behavior. To say that one implication is important does not deny that the other is also important.

Krebs seems to believe that I was trying to say something about how altruism evolved. But I did not even mean to say that self-control came before altruism in the course of human evolution; the reverse might well be true. A behavioristic theory would be silent on how altruism was selected. I did call the self-control mechanism an "innate learning mechanism" but Premack's (1969) reinforcement theory would not distinguish such a mechanism from an evolved strategy. All I intended to claim with respect to a common mechanism is that in both cases, self-control and altruism, evolution must select behavioral patterns rather than individual acts. Individual acts would be reinforced to the extent that they formed part of a valuable pattern. Whatever mechanism did this would be a "learning mechanism."

Wilson & Miller accuse me of setting up a straw man in claiming that biologists posit a specialized altruism mechanism. Then they proceed to make virtually the same claim, only now it's a "specialized form of learning." I had presumed that when Tooby and Cosmides (1992) compared the supposed specialized mechanism to the eye they were minimizing the contribution of learning. If I was wrong, I apologize. If there were a specialized form of learning, as Wilson & Miller claim, the question becomes, "What does that specialized mechanism do?" If, as I claim, it organizes low-valued particular acts into high-valued patterns, and if we had one such mechanism for self-control and one for altruism, then we would have two mechanisms doing the very same thing. It seems to me that in the absence of physiological evidence for two such redundant mechanisms we ought to assume that only one exists.

R6. Morality

Being from the Bronx, I certainly do not belong among **Hartung**'s five known pure altruists. That leaves only four remaining. Still, we should try to explain their behavior because, on a lower level, altruism is a pattern in all of our lives. The players of the prisoner's dilemma game illustrated in Figure 1 who anonymously chose Y are good examples. When I was a Boy Scout I occasionally helped old ladies across the street (I still do, although now I'm less disinterested). I did not do so in fear of hell or hope of heaven, as Hartung and **Levine** would seem to require. Despite the rhetoric, fear of hell and hope of heaven are not by themselves good explanations of altruistic behavior. We still would need to explain, in behavioral terms, how such fears and hopes work. As I said in the target article, there is a distinction to be made between altruism and morality. The behavior of the firemen and the hijackers on September 11th may have been equally altruistic but not, from our viewpoint, equally moral. **Wagstaff** cites cases where altruistic acts turn out to be socially harmful: "The man who impulsively sacrifices himself to help anyone, including gangsters and tyrants, may be acting selflessly, but he is also a social liability." Correct. This was the point of my example in the target article of the Nazi soldier sacrificing himself for his unit. Wagstaff makes the point, implicit in **Lacey**'s commentary, that no account of altruism is complete without an understanding of justice. Such an understanding would enable us to distinguish more clearly between altruism and morality. I agree.

References

[Note: The letters "a" and "r" before author's initials stand for target article and response references, respectively.]

Abelson, R. P., Frey, K. P. & Gregg, A. P. (in press) *Experiments with people: Revelations from social psychology.* Erlbaum. [CS]

Ainslie, G. (1992) *Picoeconomics.* Cambridge University Press. [aHR, JS]
(1995) A utility-maximizing mechanism for vicarious reward. *Rationality and Society* 7:393–403. [GA]
(2001) *Breakdown of will.* Cambridge University Press. [GA, HM]

Ainslie, G. & Monterosso, J. (in press) Building blocks of self-control: Increased tolerance for delay with bundled rewards. *Journal of the Experimental Analysis of Behavior.* [GA]

Alexander, R. D. (1987) *The biology of moral systems.* Aldine de Gruyter. [JS, CW, DJZ]

Apostle, H. G. (1984) Translator, Aristotle's *Nicomachean ethics.* The Peripatetic Press. [rHR]

Aristotle (1984) *The complete works of Aristotle,* ed. Barnes. Princeton University Press. [GFW]

Aron, A., Aron, E. N. & Smollan, D. (1992) Inclusion of other in the self scale and the structure of interpersonal closeness. *Journal of Personality and Social Psychology* 63:596–612. [CS]

Aron, A., Aron, E. N., Tudor, M. & Nelson, G. (1991) Close relationships as including other in the self. *Journal of Personality and Social Psychology* 60:241–53. [CS]

Axelrod, R. (1984) *The evolution of cooperation.* Basic Books. [AAS, CW, TRZ]
(1997) *The complexity of cooperation: Agent based models of competition and collaboration.* Princeton University Press. [aHR]

Axelrod, R. & Hamilton, W. D. (1981) The evolution of cooperation. *Science* 211:1390–96. [HG, TRZ]

Bacharach, M., Guerra, G. & Zizzo, D. J. (2001) Is trust self-fulfilling? An experimental study. Department of Economics Discussion Paper No. 76, University of Oxford. http://www.economics.ox.ac.uk/Research/working_papers.htm [[DJZ]

Baker, F. & Rachlin, H. (2001) Probability of reciprocation in prisoner's dilemma games. *Journal of Behavioral Decision Making* 14:51–67. [JIK, arHR]
(2002) Self-control by pigeons in the prisoner's dilemma. *Psychonomic Bulletin and Review* 9:482–88. [rHR]
(2002) Teaching and learning in a probabilistic prisoner's dilemma. *Behavioral Processes.* 57:211–26. [aHR]

Bandura, A. (1972) Modeling theory: Some traditions, trends, and disputes. In:

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

Recent trends in social learning theory, ed. R. D. Parke. Academic Press. [RCG]
Bandura, A., Ross, D. & Ross, S. A. (1963) Vicarious reinforcement and imitative learning. *Journal of Abnormal and Social Psychology* 67:601–607. [RCG]
Baron, J. (1997a) Political action vs. voluntarism in social dilemmas and aid for the needy. *Rationality and Society* 9:307–26. [JB]
(1997b) The illusion of morality as self-interest: A reason to cooperate in social dilemmas. *Psychological Science* 8:330–35. [JB]
(2001) Confusion of group-interest and self-interest in parochial cooperation on behalf of a group. *Journal of Conflict Resolution* 45:283–96. [JB]
Batali, J. (1998) Computational simulations of the emergence of grammar. In: *Approaches to the evolution of language: Social and cognitive bases,* ed. J. Hurford & M. Studdert-Kennedy. Cambridge University Press. [WZ]
Batson, C. D. (1987) Prosocial motivation: Is it ever truly altruistic? In: *Advances in experimental social psychology, vol. 20,* ed. L. Berkowitz. Academic Press. [GFW]
(1990) How social an animal? The human capacity for caring. *American Psychologist* 45:336–46. [MP]
(1991) *The altruism question: Toward a social-psychological answer.* Erlbaum. [DSW]
(1998) Altruism and prosocial behavior. In: *The handbook of social psychology, vol. 2,* ed. D. T. Gilbert, S. T. Fiske & G. Lindzey. McGraw-Hill. [GC]
Batson, C. D., Batson, J. G., Todd, M. R., Brummett, B. H., Shaw, L. L. & Aldeguer, C. M. R. (1995) Empathy and the collective good: Caring for one of the others in a social dilemma. *Journal of Personality and Social Psychology* 68:619–31. [MP]
Batson, C. D. & Oleson, K. C. (1991) Current status of the empathy-altruism hypothesis. In: *Review of Personality and Social Psychology, vol. 12: Altruism,* ed. M. Clark. Sage. [RB]
Batson, C. D., Sager, K., Garst, E., Kang, M., Rubchinsky, K. & Dawson, K. (1997) Is empathy-induced helping due to self-other merging? *Journal of Personality and Social Psychology* 73:495–509. [CS]
Batson, C. D. & Shaw, L. L. (1991) Evidence for altruism: Toward a pluralism of prosocial motives. *Psychological Inquiry* 2:159–68 or 2:107–22. [GA, RB]
Baum, W. M. (1994) *Understanding behaviorism: Science, behavior, and culture.* Harper Collins. [aHR]
Baum, W. M. & Rachlin, H. (1969) Choice as time allocation. *Journal of the Experimental Analysis of Behavior* 12:861–74. [EJF]
Baumeister, R. F. (1999) *The self in social psychology.* Psychology Press. [RAH]
(2000) Ego depletion and the self's executive function. In: *Psychological perspectives on self and identity,* ed. A. Tesser, R. B. Felson & J. M. Suls. American Psychological Association. [rHR]
Baumeister, R. F. & Leary, M. R. (1995) The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin* 117:497–529. [CS]
Bechara, A., Damasio, H., Tranel, D. & Anderson, S. W. (1998) Dissociation of working memory from decision making within the human prefrontal cortex. *Journal of Neuroscience* 18:428–37. [JRG]
Becker, G. S. (1981) Altruism in the family and selfishness in the market place. *Economica* 48:1–15. [EKL]
Beecher, M. D., Campbell, S. E. & Nordby, J. C. (2000) Territory tenure in song sparrows is related to song sharing with neighbors, but not to repertoire size. *Animal Behaviour* 59:29–37. [RS]
Beggan, J. K. (1992) On the social nature of nonsocial perception: The mere ownership effect. *Journal of Personality and Social Psychology* 62:229–37. [CS]
Bergstrom, T. C. (1995) On the evolution of altruistic ethical rules for siblings. *American Economic Review* 85:58–81. [DJZ]
Berridge, K. C. (2000) Measuring hedonic impact in animals and infants: Microstructure of affective taste reactivity patterns. *Neuroscience and Biobehavioral Reviews* 24:173–98. [RS]
Bickel, W. K. & Vuchinich, R. E. (2000) *Reframing health behavior change with behavioral economics.* Erlbaum. [aHR]
Bindra, D. (1959) *Motivation: A systematic reinterpretation.* Ronald Press. [RAH]
Boehm, C. (1997) Impact of the human egalitarian syndrome on Darwinian selection mechanics. *The American Naturalist* 34:S100-S121. [JS]
Boesch, C. & Boesch, H. (1989) Hunting behavior of wild chimpanzees in the Tai National Park. *American Journal of Physical Anthropology* 78:547–73. [RS]
Boksa, P., Wilson, D. & Rochford, J. (1998) Responses to stress and novelty in adult rats born vaginally, by caesarean section, or caesarean section with acute anoxia. *Biology of the Neonate* 74:48–59. [GC]
Bowlby, J. (1969) *Attachment and loss, vol. 1: Attachment.* Basic Books. [DSW]
Boyd, R. & Richerson, P. J. (1985) *Culture and the evolutionary process.* University of Chicago Press. [HG]
Brann, P. & Foddy, M. (1988) Trust and the consumption of a deteriorating common resource. *Journal of Conflict Resolution* 31:615–30. [aHR, RS]
Braver, T. S. & Cohen, J. D. (2000) On the control of control: The role of dopamine in regulating prefrontal function and working memory. In: *Control of processes: Attention and Performance XVIII,* ed. S. Monsell & J. Driver. MIT Press. [JRG]
Broude, G. J. (1995) *Growing up.* ABC-CLIO. [GJB]
Brown, J. (2000) Delay discounting if multiple reinforcers following a single choice. Unpublished doctoral dissertation, Psychology Department, State University of New York at Stony Brook. [aHR]
Brown, J. & Rachlin, H. (1999) Self-control and social cooperation. *Behavioral Processes* 47:65–72 [aHR]
Buck, R. (1985) Prime theory: An integrated view of motivation and emotion. *Psychological Review* 92:389–413. [RB]
(1988) *Human motivation and emotion, 2nd edition.* Wiley. [RB]
(1999) The biological affects: A typology. *Psychological Review* 106(2):729–38. [RB]
Campbell, D. T. (1975) Conflicts between biological and social evolution and between psychology and moral traditions. *American Psychologist* 30:1103–126. [SK]
Caporael, L. R., Dawes, R. M., Orbel, J. M. & van de Kragt, A. J. C. (1989) Selfishness examined: Cooperation in the absence of egoistic incentives. *Behavioral and Brian Sciences* 12:683–739. [aHR]
Carlo, G., Eisenberg, N., Troyer, D., Switzer, G. & Speer, A. L. (1991) The altruistic personality: In what contexts is it apparent? *Journal of Personality and Social Psychology* 61:450–58. [MP]
Carlo, G. & Randall, B. (2001) Are all prosocial behaviors equal? A socioecological developmental conception of prosocial behavior. In: *Advances in psychology research, vol. II,* ed. F. Columbus. Nova Science. [GC]
Caro, T. M. (1994) *Cheetahs of the Serengeti Plains: Group living in an asocial species.* University of Chicago Press. [RS]
Cavalli-Sforza, L. L. & Feldman, M. W. (1981) *Cultural transmission and evolution.* Princeton University Press. [HG]
Chapman, G. B. (1998) Sooner or later: The psychology of intertemporal choice. In: *The psychology of learning and motivation, vol. 38,* ed. D. L. Medin. Academic Press. [RCG]
Charles, R. (1962) Them that got. In: *Ray Charles Greatest Hits* (record album; for lyrics, visit http://www.thepeaches.com/music/raycharles/ThemThatGot.txt). [JH]
Charness, G. & Rabin, M. (2000) Social preferences: Some simple tests and a new model. Department of Economics Working Paper No. 283, University of California at Berkeley. http://www.haas.berkeley.edu/groups/iber/wps/econwp.html [DJZ]
Chartrand, T. L. & Bargh, J. A. (1999) The chameleon effect: The perception-behavior link and social interaction. *Journal of Personality and Social Psychology* 76:893–910. [RS]
Chisholm, J. S. (1999) *Death, hope and sex.* Cambridge University Press. [DSW]
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. *Journal of Personality and Social Psychology* 73:481–94. [CS]
Cialdini, R. B., Daumann, D. J. & Kendrick, D. T. (1981) Insights from sadness: A three step model of the development of altruism as hedonism. *Developmental Review* 1:207–23. [GFW]
Clements, K. C. & Stephens, D. W. (1995) Testing models of non-kin cooperation: Mutualism and the Prisoner's Dilemma. *Animal Behavior* 50:527–35. [RS]
Colby, A. & Damon, W. (1995) The development of extraordinary moral commitment. In: *Morality in everyday life,* ed. M. Killen & D. Hart. Cambridge University Press. [RAH]
Cookson, R. (2000) Framing effects in public good experiments. *Experimental Economics* 3:55–79. [DJZ]
Cosmides, L. & Tooby, J. (1992) Cognitive adaptations for social exchange. In: *The adapted mind,* ed. J. H. Barkow, L. Cosmides & J. Tooby. Oxford University Press. [GJB]
Danielson, P. (2002) Competition among cooperators: Altruism and reciprocity. *Proceedings of the National Academy of Sciences USA* 99 (Suppl. 3):7237–42. [PD]
Darley, J. M. & Batson, C. D. (1973) From Jerusalem to Jericho: A study of situational and dispositional variables in helping behavior. *Journal of Personality and Social Psychology* 27:100–108. [JIK]
Daumann, D. J., Cialdini, R. B. & Kendrick, D. T. (1981) Altruism as hedonism: Helping and self-gratification as equivalent responses. *Journal of Personality and Social Psychology* 40:1039–46. [GFW]
Davis, M. H., Luce, C. & Kraus, S. J. (1994) The heritability of characteristics associated with dispositional empathy. *Journal of Personality* 62:362–91. [JH]
Dawes, R. (1980) Social dilemmas. *Annual Review of Psychology* 31:169–93. [aHR]
Dawkins, R. (1976/1989) *The selfish gene.* (2nd edition, 1989). Oxford University Press. [aHR, RS, GFW]
De Bruin, E. N. M. & van Lange, P. A. M. (1999) Impression formation and cooperative behavior. *European Journal of Social Psychology* 29:305–28. [JIK]

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

Deguchi, H. (1984) Observational learning from a radical-behavioristic viewpoint. *The Behavior Analyst* 7:83–95. [RCG]
Dellu, F., Mayo, W., Vallèe, M., Piazza, P. V., Le Moal, M. & Simon, H. (1996) Behavioral reactivity to novelty during youth as a predictive factor of stress-induced corticosterone secretion in the elderly – a life-span study in rats. *Psychoneuroendocrinology* 21:441–53. [GC]
Dennett, D. C. (1984) *Elbow room: The varieties of free will worth wanting.* MIT Press. [aHR]
Derryberry, D. & Rothbart, M. K. (1988) Arousal, affect, and attention as components of temperament. *Journal of Personality and Social Psychology* 55:958–66. [GC]
De Waal, F. B. M. & Aureli, F. (1997) Conflict resolution and distress alleviation in monkeys and apes. In: *Annals of the New York Academy of Sciences, vol. 807: The integrative neurobiology of affiliation,* ed. C. S. Carter, I. I. Lederhendler & B. Kirkpatrick. The New York Academy of Sciences. [RB]
Dovidio, J. F., Gaertner, S. L., Isen, A. M. & Lowrance, R. (1995) Group representations and intergroup bias: Positive affect, similarity, and group size. *Personality and Social Psychology Bulletin* 21:856–65. [JRG]
Duckitt, J. (1992) *The social psychology of prejudice.* Praeger. [CS]
Dugatkin, L. A. (1997) *Cooperation among animals: An evolutionary perspective.* Oxford University Press. [RS]
Dunning, D. (1993) Words to live by: The self and definitions of social concepts and categories. In: *Psychological perspectives on the self, vol. 4,* ed. J. Suls. Erlbaum. [CS]
Dyson, F. (1970) Time without end: Physics and biology in an open universe. *Reviews of Modern Physics* 51:447–60. [JH]
Edney, J. J. (1980) The commons problem: Alternative perspectives. *American Psychologist* 35:131–50. [aHR]
Eisenberg, N. & Fabes, R. A. (1991) Prosocial behavior and empathy: A multimethod developmental perspective. In: *Prosocial behavior,* ed. M. S. Clark. Sage. [RB]
(1998) Prosocial development. In: *Handbook of child psychology, vol. 3: Social, emotional, and personality development, 5th edition,* ed. N. Eisenberg. (Series ed. W. Damon). Wiley. [GC]
Eisenberg, N. & Miller, P. (1987) The relation of empathy to prosocial and related behaviors. *Psychological Bulletin* 101:91–119. [RB]
Elster, J. (2000) *Ulysses bound.* Cambridge University Press. [ELK]
Falk, A. & Fischbacher, U. (1998) A theory of reciprocity. IEW Working Paper No. 6, University of Zurich. http://www.iew.unizh.ch/home/fischbacher [DJZ]
Fantino, E. & Stolarz-Fantino, S. (2002) From patterns to prosperity: A review of Rachlin's *The science of self-control. Journal of the Experimental Analysis of Behavior* 78:117–25. [EJF]
Fehr, E. & Gächter, S. (2000) Fairness and retaliation: The economics of reciprocity. *Journal of Economic Perspectives* 14:159–81. [ELK]
Fehr, E. & Schmidt, K. (1999) A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics* 114:817–68. [DJZ]
(2002) Altruistic punishment in humans. *Nature* 415:137–40. [HG, CW]
Field, A. (2001) *Altruistically inclined? The behavioral sciences evolutionary theory and the origins of reciprocity.* University of Michigan Press. [GA]
Fiske, A. P. (1991) The cultural relativity of selfish individualism: Anthropological evidence that humans are inherently sociable. In: *Review of Personality and Social Psychology, vol. 12: Altruism and social behavior,* ed. M. Clark. Sage. [GA]
Foddy, M., Smithson, M., Schneider, S. & Hogg, M. A., eds. (1999) Resolving social dilemmas: Dynamics, structural, and intergroup aspects. *Psychology Press.* [AAS]
Foddy, M. & Veronese, D. (1996) Does knowing the jointly rational solution make you want to pursue it? In: *Frontiers in social dilemmas research,* ed. W. B. G. Liebrand & D. M. Messick. Springer. [AAS]
Forsythe, R., Horowitz, J., Savin, N. & Sefton, M. (1994) Replicability, fairness and play in experiments with simple bargaining games. *Games and Economic Behavior* 6:347–69. [RS]
Frank, R. H. (1988) *Passions within reason: The strategic role of the emotions.* W. W. Norton. [RS]
Frank, R. H., Gilovich, T. & Regan, D. (1993) Does studying economics inhibit cooperation? *Journal of Economic Perspectives* 7:159–71. [GA]
Frautschi, S. (1987) Entropy in an expanding universe. *Science* 217:593–99. [JH]
Fry, P. S. (1977) Success, failure, and resistance to temptation. *Developmental Psychology* 13:519–20. [JRG]
Fudenberg, D. & Maskin, E. (1990) Evolution and cooperation in noisy repeated games. *New Developments in EconomicTheory* 80:274–79. [aHR]
Gallucci, M. & Perugini, M. (2000) An experimental test of a game-theoretical model of reciprocity. *Journal of Behavioral Decision Making* 13:367–89. [MP]
Gauthier, D. (1986) *Morals by agreement.* Oxford University Press. [PD]
Gilligan, C. & Attanucci, J. (1988) Two moral orientations: Gender differences and similarities. *Merrill-Palmer Quarterly* 34(3):223–37. [RB]
Gintis, H., Smith, E. A. & Boules, S. (2001) Costly signalling and cooperation. *Journal of Theoretical Biology* 213:103–19. [DR, CW]
Goldschmidt, W. (1990) *The human career: The self in the symbolic world.* Blackwell. [SK]
Gray, J. R. (1999) A bias toward short-term thinking in threat-related negative emotional states. *Personality and Social Psychology Bulletin* 25:65–75. [JRG]
(2001) Emotional modulation of cognitive control: Approach-withdrawal states double-dissociate spatial from verbal two-back task performance. *Journal of Experimental Psychology: General* 130:436–52. [JRG]
Gray, J. R., Braver, T. S. & Raichle, M. E. (2001) Integration of emotion and cognition in lateral prefrontal cortex. *Proceedings of the National Academy of Sciences USA* 99:4115–20. [JRG]
Green, L., Price, P. C. & Hamburger, M. E. (1995) Prisoner's dilemma and the pigeon: Control by immediate consequences. *Journal of the Experimental Analysis of Behavior* 64:1–17. [RS]
Green, L. & Rachlin, H. (1996) Commitment using punishment. *Journal of the Experimental Analysis of Behavior* 65:593–601. [aHR]
Grim, P. (1995) Greater generosity of the spatialized prisoner's dilemma. *Journal of Theoretical Biology* 173:353–59. [PG]
(1996) Spatialization and greater generosity in the stochastic prisoner's dilemma. *BioSystems* 37:3–17. [PG]
Grim, P., St. Denis, P. & Kokalis, T. (2002) Learning to communicate: The emergence of signaling in spatialized arrays of neural nets. *Adaptive Behavior.* (forthcoming). [PG]
Gul, F. & Pesendorfer, W. (2001) Temptation and self-control. *Econometrica* 69:1403–35. [JS]
Gurven, M., Allen-Arave, W., Hill, K. & Hurtado, M. (2000) "It's a wonderful life": Signaling generosity among the Ache of Paraguay. *Evolution and Human Behavior* 21:263–82. [DR]
Guth, W., Marchand, N. & Rullière, J.-L. (1998) Equilibration et dépendance du context: Une évaluation expérimentale du jeu de négociation sous ultimatum. *Revue Economique* 49:785–94. [DJZ]
Hake, D. F. & Vukelich, R. (1972) A classification and review of cooperation procedures. *Journal of the Experimental Analysis of Behavior* 18:333–43. [RS]
Hamilton, W. D. (1964) The evolution of social behavior. *Journal of Theoretical Biology* 7:1–52. [DR, JS]
Hardin, G. (1968) The tragedy of the commons. *Science* 162:1243–48. [CW]
Harlow, H. F. (1953) Mice, monkeys, men and motives. *Psychological Review* 60:23–32. [SK]
Harper, L. (1989) *The nurture of human behavior.* Ablex. [GJB]
Harter, S. (1998) The development of self-representations. In: *Handbook of child psychology: Social, emotional and personality development, vol. 3,* ed. W. Damon & N. Eisenberg. Wiley. [RAH]
Hartung, J. (1995) Love thy neighbor: The evolution of in-group morality. *Skeptic* 3(4):86–98. [JH]
(1996) Prospects for existence: Morality and genetic engineering. *Skeptic* 4(2):62–71. [JH]
(2001) The message of evolution. http://members.aol.com/_ht_a/toexist/Message.html [JH]
Hauert, C. & Schuster, H. G. (1997) Effects of increasing the number of players and memory size in the iterated Prisoner's Dilemma: A numerical approach. *Proceedings of the Royal Society of London B* 264:513–19. [CW]
Hawkes, K., O'Connell, J. F. & Blurton Jones, N. G. (2001) Hadza meat sharing. *Evolution and Human Behavior* 22:113–42. [RAH]
Hayes, S. C. (1989) *Rule-governed behavior: Cognition, contingencies, and instructional control.* Plenum Press. [aHR]
Herrnstein, R. J. (1991) Experiments on stable suboptimality in individual behavior. *American Economic Review* 81:360–64. [HM, aHR]
Herrnstein, R. J. & Prelec, D. (1992) A theory of addiction. In: *Choice over time,* ed. G. Loewenstein & Elster. Russell Sage Foundation. [aHR]
Herrnstein, R. J., Prelec, D. & Vaughan, W, Jr. (1986) An intra-personal prisoners' dilemma. Paper presented at the IX Symposium on Quantitative Analysis of Behavior: Behavioral Economics, Harvard University. [aHR]
Heyman, G. M. (1996) Resolving the contradictions of addiction. *Behavioral and Brain Sciences* 19:561–610. [aHR]
Hinde, R. A. (2002) *Why good is good.* Routledge. [RAH]
Hoffman, M. L. (1975) Developmental synthesis of affect and cognition and its implications for altruistic motivation. *Developmental Psychology* 11:607–22. [RB]
(1976) Empathy, role-taking, guilt, and development of altruistic motives. In: *Moral development and behavior: Theory, research, and social issues,* ed. T. Lickona. Holt, Rinehart and Winston. [RB]
(1991) Empathy, social cognition, and moral action. In: *Handbook of moral behavior and development, vol. 1: Theory,* ed. W. M. Kurtines & J. L. Gewirtz. Erlbaum. [GC]
Hoyle, F., Burbridge, G. & Narlikar, J. (2000) *A different approach to cosmology:*

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

From a static universe through the big bang towards reality. Cambridge University Press. [JH]
Hurford, J. R. (2002) Expression/induction models of language. In: *Linguistic evolution through language acquisition: Formal and computational models,* ed. T. Briscoe. Cambridge University Press. [WZ]
Irwin, F. W. (1971) *Intentional behavior and motivation: A cognitive theory.* Lippincott. [JB]
Isen, A. M. (1972) Effects of feeling good on helping others: Cookies and kindness. *Journal of Personality and Social Psychology* 21:382–88. [JRG]
James, W. (1892) *Psychology: The briefer course.* Holt. [SK]
Kahneman, D., Knetsch, J. L. & Thaler, R. H. (1990) Experimental test of the endowment effect and the coase theorem. *Journal of Political Economy* 98:1325–47. [CS]
Kaplan, S. (1995) The restorative benefits of nature: Toward an integrative framework. *Journal of Experimental Psychology* 15:169–82. [SK]
Kehoe, P., Shoemaker, W. J., Triano, L., Callahan, M. & Rappolt, G. (1998) Adult rats stressed as neonates show exaggerated behavioral responses to both pharmacological and environmental challenges. *Behavioral Neuroscience* 112:116–25. [GC]
Kelley, H. H. & Thibaut, J. W. (1978) *Interpersonal relations: A theory of interdependence.* Wiley. [MP]
Khalil, E. L. (1999) Sentimental fools: A critique of Amartya Sen's notion of commitment. *Journal of Economic Behavior and Organization* 40:373–86. [ELK]
(2000) Types of metaphor and identificational slips in economic discourse. *Research in the History of Economic Thought and Methodology* 18A:83–105. [ELK]
(2001) Adam Smith and three theories of altruism. *Recherches Économiques de Louvain* (*Louvain Economic Review*) 67:421–35. [rHR]
(2002) What is altruism? *Journal of Economic Psychology* 23. (in press). [ELK]
(in press) Why does trustworthiness pay? Three explanations: An introduction. In: *Trust,* ed. E. L. Khalil. Edward Elgar. [ELK]
Kirby, K. N. (1997) Bidding on the future: Evidence against normative discounting of delayed rewards. *Journal of Experimental Psychology: General* 126:54–70. [GA, RCG]
Kirby, K. N. & Guastello, B. (2001) Making choices in anticipation of similar future choices can increase self-control. *Journal of Experimental Psychology: Applied* 7:154–64. [GA]
Kirby, S. (2000) Syntax without natural selection: How compositionality emerges from vocabulary in a population of learners. In: *The evolutionary emergence of language: Social function and the origins of linguistic form,* ed. C. Knight, J. Hurford & M. Studdert-Kennedy. Cambridge University Press. http://ling.ed.ac.uk/anonftp/pub/staff/kirby/evo198.ps.gz [WZ]
Kitayama, S. & Karasawa, M. (1997) Implicit self-esteem in Japan: Name letters and birthday numbers. *Personality and Social Psychology Bulletin* 23:736–42. [CS]
Komorito, S. S., Chan, D. K.-S. & Parks, C. D. (1993) The effects of reward structure and reciprocity in social dilemmas. *Journal of Experimental Social Psychology* 29:252–67. [aHR]
Komorito, S. S. & Parks, C. D. (1994) *Social dilemmas.* Brown and Benchmark. [aHR]
Konow, J. (2000) Fair shares: Accountability and cognitive dissonance in allocation decisions. *American Economic Review* 90:1072–91. [DJZ]
Krebs, J. R. (1982) Territorial defense in the great tit (*Parus major*): Do residents always win? *Behavioral Ecology and Sociobiology* 11:185–94. [RS]
Krebs, J. R. & Davies, N. B. (1993) *An introduction to behavioural ecology, third edition.* Blackwell. [RS]
Kreps, D. M., Milgrom, P., Roberts, J. & Wilson, R. (1982) Rational cooperation in the finitely repeated Prisoners' Dilemma. *Journal of Economic Theory* 27:245–52. [DJZ]
Krueger, J. (1998) On the perception of social consensus. *Advances in Experimental Social Psychology* 30:163–240. [JIK]
Kudadjie-Gyamfi, E. (1998) Patterns of behavior: Self-control choices among risky alternatives. Unpublished doctoral dissertation, Psychology Department, State University of New York at Stony Brook. [aHR]
Kudadjie-Gyamfi, E. & Rachlin, H. (1996) Temporal patterning in choice among delayed outcomes. *Organizational Behavior and Human Decision Processes* 65:61–67 [aHR]
Lacey, H. (1995a) Teleological behaviorism and the intentional scheme. *Behavioral and Brain Sciences* 18:134–35. [HL]
(1995b) Behaviorisms: Theoretical and teleological. Review of John Staddon's *Behaviorism: Mind, mechanism and society* and Howard Rachlin's *Behavior and mind: The roots of modern psychology. Behavior and Philosophy* 23:61–78. [HL]
Lacey, H. & Schwartz, B. (1996) The formation and transformation of values. In: *The philosophy of psychology,* ed. W. O'Donohue & R. F. Kitchener. Sage. [HL]
Larrick, R. P. & Blount, S. (1997) The claiming effect: Why players are more generous in social dilemmas than in ultimatum games. *Journal of Personality and Social Psychology* 72:810–25. [MP]
Lawler, E. J., Thye, S. R. & Yoon, J. (2000) Emotion and group cohesion in productive exchange. *American Journal of Sociology* 106:616–57. [JRG]
Leary, M. R. & Baumeister, R. F. (2000) The nature and function of self-esteem: Sociomotor theory. In: *Advances in experimental social psychology, vol. 32,* ed. M. P. Zanna. Academic Press. [CS]
Levine, D. K. (1998) Modeling altruism and spitefulness in experiments. *Review of Economic Dynamics* 1:593–622. [PD]
Lewin, K. (1936) *Principles of topological psychology.* McGraw-Hill. [aHR]
Lewis, M. (1979) The self as a developmental concept. *Human Development* 22:416–19. [ML]
(1992) *Shame: The exposed self.* The Free Press. [ML]
(1995) Self-conscious emotions. *American Scientist* 83:68–78. [ML]
Lichbach, M. I. (1996) *The cooperator's dilemma.* University of Michigan Press. [AAS]
Liu, D., Diorio, J., Tannanbaum, B., Caldji, C., Francis, D., Freedman, A., Sharma, S., Pearson, D., Plotsky, P. M. & Meaney, M. J. (1997) Maternal care, hippocampal glucocorticoid receptors, and hypothalamic-pituitary-adrenal responses to stress. *Science* 277:1659–62. [GC]
Logue, A. W. (1988) Research on self-control: An integrating framework. *Behavioral and Brain Sciences* 11:665–709. [RS]
Logue, A. W., King, G. R., Chavarro, A. & Volpe, J. S. (1990) Matching and maximizing in a self-control paradigm using human subjects. *Learning and Motivation* 21:340–68. [RS]
Lowenstein, G. F., Weber, E. U., Hsee, C. K. & Welch, N. (2001) Risk as feelings. *Psychological Bulletin* 127:267–86. [RB]
Lumsden, C. J. & Wilson, E. O. (1981) *Genes, mind, and culture: The coevolutionary process.* Harvard University Press. [HG]
MacCorquodale, K. & Meehl, P. E. (1954) Edward C. Tolman. In: *Modern learning theory,* ed. W. K. Estes, S. Koch & et al. Appleton-Century-Crofts. [RAH]
MacKay, D. M. (1965) Man as mechanism. In: *The open mind and other essays,* ed. D. M. MacKay & M. Tinker. Inter-Varsity Press. [SK]
Margolis, H. (1982) *Selfishness, altruism and rationality.* Cambridge University Press. (Paper reprint 1984, University of Chicago Press). [HM]
(2000) Self-interest and social motivation (Working Paper). http://www.harrisschool.uchicago.edu/wp/00-5.html [HM]
Marinoff, L. (1998) The failure of success: Intrafamilial exploitation in the prisoner's dilemma. In: *Modeling rationality, morality, and evolution, vol. 7,* ed. P. Danielson. Oxford University Press. [PD]
Maynard Smith, J. (1978) *The evolution of sex.* Cambridge University Press. [RS]
(1982) *Evolution and the theory of games.* Cambridge University Press. [WZ]
Mazur, J. E. (1987) An adjusting procedure for studying delayed reinforcement. In: *Quantitative analysis of behavior, 5: The effects of delay and of intervening events on reinforcement value,* ed. M. L. Commons, J. E. Mazur, J. A. Nevin & H. Rachlin. Erlbaum. [aHR]
McClennen, E. (1998) Rationality and rules. In: *Modeling rationality, morality, and evolution, vol. 7,* ed. P. Danielson. Oxford University Press. [PD, PW]
McClintock, C. G. (1972) Social motivation: A set of propositions. *Behavioral Science* 17:438–54. [AAS]
McGuire, W. J. & McGuire, C. V. (1988) Content and process in the experience of self. *Advances in Experimental Social Psychology* 21:97–144. [RAH]
Medawar, P. B. & Medawar, J. S. (1983) *Aristotle to Zoos: A philosophical dictionary of biology.* Harvard University Press. [DR]
Meltzoff, A. N. & Moore, M. K. (1977) Imitation of facial and manual gestures by human neonates. *Science* 198:75–78. [RS]
Messick, D. M. & McClelland, C. L. (1983) Social traps and temporal traps. *Personality and Social Psychology Bulletin* 9:105–10. [ELK, aHR]
Mesuluam, M. (1985) *Principles of behavioral neurology.* F. A. Davis. [SK]
Midgley, M. (1978) *Beast and man.* Cornell University Press. [GJB]
Milinski, M., Semmann, D., Bakker, T. C. M. & Krambeck, H.-J. (2001) Cooperation through indirect reciprocity: Image scoring or standing strategy. *Proceedings of the Royal Society of London B* 268:2495–501. [CW]
Milinski, M., Semmann, D. & Krambeck, H.-J. (2002) Reputation helps solve the "tragedy of the commons." *Nature* 415:424–26. [CW]
Milinski, M. & Wedekind, C. (1998) Working memory constrains human cooperation in the Prisoner's Dilemma. *Proceedings of the National Academy of Sciences USA* 95:13755–58. [CW]
Monterosso, J., Ainslie, G., Toppi-Mullen, P. & Gault, B. (2002) The fragility of cooperation: A false feedback study of a sequential iterated prisoner's dilemma. *Journal of Economic Psychology* 23:437–48. [GA]
Nagel, T. (1970) *The possibility of altruism.* Clarendon Press. [PD]
Nesse, R. (2001) The evolution of subjective commitment. In: *Evolution and the capacity for commitment,* ed. R. Nesse. Russell Sage. [JS]
Neuberg, S. L., Cialdini, R. B., Brown, S. L., Luce, C. & Sagarin, B. J. (1997) Does empathy lead to anything more than superficial helping? Comment on Batson et al. 1997. *Journal of Personality and Social Psychology* 73:510–16. [CS]

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

Nowak, M. A., May, R. M. & Sigmund, K. (1995) The arithmetics of mutual help. *Scientific American* 272:76–81. [CW]
Nowak, M. A., Page, K. M. & Sigmund, K. (2000) Fairness versus reason in the Ultimatum Game. *Science* 289:1773–75. [CW]
Nowak, M. A. & Sigmund, K. (1992) Tit-for-tat in heterogeneous populations. *Nature* 355:250–53. [CW]
(1993) A strategy of Win-Stay-Lose-Shift that outperforms Tit-for-Tat in the Prisoner's Dilemma game. *Nature* 364:56–58. [aHR, CW]
(1998a) The dynamics of indirect reciprocity. *Journal of Theoretical Biology* 194:561–74. [CW]
(1998b) Evolution of indirect reciprocity by image scoring. *Nature* 393:573–77. [CW]
Nozick, R. (1993) *The nature of rationality.* Princeton University Press. [PW]
Nuttin, J. M. (1987) Affective consequences of mere ownership: The name letter effect in twelve European languages. *European Journal of Social Psychology* 17:381–402. [CS]
Ochsner, K. N. & Lieberman, M. D. (2001) The emergence of social cognitive neuroscience. *American Psychologist* 56:717–34. [JRG]
Ostrom, E. (1998) A behavioral approach to the rational choice theory of collective action. *American Political Science Review* 92:1–22. [AAS]
Packer, C., Scheel, D. & Pusey, A. E. (1990) Why lions form groups: Food is not enough. *American Naturalist* 136:1–19. [RS]
Paine, T. (1776) *The American crisis.* Pamphlet No. 1. Thomas Paine. [JH]
Palameta, B. & Brown, W. M. (1999) Human cooperation is more than by-product mutualism. *Animal Behaviour* 57:F1-F3. [RS]
Palfrey, T. R. & Prisbrey, J. E. (1997) Anomalous behavior in public goods experiments: How much and why? *American Economic Review* 87:829–46. [DJZ]
Panksepp, J. (1998) *Affective neuroscience: The foundations of human and animal emotions.* Oxford University Press. [GC]
Panksepp, J., Nelson, E. & Bekkedal, M. (1997) Brain systems for the mediation of social separation-distress and social-reward. In: *The integrative neurobiology of affiliation,* ed. C. S. Carter, I. Lederhendler & B. Kirkpatrick. New York Academy of Sciences. [RS]
Partif, D. (1971) Personal identity. *Philosophical Review* 80:3–27. [aHR]
Perugini, M. & Gallucci, M. (2001) Individual differences and social norms: The distinction between reciprocators and prosocials. *European Journal of Personality* 15:S19-S35. [MP]
Phillips, S. T. & Ziller, R. C. (1997) Toward a theory and measure of the nature of prejudice. *Journal of Personality and Social Psychology* 72:420–34. [CS]
Piliavin, J. A. & Callero, P. L. (1991) *Giving blood: The development of an altruistic identity.* Johns Hopkins University Press. [AAS]
Pinker, S. & Bloom, P. (1990) Natural language and natural selection. *Behavioral and Brain Sciences* 13:707–84. [WZ]
Platow, M. J. (1993) Observing social value orientations: A social interdependence approach. *New Zealand Journal of Psychology* 22:101–109. [AAS]
Platow, M. J., Durante, M., Williams, N., Garrett, M., Walshe, J., Cincotta, S., Lianos, G. & Barutchu, A. (1999) The contribution of sport fan social identity to the production of prosocial behavior. *Group Dynamics: Theory, Research and Practice* 3:161–69. [AAS]
Platt, J. (1973) Social traps. *American Psychologist* 28:641–51. [ELK, aHR]
Posch, M. (1999) Win-stay, lose-shift strategies for repeated games – memory length, aspiration levels and noise. *Journal of Theoretical Biology* 198:183–95. [CW]
Poulos, C. X., Parker, J. L. & Lê, D. A. (1998) Increased impulsivity after injected alcohol predicts later alcohol consumption in rats: Evidence for "loss-of-control drinking" and marked individual differences. *Behavioral Neuroscience* 112:1247–57. [GC]
Premack, D. (1965) Reinforcement theory. In: *Nebraska Symposium on Motivation,* ed. D. Levine. University of Nebraska Press. [aHR]
(1971) Catching up with common sense or two sides of generalization: Reinforcement and punishment. In: *The nature of reinforcement,* ed. R. Glaser. Academic Press. [rHR]
Preston, S. D. & de Waal, F. B. M. (2002) Empathy: Its ultimate and proximate bases. *Behavioral and Brain Sciences* 25(1):1–70. [GA, RB]
Prins, K. S., Buunk, B. P. & van Yperen, N. W. (1993) Equity, normative disapproval and extra martial relationships. *Journal of Social and Personal Relationships* 10:39–53. [RAH]
Quartz, S. R. & Sejnowksi, T. J. (1997) The neural basis of cognitive development: A constructivist manifesto. *Behavioral and Brain Sciences* 20:537–96. [DJZ]
Quattrone, G. A. & Tversky, A. (1984) Causal versus diagnostic contingencies: On self-deception and the voter's illusion. *Journal of Personality and Social Psychology* 46:237–48. [JB, JIK]
Rachlin, H. (1994) *Behavior and mind: The roots of modern psychology.* Oxford University Press. [HL, aHR]
(1995a) Self-control: Beyond commitment. *Behavioral and Brain Sciences* 18:109–59. [HL, aHR, DR]
(1995b) The value of temporal patterns in behavior. *Current Directions in Psychological Science* 4:188–91. [arHR]
(1997) Four teleological theories of addiction. *Psychonomic Bulletin and Review* 4:462–73. [arHR]
(2000) *The science of self-control.* Harvard University Press. [EJF, aHR]
Rachlin, H., Battalio, R., Kagel, J. & Green, L. (1981) Maximization theory in behavioral psychology. *Behavioral and Brian Sciences* 4:371–417. [aHR]
Rachlin, H. & Green, L. (1972) Commitment, choice and self-control. *Journal of the Experimental Analysis of Behavior* 17:15–22. [EJF]
Rapoport, A. & Chammah, A. M. (1965) *Prisoner's dilemma.* University of Michigan Press. [aHR]
Rawls, J. (1971) *A theory of justice.* Harvard University Press. [PW]
Read, D. (2001) Intrapersonal dilemmas. *Human Relations* 54:1093–117. [DR]
Read, D., Loewenstein, G. & Rabin, M. (1999) Choice bracketing. *Journal of Risk and Uncertainty* 23:5–32. [DR]
Reis, H. T. & Judd, C. M. (2000) *Handbook of research methods in social and personality psychology.* Cambridge University Press. [CS]
Rescorla, R. A. & Solomon, R. L. (1967) Two-process learning theory: Relations between Pavlovian conditioning and instrumental learning. *Psychological Review* 74:151–82. [rHR]
Rheingold, H. & Hay, D. (1980) Prosocial behavior of the very young. In: *Morality as a biological phenomenon,* ed. G. S. Stent. University of California Press. [RAH]
Robins, L. N. (1974) *The Vietnam drug user returns.* Special Action Office Monograph, Series A, No. 2, United States Government Printing Office. [aHR]
Roelofsma, P. H. M. P. & Read, D. (2000) Intransitive intertemporal choice. *Journal of Behavioral Decision Making* 13:161–77. [DR]
Rothbart, M. K., Ahadi, S. A. & Hershey, K. L. (1994) Temperament and social behavior in childhood. *Merrill-Palmer Quarterly* 40:21–39. [GC]
Rushton, J. P., Eysenck, H. J., Fulker, D. W., Neale, M. C. & Nias, D. K. B. (1986) Altruism and aggression: The heritability of individual differences. *Journal of Personality and Social Psychology* 50:1192–98. [JH]
Rushton, J. P., Fulker, D. W., Neale, M. C., Nias, D. K. B. & Eysenck, H. J. (1986) Altruism and aggression: The heritability of individual differences. *Journal of Personality and Social Psychology* 50:1192–98. [DJZ]
Ryle, G. (1949) *The concept of mind.* Hutchinson House/Barnes and Noble. [RB, rHR]
Sabini, J. (1995) *Social psychology.* W. W. Norton. [GFW]
Scheel, D. & Packer, C. (1991) Group hunting behaviour of lions: A search for cooperation. *Animal Behaviour* 41:697–709. [RS]
Schelling, T. (1971) The ecology of micromotives. *Public Interest* 25:61–98. [aHR]
(1984) *Choice and consequence.* Harvard University Press. [JS]
Schuster, R. (2001) An animal model of cooperating dyads: Methodological and theoretical issues. *Revista mexicana de análysis de la conducta* (*Mexican Journal of Behavior Analysis*) 27:165–200. [RS]
(in press) Cooperative coordination as a social behavior: Experiments with an animal model. *Human Nature.* [RS]
Schuster, R., Berger, B. D. & Swanson, H. H. (1993) Cooperative social coordination and aggression: II. Effects of sex and housing among three strains of intact laboratory rats differing in aggressiveness. *Quarterly Journal of Experimental Psychology* 46B:367–90.[RS]
Schuster, R., Rachlin, H., Rom, M. & Berger, B. D. (1982) An animal model of dyadic social interaction: Influence of isolation, competition and shock-induced aggression. *Aggressive Behavior* 8:116–21. [RS]
Sedikides, C. (1993) Assessment, enhancement, and verification determinants of the self-evaluation process. *Journal of Personality and Social Psychology* 65:317–38. [CS]
Sedikides, C. & Green, J. D. (2000) On the self-protective nature of inconsistency/negativity management: Using the person memory paradigm to examine self-referent memory. *Journal of Personality and Social Psychology* 79:906–22. [CS]
Sedikides, C. & Gregg, A. P. (in press) Self matters. In: *Sage handbook of social psychology,* ed. M. A. Hogg & J. Cooper. Sage. [CS]
Sedikides, C. & Skowronski, J. J. (1991) The law of cognitive structure activation. *Psychological Inquiry* 2:169–84. [CS]
Sedikides, C. & Strube, M. J. (1997) Self-evaluation: To thine own self be good, to thine own self be sure, to thine own self be true, and to thine own self be better. In: *Advances in experimental social psychology 29,* ed. M. P. Zanna. Academic Press. [CS]
Seinen, I. & Schramm, A. (2001) Social status and group norms: Indirect reciprocity in a helping experiment. *Discussion Paper* TI2001–003/1. Tinbergen Institute, Amsterdam. [CW]
Sethi, R. & Somanathan, E. (2001) Preference evolution and reciprocity. *Journal of Economic Theory* 97:273–97. [PD]
Shafir, E. & Tversky, A. (1992) Thinking through uncertainty: Nonconsequential reasoning and choice. *Cognitive Psychology* 24:449–74. [JB, JIK]

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press

Shaw, L. L., Batson, C. D. & Todd, R. M. (1994) Empathy avoidance: Forestalling feeling for another to escape the motivational consequences. *Journal of Personality and Social Psychology* 67:879–87. [GA]
Sidgwick, H. (1893) *Methods of ethics, 5th edition.* Macmillan. [PD]
Sidman, M. (1997) Equivalence relations. *Journal of the Experimental Analysis of Behavior* 68:258–66. [aHR]
Siegel, E. & Rachlin, H. (1996) Soft commitment: Self-control achieved by response persistence. *Journal of the Experimental of the Experimental Analysis of Behavior* 64:117–28. [aHR]
Sigmund, K., Hauert, C. & Nowak, M. A. (2001) Reward and punishment. *Proceedings of the National Academy of Sciences USA* 98:10757–62. [CW]
Silverstein, A., Cross, D., Brown, J. & Rachlin, H. (1998) Prior experience and patterning in a prisoner's dilemma game. *Journal of Behavioral Decision Making* 11:123–38. [aHR]
Simon, H. A. (1993) Altruism and economics. *American Economic Review* 83(2):156–61. [HG]
Skinner, B. F. (1938) *The behavior of organisms: An experimental analysis.* Appleton-Century-Crofts. [aHR]
(1953) *Science and human behavior.* Macmillan. [RS]
Smith, E. R. & Henry, S. (1996) An in-group becomes part of the self: Response time evidence. *Personality and Social Psychology Bulletin* 22:635–42. [CS]
Smuts, B. (1999) Multilevel selection, cooperation, and altruism. *Human Nature* 10(3):311–27. [JS]
Snyder, M., Clary, E. G. & Stukas, A. A. (2000) The functional approach to volunteerism. In: *Why we evaluate: Functions of attitudes,* ed. G. R. Maio & J. M. Olson. Erlbaum. [AAS]
Sobel, J. (2001) *Interdependent preferences and reciprocity.* [JS]
Sober, E. & Wilson, D. S. (1998) *Unto others: The evolution and psychology of unselfish behavior.* Harvard University Press. [HG, aHR, RS, JS, DSW]
Steels, L. (1999) The puzzle of language evolution. *Kognitionswissenschaft* 8(4):143–50. http:www.csl.sony.fr/downloads/papers/1999/steels-kogwis1999.pdf [WZ]
Stipek, D., Recchia, S. & McClintic, S. (1992) Self-evaluation in young children. *Monograph of the Society for Research in Child Development* 57 (1, Serial No. 226). [ML]
Stout, R. (1996) *Things that happen because they should.* Oxford University Press. [aHR]
Tobin, H., Logue, A. W., Chelonis, J. J., Ackerman, K. T. & May, J. G., III. (1996) Self-control in the monkey (*Macaca fasicularis*). *Animal Learning and Behavior* 24:168–74. [RS]
Tooby, J. & Cosmides, L. (1992) The psychological foundations of culture. In: *The adapted mind: Evolutionary psychology and the generation of culture,* ed. J. H. Barkow, L. Cosmides & J. Tooby. Oxford University Press. [rHR, DSW]
(1996) Friendship and the banker's paradox: Other pathways to the evolution of adaptations for altruism. In: *Evolution of social behavior patterns in primates and man. Proceedings of The British Academy, vol. 88,* pp. 119–44, ed. W. G. Runciman, J. Maynard Smith & R. I. M. Dunbar. Oxford University Press. [aHR]
Trivers, R. L. (1971) The evolution of reciprocal altruism. *Quarterly Review of Biology* 46:35–57. [HG, DR, JS]
Turiel, E. (1998) The development of morality. In: *Handbook of child psychology: Social, emotional and personality development, vol. 3,* ed. W. Damon & N. Eisenberg. Wiley. [RAH]
Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D. & Wetherell, M. S. (1987) *Rediscovering the social group: A self-categorization theory.* Blackwell. [AAS]
Tversky, A. & Kahneman, D. (1981) The framing of decisions and the rationality if choice. *Science* 211:453–58. [aHR]
Vallacher, R. R. & Wegner, D. M. (1987) What do people think they're doing? Action identification and human behavior. *Psychological Review* 94:3–15. [AAS]
Van der Steen, W. J. & Ho, V. K. Y. (2001) Methods and morals in the life sciences: A guide for analyzing and writing texts. Praeger. [WJvdS]
Van Lange, P. A. M. (1999) The pursuit of joint outcomes and equality in outcomes: An integrative model of social value orientation. *Journal of Personality and Social Psychology* 77:337–49. [MP]
Wagstaff, G. F. (2001) *An integrated psychological and philosophical approach to justice: Equity and desert.* Edwin Mellen Press. [GFW]
Wallach, M. A. & Wallach, L. (1983) *Psychology's sanction for selfishness: The error of egoism in theory and therapy.* W. H. Freeman. [SK]
Wedekind, C. & Braithwaite, V. A. (submitted) The long-term benefits of human generosity on indirect reciprocity. [CW]
Wedekind, C. & Milinski, M. (1996) Human cooperation in the simultaneous and the alternating Prisoner's Dilemma: Pavlov versus generous tit-for-tat. *Proceedings of the National Academy of Sciences USA* 93:2686–89. [CW]
(2000) Cooperation through image scoring in humans. *Science* 288:850–52. [CW]
Weirich, P. (2001) *Decision space: Multidimensional utility analysis.* Cambridge University Press. [PW]
(forthcoming) Economic rationality. In: *Handbook of rationality,* ed. A. Mele & P. Rawling. Oxford University Press. [PW]
Wilson, E. O. (1975) *Sociobiology: The new synthesis.* Harvard University Press. [HM, DJZ]
Wilson, M. & Daly, M. (1997) Life expectancy, economic inequality, homicide, and reproductive timing in Chicago neighborhoods. *British Medical Journal* 314:1271–74. [DSW]
Youniss, J. & Yates, M. (1999) Youth service and moral civic identity: A case for everyday morality. *Educational Psychology Review* 11:361–76. [RAH]
Zahavi, A. (1995) Altruism as a handicap: The limitation of kin selection and reciprocity. *Journal of Avian Biology* 26:1–3. [CW]
Zizzo, D. J. (2000a) Money burning and stealing in the laboratory: How conflicting ideologies emerge. Department of Economics Discussion Paper No. 40, University of Oxford. http://www.economics.ox.ac.uk/Research/working_papers.htm [DJZ]
(2000b) Relativity-sensitive behaivour in economics. Unpublished doctoral dissertation, University of Oxford. [DJZ]
(in press) Empirical evidence on interdependent preferences: Nature or nurture? *Cambridge Journal of Economics.* [DJZ]
Zuidema, W. (2003) How the poverty of stimulus solves the poverty of stimulus. In: *Advances in Neural Information Processing Systems 15,* ed. S. Becker, S. Thrun, & K. Obermayer, MIT Press (forthcoming). http://arti.vub.ac.be/~jelle/research/zuidema02nips.pdf [WZ]
Zuidema, W. & Hogeweg, P. (2000) Selective advantages of syntactic language: A model study. In: *Proceedings of the 22nd Annual Meeting of the Cognitive Science Society,* pp. 577–82. Erlbaum. http://arti.vub.ac.be/~jelle/research/cogsci2000.pdf [WZ]

https://doi.org/10.1017/S0140525X0228005X Published online by Cambridge University Press