Evolutionary game theory and the rationalist/behaviorist controversy[1]

Lloyd J. Carr (latest: 12/2/12)
Philosophy Dept.
Rivier University
Website: http://www.rivier.edu/faculty/lcarr

1. Introduction

The rationalist-behaviorist debate arises at several key junctures within rational choice theory, indicating that it isn't a narrow or local problem within this theory, but rather a foundational problem with wide and deep philosophical roots. One place where we see this debate concerns a concept central to rational choice theory: preference. The issue is this: how should preference ordering principles be understood? Do they have the status of ideal rational norms we should (try to) conform to, or are they open to empirical testing and consequently subject to being adjusted to conform to the reality of human behavior? That is: do we put our effort into adjusting our decisions so as to satisfy these principles, as the rationalists argue, even though humans seem doomed to fail to live up to such norms (similar, for example, to our beliefs about moral norms such as the golden rule)? Or, do we take the fact that human decisions continually violate preference ordering principles as evidence that they are empirically incorrect, as behaviorists argue, and so should be adjusted to fit this fact?

The controversy within rational choice theory concerning optimizing (maximizing) verses satisficing methods of practical reasoning can be seen to harbor the same problem. Here the issue is: should optimizing be an ideal of practical reasoning that decision makers should try to achieve, even though most real agents most of the time seem to fail to do so, or should satisficing (bounded rationality) be the standard of practical reasoning given that it seems to describe better the way most real agents actually make choices? Let's briefly review these opposing positions before seeing how this controversy reappears in the area of potentially cooperative games.

The rationalist position is that both preference ordering principles, for example the principles of transitivity for preference and for indifference, and optimizing are *norms* of practical rationality that people should try to live up to in making decisions. They are ideals, discovered and justified on rational grounds independent of how people actually make decisions. As such, they help form the standards by which actual choices can rightfully be justified as rational or judged as not. If it turns out that most people most of the time actually make bad (i.e. irrational, by the standards in question) choices, the problem does not lie with the principles and standards that we depend on to discover this fact, the problem lies with the way people go about making their real decisions. That is, the practical reasoning leading to their choices is flawed, it violates one or more of these preference principles or fails to optimize (or fails some other rational norm). So, the rationalist argues, people should try to improve their reasoning to better conform to the methods and norms of rational choice. By analogy, the fact that most people have and will continue to tell lies is no reason to give up the moral ideal of honesty, for the latter is justified as a moral norm independently of human behavior.

The behaviorist position, on the other hand, is that the theory of rational choice, if it is to apply to real human decision making, must be based in reality, not in abstractions and unrealistic ideals. The principles of practical rationality, including such things as preference ordering rules and ideals of optimizing, have (or should have) empirical status, which means that they must be verified or falsified on the basis of observing how real people actually make decisions, not on how ideal rational agents would make decision. If we find that enough real people violate one or more of these preference principles in how they go about making decisions, or if we find that real people are in fact satisficers rather than optimizers, it is wrong to judge them irrational or poor decision-makers; rather, this shows that the principles and ideals in question are false and unrealistic, and must be changed to better fit human behavior, the ruling “facts on the ground.” The theory should fit the data, whereas rationalists wrongly believe the data should change to fit the theory.[2]

This controversy is not limited to the status of preference ordering principles and the issue of optimizing verses satisficing as methods of practical reasoning. It gains deeper and wider importance for understanding rational choice in light of two discoveries concerning potentially cooperative decisions. (1) From the analysis and evaluation of potentially cooperative games, it looks as if practical reasoning has broadly failed to justify cooperation as the rational choice in cases where it seems obvious that cooperation would yield the best outcome, the outcome having maximum goal achievement. The same methods and principles of practical reasoning that proved so powerful in arriving at rational choice solutions in competitive decision problems seem hopelessly inadequate when it comes to some non-zero sum games – namely, those games that are especially valuable for understanding a wide range of human cooperative decision making. We can see this by comparing five widely-studied games, after which we will discuss the second discovery: (2) a surprising alternative way of achieving cooperative solutions that does not depend on rationality at all but comes from biology

2. A failure of practical rationality?

In the game *harmony*, agents cooperate without any non-rational "outside assistance" in the form of affection or good will toward one another, or a third-party enforcer such as government, or as a matter religious or moral obligation. In *harmony* agents cooperate as the rational choice, arrived at by clear methods of practical reasoning (i.e. by dominance, by maximin reasoning, and by Nash equilibrium).[3] This is not what happens, however, in the following four potentially cooperative games. Two, the *strong clash of wills* and *chicken*, are similar in that both suffer from "too much." They each have one-too-many rational choice solutions; that is, they each have more than one equally rational Nash equilibrium points, and there seems to be no practical reasoning way for agents to cooperate one way or the other. As far as practical reasoning goes, there is no single rational choice solution. Instead, unsatisfying (at least from the rationalist point of view) recourses to non-rational "outside" decisional help must take place for one or both agents to maximally achieve the goal (e.g., flipping a coin in the *clash of wills* or skillful use of deception in *chicken*). In these two important games, then, practical reasoning leaves agents with an **equilibrium selection problem**, threatening the prospects for cooperation with a stalemate or deadlock.[4]

In contrast, the other two, the *stag hunt* and the *prisoner's dilemma*, are similar in that both seem to suffer from "too little." Each has a single rational choice solution by one or more standard methods of practical reasoning, but these solutions seem to fall so far short of what mutual cooperation would achieve of the goal as to be unacceptable to reasonable agents. Not only many rational choice theorists, but also the strong intuitions of non-experts, reject (or at least would like to reject) mutual defection, the rational choice in each case, as too sub-optimal, yielding only the third best (second worse) outcome out of four possibilities. Surely, they argue (and we all feel), humans can – and often do – do better than that. In these two important games, then, practical reasoning leaves agents with a **sub-optimal outcome problem**, and the benefits of cooperation are lost.[5]

In sum, except for harmony (and perhaps the special cases of the *weak clash of wills* and the *iterated weakened prisoner's dilemma*), practical reasoning and making rational choices appears to be an obstacle and not a path to human cooperation, at least for all those cases of possible human interaction that can be represented by these four games. The first discovery that adds fuel to the rationalist/behaviorist controversy, then, is this apparent failure of practical rationality to arrive at and justify the decision to cooperation in important potentially cooperative interactions.

3. Biological conflict and cooperation

(2) The second discovery happened in the field of biology. In the 1970's it was discovered that non-human creatures – plants, microbes, insects, and animals – seem to behave and interact in ways that can be interpreted and modeled as games. Now at first sight this might not appear to be very remarkable. Because there seems to be a lot of conflict and competition that takes place among non-human creatures, it would be perfectly natural, one might think, if competitive (zero sum) games are helpful in

understanding non-human interactive behavior. It seems like there ought to be a close connection between the struggle to survive (“survival of the fittest”) among living things and the idea of a “zero sum conflict.” But we must keep in mind that non-human creatures are not using methods or principles of practical reasoning and are not making rational choices in their competitive interactions. Something else must be going on that enables living things to compete with each other with enough success (i.e., goal achievement) to survive and prosper at least on the level of species if not always in the case of the individual organism. So: how could the competitive behavior of completely non-rational life forms ever be understood as zero sum games? Are dominant strategies a matter of instinct perhaps? Is maximin behavior genetically “programmed”? Are the brains of animals or the cells of microbes and plants somehow “hardwired” or “coded” for optimal mixed strategy behavior? The possibility of non-rational zero-sum games starts to appear quite remarkable. But notice how such questions, such possibilities, do not belong in the rationalist camp. They fall squarely in what we have called the behaviorist position; that is, these are empirical, naturalistic questions to be answered by experiments and observations, and are not to be answered by considering the decisions of ideal rational agents.

It gets better for the behaviorist position when we consider the non-competitive interactions that seem to take place among non-human living things. Just think of all the cooperation going on, say, within a family of wolves or chimpanzees, not to mention the amazing amount of cooperation to be found within a bee hive or an ant colony. There is even cooperation between members of widely different species: plants and insects, microbes and mammals, fish and reptiles, for example. It turns out that some of this non-human cooperative behavior can be analyzed in terms of potentially cooperative games. And when analyzed as games these non-human interactions appear to achieve and sustain patterns of mutual cooperation – the very thing that (except for harmony) practical reasoning failed to achieve in the case of humans! This is very surprising. Clearly, such non-human creatures are not using methods or principles of practical reasoning and are not making rational choices in their cooperative interactions (and good thing, one might add, in light of the first discovery above, for if their behavior depended on practical reasoning they would likely not be cooperating as much as they are observed to). So, how is all this cooperation happening? According to biology, and according to wide-spread popular opinion, the ways in which non-human living things behave and interact is the result of evolution over many generations. It would appear that evolution, a purely natural process of selection and not one that works by rational decisions, somehow has brought about patterns of cooperative behavior among living things (including, perhaps, the huge range of human cooperation) that clearly cannot be accounted for in terms of practical reasoning and principles of rational choice.

These two discoveries (that is: (1) the failure of practical reasoning to justify a single, optimal (cooperative) outcome as the rational choice in several important potentially cooperative games, and (2) evolution based cooperation in non-human "games") give great weight to the behaviorist position. This is a major challenge to the rationalist position. Perhaps the theory of rational choice not only fails to provide ideals and norms of practical rationality for important cooperative decisions, maybe it is a complete dead end as an analysis and justification of the entire range of human decision-making. Instead, perhaps the theory of evolution, appealing not to practical reasoning but only to natural biological processes and the (observed) behavior of living things – including humans – will provide a powerful understanding of all forms of human decision-making, including the all-important decision to cooperate. Let's look a bit more deeply into this version of the rationalist/behaviorist controversy to see how the behaviorist (that is, the evolutionary) account of cooperation works.

In what follows, (1) we will first set up a touchstone example of a non-human cooperative interaction analyzed as an evolutionary game. Next, (2) we will form a general picture of evolutionary game theory, focusing on those evolutionary mechanisms that appear to offer a plausible account of how such cooperative behavior could have evolved. Evolutionary game theory, as here presented, will represent the behaviorist alternative to rational choice theory, and the process of evolution (albeit simulated) will represent a naturalistic alternative to practical reasoning. Then (3) we will end this study by suggesting, on behalf of rational choice theory, a rationalist response to this broadly behaviorist rival theory.

4. Imaginary monkeys

We start with an example that we can easily relate to, but bear in mind at this stage that these ideas are meant to apply to the interactions of a wide variety of non-rational living things, including plants and microbes, and the fact that an animal with a degree of intelligence serves as our example will be irrelevant. Imagine, then, a species of monkey, ones with a lot of body hair. The hair collects dirt, easily gets matted, and makes a perfect environment for all kinds of infestations: mites, body lice, bugs, parasites, etc. It is unhealthy, a weakened condition, for any of these monkeys to have infested body hair. Let's suppose that there are two behaviors these monkeys engage in that are related to the condition of their hair. One is mating behavior: mating is less successful the more an individual monkey's hair is infested – the infested monkey, let's say, is less healthy and so less attractive to potential mates, and weakened to the point of disadvantage in competing for mates; consequently, it has a below average mating rate. A well-groomed monkey, on the other hand, has an easier time attracting mates and is stronger in competing for mates than an infested monkey, and consequently has an above average mating rate.

The other is grooming behavior: a monkey that is groomed has clean body hair and is in this regard healthy and strong, but an un-groomed monkey will quickly become infested. These monkeys, let's suppose, can't groom themselves very well; imagine that it takes two hands to un-matt body hair, it takes the ability to see the skin area closely to find the parasites, it takes the ability to groom the hair on hard-to-reach body parts, etc., none of which any of these monkeys can do for itself. But they can groom each other's body hair pretty thoroughly. It is important, here, not to think of grooming as a choice a monkey makes. Instead, think of grooming as a genetically determined behavior (if not directly, then perhaps as a learned or copied behavior for which these monkeys have a sufficient genetic basis). This means two things: (1) grooming is a behavior a monkey carrying certain genes automatically does to another monkey given the appropriate stimulus, or fails to do in the case of monkeys carrying the genes for non-grooming behavior, and (2) it's a behavior (grooming or not grooming, whichever it is) that gets passed on to all offspring inheriting those genes. For each monkey doing the grooming, however, this behavior is an investment of precious time that benefits another monkey, time that might have gone toward mating behavior, searching for food, resting, and other self-centered behaviors. The best case for each monkey, clearly, is to be groomed and not spend time grooming any other in return (i.e. getting a “free ride”). The worst case, equally clearly, is for a monkey to groom others and never get groomed in return (i.e. being a “sucker”).

On the basis of the information so far, we want to keep track of two things in this imaginary species of monkey: grooming and mating. Grooming (or not grooming) represents a fixed (directly or indirectly genetically determined) behavior that, by increasing (or decreasing) a monkey’s health, will increase (or decrease) mating opportunities for groomed (or un-groomed) monkeys. Mating opportunities in a group of monkeys, in turn, will result in increased (or decreased) numbers of offspring in the next generation carrying the grooming (or the non-grooming) genes. In this next generation of offspring there will be grooming and mating resulting in a new generation of offspring, and so on. As this cycle continues, you can see the potential for changes, for in each generation some monkeys will be leaving more, and some fewer, offspring to populate the next generation, and these offspring will carry the genes for grooming or for non-grooming behavior.

In any given generation of our imaginary species of monkey, then, each individual will either groom other monkeys or fail to do any grooming depending on its genetic makeup. This means that for each monkey interacting with another there are 4 grooming possibilities, from best to worse case:

1) Be groomed by other monkeys, but not spend time grooming any other monkey (= best case for mating; this monkey would have the non-grooming gene in a group of monkeys having the grooming gene).

2) Be groomed by others, and spend time grooming others in return (= 2nd best case for mating; this monkey would have the grooming gene in a group of monkeys likewise having the grooming gene).

3) Not be groomed by others, and not spend time grooming others (= 3rd best case for mating; this monkey would have the gene for non-grooming in a group of monkeys likewise having the non-grooming gene).

4) Not be groomed by others, and spend time grooming other monkeys (= worse case for mating; this monkey would have the grooming gene in a group of monkeys having the non-grooming gene).

Let’s put a “mating value” in terms of offspring on each of these 4 possibilities. (We’ll use values that are unrealistic but that will let us speed things up once evolution is brought into the picture). We’ll suppose that getting groomed is worth many mating opportunities the result of which is on average, say, 12 offspring in the next generation. And let’s put a cost for the time spent grooming others that could have been used mating; let’s say it costs a monkey on average 2 potential offspring to give up mating behavior and instead use the time to groom another. For an infested monkey, one that hasn’t been groomed, let’s imagine a significant decrease in mating opportunities; let’s say an un-groomed (infested) monkey on average produces only 1 offspring.

So, in line with the above 4 grooming possibilities we have:

1) the best case = full mating value of 12 offspring,

2) the 2nd best case = 10 offspring (12 offspring minus cost of 2 offspring),

3) the 3rd best case = 1 offspring, and

4) the worst case = 0 offspring (1 offspring minus grooming cost of 2 offspring = -1 which, in our imaginary example, amounts to 0).

By design, then, this example can now be set up as a potentially cooperative grooming game with offspring as outcome values. Each of 2 interacting monkeys (monkey “row” and monkey “col”) will either groom the other or not, depending on the genetic makeup of each.

		Col:	
		C1: groom	C2: don’t groom
Row:	R1: groom	10, 10	0, 12
	R2: don’t groom	12, 0	1, 1

We can easily recognize that this is a prisoner's dilemma, except that in this case neither monkey has a choice but is genetically "hardwired" – directly or indirectly via learning/copying ability – to one strategy or the other. The grooming gene represents the cooperative option, while the non-grooming gene stands for the defection option. If our imaginary monkeys had a choice and could use practical reasoning, they would choose (R2,C2), and thereby fail to groom each other. This is just the sub-optimal problem the rational choice theorist struggles with, and the behaviorist looks to evolutionary game theory (and not to rational choice theory) for a solution.

Let's now add evolution to this picture (to be more exact: population dynamics – the change of a population over time relative to certain conditions and forces). We'll imaginatively let our population of monkeys interact with respect to grooming behaviors and mating outcomes, keeping to the mating values assigned to grooming and non-grooming behaviors, and watch what happens to the frequency of the grooming and the non-grooming genes as they go through several generations. This will give us the general idea of how natural selection might work to bring about, maintain, and even increase cooperation.

To keep our example simple, take a small starting group of 10 monkeys (say, 1 male and 9 females) interacting with each other. It is easy to see what would happen if all have the non-grooming gene. In the initial population no monkey would get groomed and no monkey would spend time grooming. Each would, therefore, produce (on average) only one offspring inheriting the non-grooming gene, resulting in a 1st generation of 10 new monkeys (the initial generation dies), none engaging in grooming behavior. The 2nd, 3rd,…,nth generations would be similar to the initial population, each monkey in each generation producing only 1 copy of itself. In such a group, evolution is completely conservative; everything stays the same except for changes in individual monkeys. In this imaginary group, mutual defection doesn't increase or decrease the spread of the gene for non-grooming.

Let's reverse things and start with an initial group of 10 monkeys (again, 1 male and 9 females) all having the grooming gene. What would happen? Each monkey would both groom and be groomed, and each would produce (on average) 10 offspring all having the grooming gene yielding a 1st generation of 100 groomers. The 2nd generation would number 1000 groomers, and by the 3rd generation there would be 10,000. In such a group, as long as they interacted with each other regarding grooming behavior, our imaginary grooming gene (that is: mutual cooperation) would remain 100% in frequency within a group exploding in population.

Next, let's see what happens in a possible mixed group. Suppose that by some genetic mutation, in a group of 10 non-grooming monkeys, 1 female monkey's non-grooming gene turns into a grooming gene which she will pass on to offspring. The population is "invaded" by a single mutant, a groomer, while the other 9 in the group retain the non-grooming gene. What would happen? In the initial generation, this mutant monkey would groom the other 9 but not be groomed in return (something like a "slave" monkey working away to make the other 9 healthy. And to keep the example simple, we'll skip the potential complication that each non-groomer monkey might have only a 1 in 9 chance of being groomed when interacting with others in the group, and assume that in such a small group of 10 the groomer gets to service all the others in the group.) This mutant monkey would leave no offspring and the grooming gene would be extinct within 1 generation. The other 9 monkeys would get the initial benefit of the free ride outcome for a total of 108 offspring, each carrying the non-grooming gene. Subsequent generations, however, would then be like this 1st generation of 108, the non-grooming gene no longer increasing or decreasing. So, a population of non-groomers turns out to be *stable* in the sense that it can't be invaded or taken over by a single mutant grooming monkey.

Reflect for a moment on what we've just seen. In only 3 generations of our imaginary species of monkey, if they form 3 small sub-groups or families of 10 interacting individuals each such that in 1 group no grooming gene exists, in a 2nd group only 1 mutant has the grooming gene, and in a 3rd group all 10 have the grooming gene, the grooming behavior (given our assigned values) would explode so powerfully that it would "pull" or "drive" the entire species in its direction. The 3 initial populations started off with a total of 19 non-grooming genes and 11 grooming genes, and ended up in 3 generations with a ratio of 118 non-grooming genes to 10,000 grooming genes. This required, as indicated, the unrealistic condition that each group be isolated into neighborhoods or territories so that its members only interact with others in their own sub-group regarding grooming behavior and not with members of another group. Nevertheless, the behaviorist's idea seems to show promise that the machinery of evolution (better: population dynamics), at least under certain conditions, has the power to "grow" or "spread" certain genes (or "tendencies") for cooperative behaviors and to "limit" other genes (or "tendencies") for defection, even if the latter don't ever completely go extinct. The principles of evolution embedded in our evolutionary game example appear, so far, to be a lot more favorable in allowing and spreading mutual cooperation (that is: addressing the sub-optimal outcome problem) than are the principles of rational choice and the practical reasoning that uses these principles in human games.

Now let's return to our imaginary "experiment." What would happen in a small group of 10 monkeys if they were a mixed population of, say, 9 carrying the grooming gene and 1 (again, because of a mutation) carried the non-grooming gene? Can groomers be invaded by a single non-groomer? Initially, all 10 would be groomed, but only 9 would produce on average 10 offspring each, the remaining non-groomer would spread that gene to 12 offspring (the "free ride" outcome being worth 2 more offspring than the mutual cooperation outcome). So, the 1st generation would contain 90 groomers and 12 non-groomers. How about the 2nd generation? Assuming that the 1st generation group is still small enough so that all members interact with each other concerning grooming, the 2nd generation would contain 900 groomers and 144 non-groomers. This 2nd generation would produce a 3rd generation of 9,000 groomers and 1,728 non-groomers. Clearly, the frequency (proportion) of the non-grooming gene is growing at a slightly faster rate than the frequency (proportion) of the grooming gene. If we project evolution at these rates out over many generations, non-groomers will catch up with and overtake the number of groomers. Clearly, a population of groomers can be successfully invaded by a single mutant non-groomer. What happens to mutual cooperation in such a scenario? Will the grooming gene go extinct?

If we relax the assumption that each monkey in such a large group interacts with every other member of the group, then there are 3 forces that work to stabilize the percentage of groomers and non-groomers: (1) the frequency with which groomers interact with other groomers, (2) the frequency with which non-groomers interact with other non-groomers, and (3) the frequency with which groomers and non-groomers interact. As the number of non-groomers swells there is a greater and greater chance that they will only interact with their own kind, resulting in the meager mutual defection outcome. This is a limiting factor on the spread of the non-grooming gene – good news for groomers. However, as the number of non-groomers increases there is also an increased chance that a groomer interacts with a non-groomer and thereby has its average number of offspring lowered by the sucker's payoff, and boosting the non-groomer's average number of offspring by the free ride payoff -- good news for non-groomers. We can expect, informally, that these 3 forces will reach a balance point, a ratio of groomers to non-groomers, that will adjust to each other over generations unless other evolutionary forces (for example, new kinds of mutants) arise to de-stabilize the population. The lesson the behaviorist wants to get across, then, is that mutual cooperation does not go extinct in such a mixed population, it holds its own so long as groomers can interact with their own kind with a minimal frequency.

As a final variation in this species of monkey, suppose we don't imagine a population that is a mixture of monkeys each of which only grooms or not, but rather imagine a population containing a new mutation: a mixed-strategy gene. Such monkeys are genetically determined, let's suppose, to behave as a groomer with other groomers and with each other, and behave as a non-groomer with other non-groomers. (We can call this new genetic mutation and new behavior strategy by several revealing names: "do unto others as they do unto you," or "eye for an eye, tooth for a tooth," or "you scratch my back and I'll scratch yours," or "tit-for-tat," or "one good turn deserves another," or "when in Rome do as the Romans," or "reciprocate in kind," or even "blending in with the crowd.") We can see that such mutants reproductively gain the full benefit of mutual cooperation with each other. And, while it's true that they never exploit a groomer and so miss out on the free-ride payoff, it is much more important to consider that they avoid being exploited and the sucker's payoff of 0 offspring. A genetic mutation such as this, once established, would rapidly spread and sustain mutual cooperation.

At this point, even though the example we have been working with is very artificial and designed for a special purpose (and by no means represent the way real biological evolution takes place), we should be at least a little impressed with the way some of the machinery of evolution might serve to generate mutual cooperation and work to solve the sub-optimal outcome problem. The behaviorist's use of "evolution" (i.e. mutation plus population dynamics) to arrive at a cooperative solution to important non-zero sum games – games that proved to be stumbling blocks for rational choice theory – looks in comparison to be quite fruitful. As in the case of the spread of a cooperative grooming mutation in the *prisoner's dilemma* monkey example we have been considering, we can ask: might the evolution of trust outweigh the rational pull risk has towards mutual defection in the *stag hunt* and lead to the spread of mutual cooperation, or would evolution still favor mutual defection? Could the evolution of a "confrontation-avoidance" or a "giving-in" strategy in some populations bring about a solution of the equilibrium selection problem in the *clash of wills*, or would evolution be unable to weed out this problem? How about *chicken*: would the evolution of certain kinds of deception strategies or "bluff genes" solve this game's equilibrium selection problem, or would the problem prove too stubborn for evolution to solve? The behaviorist research program, then, is to use evolutionary mechanisms (rather than rational choice) within game theory to explore how optimal single-equilibrium cooperation comes about – as it seems to do within and among many kinds of living things, including humans.

5. Evolutionary game theory

We'll now turn to forming a general idea of **evolutionary game theory**. You can see from our example of the grooming monkeys that this theory is a blend of some key ideas from the theory of evolution and from rational choice theory. Let's start by briefly reviewing how biological evolution takes place according to the theory of evolution. (Of necessity, this will be a very simple overview of standard Darwinian evolution, but it will give us what we need.)[6] Specifically, there are three central mechanisms to biological evolution that play a role in evolutionary game theory: variation, selection, and replication.

First, to have biological evolution there must be *variety*, some degree of difference, among individuals. Whether it is plants, microbes, insects, or animals, not all individuals within a population can be exactly identical to one another in their traits and behaviors, for then there would be no possibility for change to take place from generation to generation. All living things reproduce, and it turns out that the methods of reproduction yield offspring that are in fact more or less different from each other, and therefore more or less different from the parent generation. How might reproduction let diversity happen? Let's assume for the sake of simplicity that many of a creature's traits and behaviors are directly or indirectly determined to a significant degree by the creature's genes. One way to have diversity is for such genetic material to be broken down and variously combined with other genetic material, as in sexual reproduction where a male and a female each contributes half their genes to make an offspring. Or, perhaps such genetic material is fragile enough to undergo random mutations sufficient to make an individual more or less unique. Whatever the specific genetic mechanism, when reproduction results in genetic differences among offspring this will yield unique individuals; that is, they differ to some "noticeable" (by the environment) degree from each other in their traits and behaviors.

Second, to have biological evolution there must be a *selection* mechanism. In biological evolution, natural selection – sometimes popularly called "survival of the fittest" – is the mechanism. Contrast this with breeding livestock or pets for certain desirable traits in which artificial (not natural) selection by breeders is the mechanism. Natural selection works on the diverse individuals that make up populations. An organism's environment will contain dangers and opportunities, threats and comforts, conflicts and benefits. Each living thing is subject to a variety of environmental pressures, for example predators, climate events, diseases, and rivals for the resources that it needs to live. Biologically, "fitness" is measured by relative reproductive success; the more surviving offspring a living thing produces (relative to the number of surviving offspring other competing living things produce), the more fit it is. And vice-versa: the more fit a living thing is, the greater its expected number of surviving offspring. Natural selection, then, means that some individuals don't reproduce at all, or on average don't get to produce many offspring, because the individuals are formed in such a way and behave in such a way that they can't survive the environment they live in. Natural selection also means that other individuals are "favored." They behave in such a way that they survive the pressures in their environment; that is, they produce average or above average numbers of offspring that in turn reproduce.

It is important to note that natural selection, unlike artificial selection in livestock or pet breeding programs, is a “dumb” or (better!) a “blind” natural process. The environment is not trying to select certain individuals to ‘reward” with successful reproduction rates, and it is not trying to kill off other individuals as unfit to live, as if there were a plan in place. It is better to think of natural selection as a complex series of independent events that happen to coincide. Here is an imaginary example. Suppose the earth wobbles in its axis and as a result there is an ice age. Meanwhile a little creature purely by chance undergoes a genetic mutation that makes its body temperature go from 95 to 96 F^0. The wobble and the ice age “know” nothing of this little creature’s genetic mutation, and the genetic mutation “knows” nothing of the wobble and the resulting ice age. Yet, as the climate gradually becomes colder and the environment changes, the little creature’s long line of genetic offspring have a slightly easier time keeping warm and as a result can reproduce slightly more offspring than it could have, had the original mutation not happened. Its lineage survives; it got a lucky break. Suppose however the wobble resulted in the earth heating up. The little mutant’s long line of offspring metabolizing at 96 F^0 now have a slightly harder time keeping cool and consequently can’t reproduce as successfully. After many generations of below average reproductive rates the population carrying this mutation goes extinct. There is no plan or design in this example of natural selection; the wobble is an astronomical event that happens independently of the mutation, and the mutation is a glitch, an “error,” in the little creature’s genetic code that happened in complete isolation from what’s going on astronomically with the earth’s tilt. But the two events end up coinciding in a way that the mutation resulting in a 96 F^0 body temperature is “naturally selected” by the new environment for reproductive success and survival (if an ice age) or for reproductive failure and extinction (if global warming).

Third, to have biological evolution there must be *replication*. Each living thing possesses a variety of traits and behaviors. In reproduction, there must be a way for some traits and some behaviors to be passed on to its offspring; that is, traits and behaviors are replicated in offspring, its offspring inherit some of them. Of course, a living thing can’t pass all its traits and behaviors on to its entire offspring such that there is no difference between parent and offspring, or the first requirement, diversity, will not happen; all offspring would be identical to the parent and therefore identical to each other and evolution by natural selection couldn't happen as it does. How might replication happen? Let's continue to suppose that some of a living thing's traits and behaviors are to some degree genetically determined; that is, suppose that a creature's genes cause certain traits and certain behaviors (if not always, then with sufficient probability). Now if reproduction took place in such a way that a living thing could pass some of its genetic material directly to its offspring, including genetic material resulting from recombination and mutation, then the parents' traits and behaviors caused by these genes will be replicated in the offspring. (If not by direct genetic determination, the cause might work indirectly by way of a genetic base for, say, a mimicking or learning instinct by which replication of behavior takes place.) This, as we know, is the case. If a female contributes only half its genes to an offspring, then each of its genes will require the female to produces 2 offspring (on average) to replicate. But whatever the specific genetic mechanism, reproduction involves the replication of genetic material, and this means the replication of traits and behaviors in at least some offspring; they directly or indirectly inherit them from the earlier generations to which they are genetically related.

Biological evolution, of course, standardly takes place in small units and gradually over many, many generations. For our purposes it can be conveniently summed up as (1) the replication of genetically based traits and behaviors, (2) naturally selected by the environment, (3) from a pool of diverse organisms.[7] Some creatures reproduce more offspring and others less because of their inherited traits and behaviors, given various environmental pressures. Their offspring will inherit some of these same traits and behaviors and will likewise, on average and over time, reproduce more or less successfully. Thus, some populations grow (again, popularly called "survival of the fittest") while others shrink in number, and still others shrink to the point of becoming extinct. As environmental pressures change, different genetic combinations (that is, different traits and behaviors) will be "favored" and former successful populations will become extinct. All the small variations that are reproductively successful accumulate over eons giving rise to populations (species) that are very different in their traits and behaviors than the earlier populations (species) from which they evolved. This then is the standard, widely accepted biological position: all life on earth is related. From single cell creatures to the most complex multi-cell organisms, the huge variety of traits and behaviors that living things – both extinct as well as surviving – have ever possessed, results from evolution.

Let's now see how this picture of biological evolution is used in the theory of choice, primarily within the framework of games, so that the engine of evolution and not practical reasoning can be thought of as "making choices" and generating certain outcomes. We'll start with the following translations or redefinitions of ideas central to rational choice theory into evolutionary ideas.

1. Instead of defining a strategy as an option (as rational choice theory defines it), in evolutionary game theory a strategy is thought of as *a genetically based pattern of behavior*. Whatever a population or a species characteristically does when interacting with its environment and interacting with other living things is, in this sense, a strategy. For example, an animal species will have characteristic strategies for getting food, holding territory, mating, avoiding predators and other harms, and strategies for competing with rivals. A strategy is genetically based, so the three mechanisms of biological evolution will apply: (i) a strategy can mutate, (ii) it can be replicated in offspring, and (iii) it is subject to the pressures of natural selection.

A strategy might be a simple pattern or "rule" of behavior, such as: when there is food, consume as much as possible. It might be more complex, such as: when there is food, consume a small amount and hide the rest. Or it might be very complex, such as: when there is food and a rival, battle the rival for the food unless the rival is a large male and you have young offspring to raise. It might even be a mixture of behaviors, such as: one third of the time there is food, consume as much as possible and the rest of the time there is food hide it all. The important point, however, is that an organism does not "choose" its strategy (as if it were an option in a menu of options requiring a decision). Quite the contrary, its strategy is "chosen" for it by natural selection; it inherits it and is genetically determined, directly or indirectly, to behave in such-and-such a way in a given environment.

2. In rational choice theory, in order for there to be a choice there must be a menu of options (at least two) and, unless they all tie as equally rational choices, there must be some options that are not chosen, or would be irrational choices if chosen. This corresponds, in the theory of evolution, to genetic variation; there must be a pool of different individuals for natural selection to operate on by allowing some to have greater opportunities, and limiting the opportunities of others, to reproduce before they all die. In evolutionary game theory, there is an equivalent idea put in terms of a *variety (at least two) of interacting strategies*. Given two strategies, x and y, one might interact with its own kind (that is, x with another x, or y with another y), or two different strategies might interact (that is, x with y, or y with x). As two strategies interact, each becomes the natural selection environment of the other, each allowing or limiting the reproductive success of the other. In evolutionary game theory, "interact" means *repeated play of the game* in question (e.g., prisoner's dilemma, clash of wills, stag hunt, or chicken – to keep to the potentially cooperative games mentioned above). For instance, in our grooming monkey example as the grooming strategy interacts with itself (that is, two monkeys groom each other) each gives the other greater mating opportunities than the non-grooming strategy interacting with itself gives to each other.

3. Instead of defining an outcome or payoff as an option's amount or degree of expected goal achievement (represented by utility values in rational choice theory), in evolutionary game theory outcome is interpreted as *the degree of reproductive success of a strategy*. All living things can be taken to have one overall goal: to reproduce (from the gene's point of view: to replicate itself). Of course, this is not a conscious goal; plants, viruses, and single-celled creatures don't consciously set goals or make plans to reproduce. Rather, think of reproduction as itself genetically determined; a living thing's genes will drive it to reproduce as much as it can (within the bounds of the environmental pressures it is subject to). Given this overall goal of maximum replication, then, a strategy's outcome "utility" – its measure of goal achievement – will actually be its expected reproductive success. In other words, "outcome" is redefined as *the expected number of offspring*, on average, a creature using the strategy can be estimated to produce (relative to the number of offspring produced by a creature using an alternative strategy). Because reproductive success corresponds to fitness, in evolutionary game theory we should think of a strategy's outcome "utility" as representing its fitness, its expected offspring, in the process of

natural selection; that is, in the process of interacting with other strategies; and that is, repeatedly playing the game in question.

4. In rational choice theory, equilibrium is a central principle of practical reasoning for discovering and justifying a combination of choices as rational. Indeed (as described above in section 1), the equilibrium selection and sub-optimal equilibrium problems of important potentially cooperative games are taken to represent a deep challenge to, if not an outright failure of, rational choice theory, and to motivate the search for a naturalistic solution in terms of biological evolution. In evolutionary game theory, outcome equilibrium is interpreted as a combination of interacting strategies that is *evolutionarily stable* in a reproducing population. As rational choice equilibrium means that two choices are the best reply to each other (neither agent can gain more of the goal by switching to another option so long as the other doesn't switch), an *evolutionarily stable strategy* is, similarly, one that always generates more copies of itself than another (mutant) does. Let's make this idea clear. Take two strategies x and y. There are 4 possible interactions, each yielding a number of offspring (on average): (i) the offspring x produces interacting with another x, (ii) the offspring x produces interacting with a y, (iii) the offspring y produces interacting with another y, and finally (iv) the offspring y produces interacting with an x. We'll abbreviate these as: (i) O(x,x), (ii) O(x,y), (iii) O(y,y), and (iv) O(y,x). A strategy x is evolutionarily stable, then, providing another strategy y can't invade (that is, take over) a population of x's. This means that for x to be a stable strategy one of two things must happen from one generation to the next if y's enter a population of x's: either (i) O(x,x) > O(y,x), or (ii) O(x,x) = O(y,x) and O(x,y) > O(y,y). If we return to our grooming monkey example and let x = grooming behavior and y = non-grooming behavior, you will easily see by assigning the offspring numbers we used above that grooming is not a stable strategy relative to non-grooming (as we saw, a mutant non-groomer can invade and non-grooming can survive and spread among groomers). But if we switch it around and let x = non-grooming and y = grooming, then non-grooming is evolutionarily stable (we saw that a mutant groomer can't invade and prosper in a population of non-groomers, it goes extinct). Finally, if we let x represent the mixed strategy "reciprocate in kind" and y = non-grooming, it is easy to see that this mixed strategy is evolutionarily stable relative to y. (Take a moment and plug in the numbers to verify this for yourself.)

We are now ready to put these redefined ideas together into a general picture of **evolutionary game theory**: the behaviorist alternative to rational choice theory. Recall the plan. Practical reasoning, using the well-established methods of dominance, maximin reasoning, and Nash equilibrium, land us in the equilibrium selection and sub-optimal outcome problems, and strikingly fails to arrive at or justify mutual cooperation as the rational choice in several important potentially cooperative games. Perhaps, then, we can generate mutual cooperation, or a single equilibrium selection, from certain realistic initial conditions viewed as games by using mechanisms from the theory of evolution. In evolutionary game theory, then, the behaviorist starts with populations of cooperative and defecting strategies. These strategies are made to interact in certain environments (= repeatedly play each other in situations that can be structured as potentially cooperative games such as the *prisoner's dilemma* or the *stag hunt*). Their interactions generate copies of themselves (= they replicate according to assigned values). The behaviorist, then, using this theory tries to discover which strategies involving mutual cooperation (or involving a single equilibrium outcome) will prove to be evolutionarily stable over generations. For any that are discovered, the behaviorist needs to make a final (big!) step for the theory to have done its job: such strategies must be "linked" with (that is, shown similar to, if not actually identified with) real patterns of cooperation found in human interaction (thus, the above use of suggestive "human" names for the mixed strategy in our grooming monkey example). For any such optimum cooperative strategy, or any such single equilibrium outcome strategy, the "evolution" simulated within the experiment will have acted as if it made a choice, even though it is a purely natural, mechanical, blind process modeled on the way real biological evolution is thought to happen according to the standard theory of Darwinian evolution.

6. A possible rationalist response

It is time to bring rational choice theory (RCT) and the rationalist position back onto center stage now that an alternative theory (evolutionary game theory (EGT)) has been given a hearing and the behaviorist challenge to RCT in the area of potentially cooperative games has been describe. There are three related general philosophical principles to which the rationalist might appeal: (1) rationality implies norms, (2) is does not imply ought, and (3) ought implies can. How might the rationalist apply these principles to the rationalist-behaviorist controversy within RCT?

(1) Rationality is normative: this principle means that human reasoning, whether epistemic or practical, is an activity that is subject to guidance and control as it is being performed, and subject to evaluation when it is finished. (While some other forms of thinking might be, *reasoning* is neither a random nor an automatic mental activity taking place without the awareness of the person doing it; in addition, we accept the idea that reasoning can be done well or poorly.) If this is true, then there must be standards, rules, criteria – in a word: norms – of human rationality by which people can learn to reason well. Logic studies and discovers the norms and principles of epistemic rationality, while RCT studies and discovers the norms and principles of practical rationality. But how are norms discovered and studied? Clearly not by empirical and behavioral research alone, for without norms of rationality already assumed no such research would be able to distinguish good from bad patterns of reasoning. Rationality, then, must be a normative system able to be explored and studied by non-empirical methods of research, for example by idealizing instances of perfect rationality and discovering what principles of form, coherence, organization, or sequence might be involved. Thus, RCT can be justified, notwithstanding its shortcomings concerning some potentially cooperative games, independently of empirical/behavioral approaches to rationality.

(2) Is does not imply ought: this means that just because things *are* a certain way in reality (the “is” part of the principle) it doesn’t follow that that’s the way things *should* be (the “ought” part). For example, your car is making a noise; from this fact alone you would be wrong to conclude that your car should be making that noise and so nothing is wrong with it. You need to apply standards or norms of how a car *should* sound in order for you to judge whether your car’s noise is normal or an indication that something is wrong. In general, then, no amount of observation or description or simulation of reality by itself will ever give us norms that tell us the way things should be. EGT specifically, and behavioral choice theory in general, is primarily concerned with the origins, the development, and the way real people make real choices. This is valuable empirical research, and we look to it to tell us what is the case. But in performing

experiments with real and simulated animal and human choice behavior, EGT is not discovering norms for making good choices, nor is it justifying those real patterns of choice behavior as how people should be making choices – unless, that is, EGT makes a very controversial assumption: that nature is normative. In contrast, RCT is specifically concerned with discovering and developing norms and principles of practical rationality. In this regard, EGT is no substitute for RCT, as valuable as the former might be in explaining how certain patterns of choice behavior come to be. That RCT seems to have hit a road block when it comes to discovering norms of rational choice for cases of important potentially cooperative games should not blind us to its value in justifying decisions as rational choices in other cases. The rationalist position concerning RCT, then, seems in no way diminished by the interesting empirical research done within EGT.

(2) Ought implies can: this means that the way we say things *should* be (the "ought" part of this principle) must always be *possible* for them to be (the "can" part). If it is impossible for reality to be a certain way, then we can't reasonably claim that it *should* be that way. For example, it would be foolish to have a norm that says a car when driven should make no noise at all, for this is physically impossible and would make every and any (normal!) noise a sign that something is wrong with the vehicle. So, possibility is a necessary condition for normativity. Now when behaviorists argue (as we have made them argue) that RCT sets norms and ideals of practical reasoning that are *impossible* for real humans to achieve when making real decisions, standards according to which "perfectly good" decisions of real agents must be judged defective, if not irrational, this in effect challenges – it undermines – the overall normative status of RCT. An important question, therefore, should be asked: does "impossible" here really mean impossible, or does it mean "difficult" to do? It is one thing to claim that it is impossible for human decision makers to live up to the standards of RCT, but it is quite another thing to claim that it is hard, inconvenient, time-consuming, impractical, and even annoying to have to meet such standards. We can all sympathize with the latter claim, but such difficulties don't add up to strict practical impossibility. Quite the contrary, because behaviorists actually work out rational choices according to norms of RCT to solve realistic decision problems in an effort to demonstrate that this is *not* the way most decision-makers go about arriving at their choices, they have shown that it is indeed possible (thought often difficult, time-

consuming, and perhaps even annoying) to do so. Even though real decisions often take place under stressful conditions, deadlines, and other real world difficulties, it appears that the behaviorist charge of “impossibility” is overstated. And the normative status of RCT is not diminished one bit by the complaint of “difficult to do.” Similarly, in logic and in ethics

But our behaviorist will no doubt protest: what about the miserable failure of practical reasoning to justify mutual cooperation in such important human interactions as those represented by games like the stag hunt and the prisoner’s dilemma? Doesn’t this count against RCT? Doesn’t EGT show much more promise dealing with this problem? Fair enough, it does count against RCT. But this failure is limited to certain (admittedly) important games; it does not diminish the success in all the other kinds of decision in which rational choices are justified by clear principles, methods, and norms of practical reasoning.

7. Conclusion

The rationalist-behaviorist debate is a deep, complex, perennial philosophical controversy. It concerns the connection between the ideal and the real, between norms/standards/criteria and the reality that they “govern” or are used to evaluate, between a theory that contains abstract concepts and principles and the concrete facts to which the theory applies, between reason and behavior. We have here discussed one version of it concerning the normative status of the methods, principles, and standards of practical reasoning as found within RCT vis-à-vis results of behavioral studies of human decisions makers and evolutionary models of non-human interactions.
(This means that in our case we should all try to satisfy the norms and principles of RCT in our practical reasoning, and if we fail to do so the problem is not with the theory, it’s with the often careless way we real agents go about making our decisions.)

Footnotes:

1. This study began as an addendum to Chapters 11 and 12 of my text: *Making Good Choices: An Introduction to Practical Reasoning*. These chapters introduce cooperative (non-zero sum) games in general, and focuses on 4 – clash-of-wills, chicken, stag hunt, and prisoner's dilemma – whose analyses are controversial in a number of disciplines, e.g. philosophy, economics, psychology, and biology. My initial plan was to offer a brief presentation of evolutionary game theory's approach to these games – and through this to its approach to the problem of cooperation in general – as an example of the behaviorist (the empirical and naturalistic) alternative to rational choice theory. This "brief presentation," however, quickly ballooned to a size well beyond an addendum. Rather than fashion it as a separate chapter in a text primarily devoted to skill acquisition rather than to philosophical debate, the issue is here presented as an independent essay with references to parts of Chapters 11 and 12 as needed for background.

2. Resnick (1987), taking the rationalist point of view, offers an interesting understanding of this debate focusing on preference and preference ordering principles; see chapter 2.1. Mullen and Roth (1991) argue for a broad "interdependence" of the rationalist and the behaviorist (the "normative" and the "empirical" in their terminology) positions within the theory of rational choice; see p. 6ff. H. Simon is perhaps the leading proponent of the behaviorist position; see essay 8 "Alternative visions of rationality" in Moser (1990).

3. See Chapter 11, section 2 in *Making Good Choices* for the game *harmony*.

4. See Chapter 11, sections 3 and 4 in *Making Good Choices* for the equilibrium selection problem in the *clash-of-wills* and *chicken* games.

5. See Chapter 12, sections 1 and 2 in *Making Good Choices* for sub-optimal outcome problem in the *stag hunt* and *prisoner's dilemma* games.

6. This brief review of the standard Darwinian theory of evolution draws largely on Dawkins (1989), see especially Chapters 1 and 2.

7. Stating these 3 conditions this way highlights the tension between conditions 1 and 3; namely, to satisfy condition 1 there must be a degree of identity, “enough” similarity, between generations or there won’t be *replication*, there will only be causation from one generation to the next. But to satisfy condition 3 there must be a degree of differentiation, “enough” dissimilarity, between generations or there won’t be the *variation* in offspring from which nature selects.

Bibliography

For evolutionary game theory, Dawkins (1989) Chapter 5, and Smith (1978) are recommended to anyone with an interest in biology. Binmore’s “Evolutionary Ethics” in Leinfeller and Koehler (eds.) (1989) offers a clear overview of the evolutionary position/program. For specific games, see Skyrms’ (2004) exploration of the stag hunt from an evolutionary perspective, and Skyrms (1996) for an evolutionary approach to chicken, prisoner’s dilemma, and other cooperative decision problems. The prisoner’s dilemma has, by far, generated more literature in more disciplines than all the other games combined. Poundstone (1992) is a great read, not to be missed, and the historical background he provides is a real plus. If there is just one thing you’ll read on the prisoner’s dilemma, Axelrod (1984) gets my vote; it has become a classic. For an online source, Stanford University’s Encyclopedia of Philosophy (http://plato.stanford.edu/contents) has excellent current entries for game theory, evolutionary game theory, prisoner’s dilemma, etc. that focus primarily on philosophical issues rather than on technical (formal) problems or on experimental (behavioral) design and interpretation, containing extensive bibliographies and links to related online resources.

Axelrod, Robert (1984). *The Evolution of Cooperation*. Basic Books

Binmore, Ken (2005). *Natural Justice.* Oxford University Press

________ (1989) "Evolutionary Ethics." In Werner Leinfellner and Eckehart Koehler (eds.) (1998). *Game Theory, Experience, Rationality*. Kluwer

Carr, Lloyd J. ( ). *Making Good Choices: An Introduction to Practical Reasoning*. Website: www.rivier.edu/faculty/lcarr

Davis, Morton D. (1983). *Game Theory: A Nontechnical Introduction*. Dover

Dawkins, Richard (1989). *The Selfish Gene*. Oxford University Press

Moser, Paul (ed.) (1990). *Rationality in Action*. Cambridge University Press

Mullen, John D. and Roth, Byron M. (1991). *Decision Making: Its Logic and Practice*. Rowman & Littlefield

Poundstone, William (1992). *Prisoner's Dilemma*. Doubleday Anchor Books

Resnick, Michael D. (1987). *Choices: An Introduction to Decision Theory*. University of Minnesota Press

Selten, Reinhard (1998). "Game Theory, Experience, Rationality." In Werner Leinfellner and Eckehart Koehler (eds.) (1998). *Game Theory, Experience, Rationality*. Kluwer

Skyrms, Brian (1996). *Evolution of the Social Contract*. Cambridge University Press

_______ (2000). *Choice and Chance* (4th ed.). Wadsworth

_______ (2004). *The Stag Hunt and the Evolution of Social Structure*. Cambridge University Press

Smith, John Maynard (1978). "The evolution of behavior." *Scientific American*, vol. 239

Straffin, Philip D. (1993). *Game Theory and Strategy*. The Mathematical Association of America #36