18

Three Differences between Deliberation and Evolution

Elliott Sober

FROM: Modeling Rationality, Morality, and Evolution, Edited by Peter A. Danielson, Oxford University Press, 1998, Pages 408 to 422.

1. Introduction

The title of this paper may seem absurd. Why bother to write about differences between two processes that are so *obviously* different? Next thing you know, somebody will write a paper about the distinction between square roots and albatrosses.

Deliberation is something done by an organism that has a mind. The organism considers a range of alternative actions and chooses the one that seems to best advance the organism's goals. Evolution, on the other hand, involves a population of organisms, who may or may not have minds. When the population evolves by natural selection, the organisms display different characteristics; the trait that evolves is the one that best advances the organism's chance of surviving and reproducing.[1] Deliberation involves a change that occurs in an *individual*; evolution effects a change in the composition of a *population*, the individual members of which need never change their traits at all.

The outcome of evolution by natural selection is determined by the fitnesses of the traits that are present in the population. The outcome of rational deliberation is determined by the expected utilities of the actions that the agent considers performing. In natural selection, the fittest trait evolves; in rational deliberation, the act with the highest expected utility is the one that the agent chooses to perform. Of course, both these criteria involve simplifications; rightly or wrongly, we often decide to ignore factors that influence the two processes additional to the ones just cited.[2] But the fact remains that in considering the process of natural selection and the process of rational deliberation, we use similar rules of thumb.

This is not to deny that fitness and utility are different. Fitness is an objective property of an organism; it has nothing to do with what the organism thinks. Utility, on the other hand, is a subjective quantity, which reflects how much the agent likes or dislikes a possible outcome. Mindless organisms do not have (subjective) utilities, but it remains

true that some traits are better than others as far as survival and reproduction are concerned. And for organisms such as ourselves who do have preferences, utility can be and often is orthogonal to survival and reproductive success. Yet, in spite of these manifest differences, there seems to be an important isomorphism between the two processes. Selection and deliberation, understood in terms of the usual idealizations, are *optimizing processes*. Just as the (objectively) fittest trait evolves, so the (subjectively) best action gets performed.[3]

This isomorphism plays an important heuristic role in the way biologists think about the evolutionary process. When biologists consider which of an array of traits will evolve, they often ask themselves: If I were an organism in this population and I wanted to maximize fitness, which of these traits would I want to have? This way of thinking about evolution deploys what I will call *the heuristic of personification*:

> If natural selection controls which of traits $T, A_1, A_2, \ldots, A_n$ evolves in a given population, then T will evolve, rather than the alternatives listed, if and only if a rational agent who wanted to maximize fitness would choose T over $A_1, A_2, \ldots, A_n$.

Often this heuristic is harmless. When running speed evolves in a population of zebras, we may ask ourselves whether we would want to be fast or slow, if we were zebras who wanted to survive and reproduce. We get the right answer by this line of questioning; since *we* would want to be fast, it follows that *fast* is a fitter trait than *slow*, which means that selection will lead the population to a configuration in which all the organisms are fast and none are slow.

In this paper, I'll explore three contexts in which this heuristic yields the wrong answer. They all come from game theoretic discussion of altruism and the Prisoner's Dilemma. Whether it is applied to evolution or to rational deliberation, game theory models situations that involve *frequency dependence*. In the evolutionary case, how fit a trait is, and whether it is more or less fit than the alternatives, depends on the composition of the population (Maynard Smith 1982). In the case of rational deliberation, which act is best for the agent depends on what other actors are likely to do. As we now will see, frequency dependence can throw a monkey wrench into the convenient relationship between deliberation and evolution posited by the heuristic of personification.

2. Which Trait is Best for Me versus Which Trait Does Best on Average

Game theorists who discuss the Prisoner's Dilemma label the two actions "co-operate" and "defect." Evolutionists use the terms "altruism" and

"selfishness" instead. In a one-shot Prisoner's Dilemma, co-operating is bad for the actor, though it benefits the other player. This is the essence of evolutionary altruism, which is usually defined as an action that reduces the actor's fitness but benefits the other individual(s) in the group. For the sake of convenience, I'll use the evolutionary terminology throughout. But let us be clear that we are here describing an action's consequences for fitness or utility, not the psychological motives that produce it. Whether game theory is applied to evolution or to rational deliberation, behaviour and its attendant payoffs are ultimately what matter, not the proximate mechanisms (psychological or otherwise) that happen to produce the behaviour (Sober 1985).

The payoffs to row in a one-shot Prisoner's Dilemma may be represented as follows:

	Altruist	*Selfish*
Altruist	$x + b - c$	$x - c$
Selfish	$x + b$	x

If you are paired with someone who is an altruist, you receive the benefit *b* from this person's actions. If you yourself are an altruist, you pay a cost of *c* when you help the other person with whom you are paired.[4]

What should a rational deliberator do in this circumstance? Given the payoffs displayed, a simple dominance argument shows that the selfish behaviour is better. No matter what the other person does, you are better off by acting altruistically rather than by acting selfishly:

(1) A rational deliberator in the one-shot Prisoner's Dilemma should be selfish if and only if $c > 0$.

Now let us discuss the evolutionary case. When will selfishness have the higher average fitness in a population in which pairs of individuals are each playing a one-shot Prisoner's Dilemma? Let $Pr(A \mid S)$ represent the probability that one individual in a pair is altruist, conditional on the other individual's being selfish. This, and the other conditional probabilities $Pr(S \mid A)$, $Pr(A \mid A)$, and $Pr(S \mid S)$ allow one to describe whether individuals pair at random or tend to seek out individuals like (or unlike) themselves.[5] We now may represent the fitnesses of the two behaviours in this population of pairs of individuals as follows:

$$w(\text{Altruism}) = (x + b - c)Pr(A \mid A) + (x - c)Pr(S \mid A)$$

$$w(\text{Selfish}) = (x + b)Pr(A \mid S) + (x)Pr(S \mid S).$$

This simplifies to the following criterion for the evolution of altruism:

(2) In a population of pairs of individuals playing one-shot Prisoner's Dilemma, selfishness is the fitter trait if and only if $b[Pr(A \mid A) - Pr(A \mid S)] < c$.

Notice that (1) and (2) state different quantitative criteria. In particular, $c > 0$ is not sufficient for selfishness to be the fitter trait in (2). For example, suppose that altruists tend to pair with altruists and selfish individuals tend to pair with selfish individuals. If *like interacts with like*, then $Pr(A \mid A) - Pr(A \mid S) > 0$. In this case, altruism can be the fitter trait, even when $c > 0$.[6]

So it is quite possible for altruists to be fitter than selfish individuals, even though each individual would do better by being selfish than by being altruistic. The advice you would give to an individual, based on (1), is to be selfish. However, this does not accurately predict which trait will be fitter when you average over the entire population. The simple rule of thumb we saw before in the zebra example does not apply. We get the wrong answer if we use the heuristic of personification. *I would be better off being selfish in a one-shot Prisoner's Dilemma, but it does not follow that selfish individuals do better than altruists in a population of pairs of individuals playing a one-shot Prisoner's Dilemma.*[7]

There is a special evolutionary circumstance in which the heuristic of personification must deliver the right advice. This occurs when players are not correlated. If $Pr(A \mid A) - Pr(A \mid S) = 0$, then criterion (2) and criterion (1) are equivalent (Eells 1982; Sober 1993; Skyrms 1994). However, with positive correlation between interacting individuals, natural selection and rational deliberation can part ways.

One half of this conclusion is more controversial than the other. It is not controversial that the fitness of a trait is an average over all the individuals who have the trait. The fact that altruists are less fit than selfish individuals in every pair in which both traits are present does not tell you which trait is fitter overall. The reason is that this fact about mixed pairs fails to take into account what is true in pairs that are homogeneous.

Rather more controversial is what I have said in (1) about the decision problem. If the dominance principle is correct, selfishness is the rational act. But why buy the dominance principle? It is in conflict with some formulations of decision theory, as *aficionados* of the Newcomb problem well realize. I will not try to track this argument back to first principles, so perhaps my conclusion should be more conditional: if the dominance principle is a correct rule for rational deliberation, then the one-shot Prisoner's Dilemma provides a counter-example to the heuristic of personification.

3. Backwards Inductions in Iterated Prisoner's Dilemmas of Known Finite Length

What strategy should a player choose when playing an iterated Prisoner's Dilemma of known finite length? Luce and Raiffa (1957, pp. 94–102) proposed a backwards induction – an unraveling argument – to show that both players should choose to defect (to be selfish) on every move: On the last move, it makes sense for both players to defect. On the next-to-last move, there is no reason for them to co-operate, so both choose to defect then too. By working from the end to the beginning, the conclusion is drawn that the players should defect on every move. Game theorists have mostly accepted the correctness of this argument, and have moved on to consider games in which the game's length is not fixed beforehand, but is a matter of chance (cf. e.g., Axelrod 1984).

However, I want to argue that the backwards induction argument is invalid when formulated in a certain unconditional form. Without some set of qualifying assumptions, it reaches a conclusion about correct action that cannot be justified by strictly adhering to the policy of maximizing expected utility.

Let us begin by recalling an elementary fact about utility maximization. Consider a game in which there are two moves, X and Y. The payoffs to row are as follows:

	X	*Y*
X	9	2
Y	7	3

No dominance argument can establish whether X or Y is the better move. However, if one can assign a probability to what the column player will do, one then will be able to say whether X or Y is better. And even if no point value for this probability can be assigned, one could reach a decision provided one knew whether or not the probability that the column player will perform action X exceeds ⅓. However, with no information about this probability, no solution can be defended. After all, decision theory's criterion for action is maximizing *expected* utility; if neither action dominates the other, this theory cannot deliver a verdict when one is wholly ignorant of the probabilities involved.

Of course, in this circumstance, one could adopt a maximin strategy, and choose Y, since the worst-case outcome then is that one receives 3. Alternatively, one could adopt a maximax strategy, and choose X, since the best-case outcome then is that one receives 9. But no uncontroversial rational principle dictates whether one should maximin or maximax, so game theory has no definite recommendation to make here.

With that preamble, let us consider a three-round Prisoner's Dilemma, in which the pay-offs to row on each move are as follows:

	Altruist	*Selfish*
Altruist	2	0
Selfish	5	1

I want to consider four possible strategies that might be used in this three-round game. They are TFT3, TFT2, TFT1, and TFT0. TFT3 plays Tit-For-Tat on each of the three moves.[8] TFT2 plays Tit-For-Tat on the first two moves and then defects on the last one. TFT0 defects on all moves; it is ALLD by another name. The payoffs to row in a three-round game are as follows:

	TFT3	*TFT2*	*TFT1*	*TFT0*
TFT3	9	6	4	2
TFT2	11	7	4	2
TFT1	9	9	5	2
TFT0	7	7	7	3

What should a rational player do in this game? If the other player plays TFT3, the best response is TFT2. If the other player plays TFT2, the best response is TFT1. And if the other player plays TFT1, the best response is TFT0. All this is true, but does this show that the rational strategy is TFT0?

I would say no. There is no way to establish which strategy will maximize expected utility, for reasons captured by the simple game involving strategies X and Y. To be sure, if we consider just TFT3 and TFT2, TFT2 is the dominant strategy. And if we consider just TFT2 and TFT1, TFT1 is the dominant strategy. And if we consider just TFT1 and TFT0, TFT0 is the dominant strategy. All this is true *and irrelevant* to the game before us in which all four strategies need to be considered:

(3) In a three-round Prisoner's Dilemma in which each player may play TFT3, TFT2, TFT1, or TFT0, a rational deliberator cannot say which strategy is best without information concerning which strategy the other player will use.

In contrast, we do get a solution to the parallel evolutionary problem. If the population begins with TFT3 in the majority, that configura-

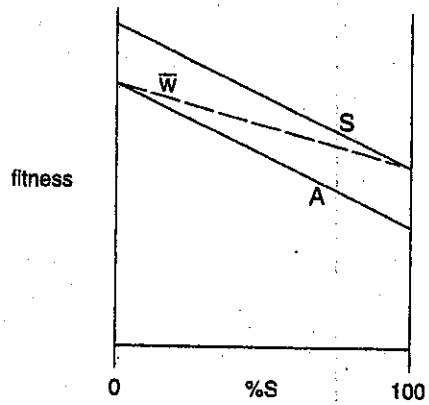

Figure 1

The upshot of natural selection in this case is straightforward, if the selection process takes place within the confines of a single group.[12] Selfishness is fitter than altruism at every population configuration, so the population evolves to a configuration of 100% selfishness. Note that this endpoint is the *minimum* value of $\bar{w}$. In this case, what evolves is not good for the group. Nor is it especially good for the individuals in the population, who would have been better off if everyone had been altruistic. We have here a "pessimistic" picture of natural selection; it reduces fitness, rather than increasing it:

(5) If selfishness is fitter than altruism at every frequency, then individual selection will drive the population to 100% selfishness.

Regardless of whether the altruism under discussion involves benefits donated exclusively to others or involves the creation of public goods, altruism cannot evolve within the confines of a single group as long as b and c are both positive.

What happens when we shift this problem into the context of rational deliberation? We imagine the agent is trying to maximize expected utility. What should the agent do? Although the evolutionary problem depicted in Figure 1 has an obvious solution, the present problem is more subtle. Figure 2 shows two ways in which the relationship between selfishness and altruism in Figure 1 might be "magnified," with *utility* substituted for *fitness*. Rather than imagining that the frequency of selfishness in the population can take any value between 0 and 1, we are thinking of the population as containing n individuals.

The agent must choose between a selfish and an altruistic act. There are already i ($0 \leq i < n$) selfish individuals in the population. If the

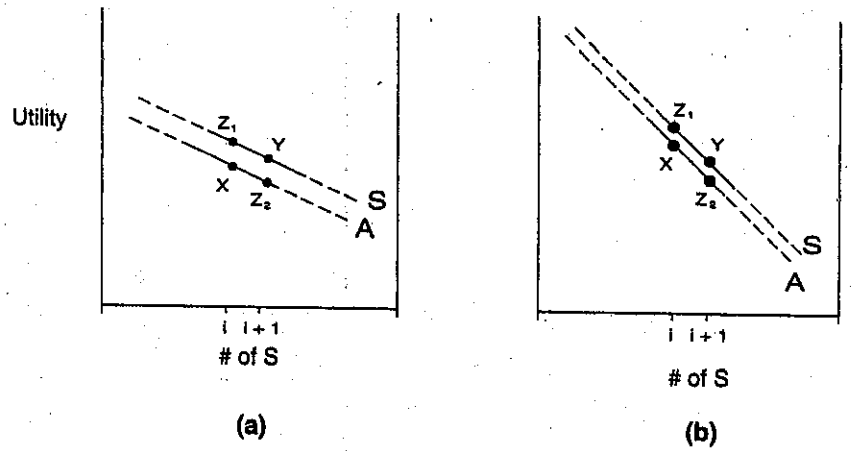

Figure 2

agent chooses selfishness, this number is augmented to $i + 1$; if the agent chooses altruism, the number of selfish individuals remains at i.

Which choice is better for the agent depends on whether the payoffs are as displayed in Figure (2a) or (2b). In (2a), point y has a higher utility than point x, so the agent should choose selfishness. However, in (2b), point y has a lower payoff than point x, so the agent should choose altruism:

(6) In an n-person Prisoner's Dilemma, a rational deliberator should behave selfishly if and only if the payoff to selfish individuals when there are $i + 1$ selfish individuals exceeds the payoff to altruists when there are i selfish individuals.

The contrast between propositions (5) and (6) reflects a difference between the process of natural selection and the process of rational deliberation. To see what happens in natural selection, you compare the fitnesses that alternative traits in the population *actually* have. However, to deliberate, you must think *counterfactually*. You must compare what would happen in one circumstance with what would happen in another. An evolutionist who looks at Figure (2a) or Figure (2b) and wishes to predict which trait will evolve will compare points x and z_1 or points y and z_2, depending on which correctly describes the population at hand. In contrast, a rational deliberator who wishes to choose an action will compare points x and y.

I began this section by contrasting altruistic donations that exclusively benefit others with altruistic donations that create public goods. Individual natural selection favours *neither* of these. But rational deliberators who aim to maximize their own fitness will sometimes choose

to create public goods even though they will never make altruistic donations that exclusively benefit others. They will decline to issue sentinel cries, but they may build stockades. It is in the context of public goods altruism that the heuristic of personification is found wanting.

In an evolutionary context, "selfishness" names the trait that evolves if individual selection is the only evolutionary force at work, while "altruism" is the name of the trait that evolves if group selection is the only cause of evolutionary change. When both types of selection act simultaneously, the outcome depends on which force is stronger. As noted earlier, the biological concepts of selfishness and altruism do not require that the organisms so labelled have minds. But now let us suppose that they do. Let us suppose that a selfish or an altruistic behaviour has evolved, and ask whether rational deliberation could be the psychological mechanism that causes individuals to produce these behaviours. We shall begin by considering the simple case of agents who want only to maximize their fitness.

If selfishness evolves and the fitnesses are those displayed in Figure (2a), this behaviour could be produced by individuals who are rational deliberators. However, if selfishness evolves and the fitnesses are those disposed in Figure (2b), matters are different. Rational deliberators will *not* choose to be selfish in this case.

A mirror-image pair of claims applies if altruism evolves. If the fitness relations are as depicted in Figure (2a), then the altruists who are the product of natural selection cannot be rational deliberators. However, if the fitness relations are as shown in Figure (2b), they can be rational deliberators. These conclusions are described in the following table. The cell entries answer the question: Could the trait that evolves be produced by a rational agent whose goal is to maximize fitness?

		What Trait Evolves?	
		Selfishness	*Altruism*
Fitnesses	*(2a)*	YES	NO
	(2b)	NO	YES

This set of conclusions does not mean that rational deliberation cannot evolve in two out of the four cases described. It means that if rational deliberation is to evolve in the upper-right- or lower-left-hand cases, then the agents' preferences must *not* be perfectly correlated with maximizing fitness.

It is commonly said as a criticism of sociobiology that human beings care about more than just staying alive and having babies. Usually, this

is taken to be a testimony to the influence of culture, which is supposed to displace biological imperatives from centre-stage. The present argument, however, shows that caring about things other than fitness can be a direct consequence of evolutionary processes. Rational deliberation may confer a biological benefit, but sometimes its evolution depends on having utilities that do *not* directly correspond to the fitness consequences of actions.

Some differences between the psychological concepts of egoism and altruism and the evolutionary concepts that go by the same names are obvious. The psychological concepts involve motives, whereas the evolutionary ones do not; the evolutionary concept concerns the fitness consequences of a behaviour, whereas the psychological categories make use of a much more general notion of welfare (Wilson 1991; Sober 1994b). However, the present discussion identifies a lack of correspondence between the two sets of categories that I think goes beyond the usual clarifying remarks. It is often supposed that a trait is evolutionarily selfish if a psychological egoist interested only in maximizing his own fitness would choose to have it. This application of the heuristic of personification is as natural as it is mistaken.[13]

5. Concluding Remarks

Darwin realized that his term "natural selection" was a metaphor drawn from the literal notion of rational deliberation. Nature is not a conscious agent, but in most instances it is harmless to think of natural selection *as if* a conscious agent were choosing traits on the basis of a fitness criterion. The imperfections of this analogy did not prevent the development of a literal theory of how natural selection works. In fact, only with this theory in hand can we return to the metaphor from which it derived and assess that metaphor's scope and limits. There is much to be said for the heuristic of personification. My goal here has been to suggest that we not overestimate its power.

I began this paper by asking why there was any point in writing it. I now can offer something by way of an answer. Because it is so natural and intuitive to use the heuristic of personification to think about the process of natural selection, it is well to have clearly in focus the fact that the heuristic can be misleading. So one reason it is worth detailing the differences between deliberation and natural selection is to hone one's understanding of the latter process. But there is a second type of illumination that this enquiry may be able to provide. To understand how the ability to deliberate evolved, it is important first to have a clear appreciation of what that ability involves. Popper (1972) is hardly the only person to have suggested that deliberation is a selection process in which our "theories die in our stead."[14] To understand what

deliberation is, it is important to see that it does not simply replicate the structure of the process of natural selection. Engaging in rational deliberation is a phenotypic trait for which an evolutionary story needs to be told. The phenotype is indeed a curious one. Why did evolution lead it to exhibit its present contours?

Acknowledgments

I am grateful to Ellery Eells, Larry Samuelson, Brian Skyrms, and David S. Wilson for comments on an earlier draft of this paper.

Notes

1 In putting matters this way, I am ignoring selection processes that can produce traits that are not good for the organism – namely, group selection on the one hand and true genic selection on the other. For an introduction to the issues involved here, see the chapter on the units of selection controversy in Sober (1993).

2 Adaptationists ignore non-selective processes because they think that they have a negligible effect on evolutionary outcomes. See Sober (1993) for discussion of the debate about adaptationism.

3 This does not mean that the processes must lead to an outcome that is optimal, for reasons that will become clear later on in the paper. I use "optimizing" to describe processes whose instantaneous laws of motion involve change in the direction of some "best" state; such processes need not, in the end, come to rest at some globally optimally state.

4 This table represents the "additive" case in which (1) altruism imposes the same cost on self, regardless of what the other player does, and (2) altruism confers the same benefit on the recipient, regardless of whether the recipient is altruistic or selfish. Non-additive payoffs can certainly be considered; the lessons I shall draw from the additive case would apply to them as well.

5 I assume here that organisms do not choose their phenotypes. They are either altruistic or selfish, and then the biology of their situation determines what the rules of pair formation are. For example, if individuals are reared in sibgroups, this will mean that altruists interact with altruists more than selfish individuals do. See Sober (1993), p. 114 for discussion.

A more complicated model would allow organisms to "choose" their phenotypes in the sense that an organism's phenotype would be conditional on some detectable environmental cue. For discussion of the issue of phenotypic plasticity, see Sober (1994a).

6 Nor is $c > 0$ necessary for selfishness to be the fitter trait. If c and b are *both* negative, the inequality stated in (2) may be satisfied. Of course, when $b < 0$, it is odd to use the term "altruism." Rather, a more general format is then being explored in which both positive and negative interventions of one organism in the affairs of another are being explored.

7 See the discussion of evidential and causal decision theories (as represented by the distinction by type-A and type-B beliefs) in Eells (1982).

8 Tit-for-Tat means that the player acts altruistically on the first move and then does on the next move whatever the partner did on the previous one.

9 As is usual in evolutionary game theory, one assumes the input of a few mutations in order to "test" the stability of monomorphic configurations.

10 In my opinion, the assumption that others are rational is nothing more than a useful heuristic that is often approximately correct. Davidson (1984) argues that the assumption of rationality is an *a priori* requirement if the beliefs and behaviours of others are to be interpreted; the empirical character of claims of rationality (and *ir*rationality) is defended in Kahnemann, Tversky, and Slovic (1982).

11 For the purposes of this example, I assume that the sentinel crow does not gain protection from the predator by sending the rest of the flock into a flurry of activity.

12 Not so if group selection occurs, but I am ignoring that possibility here.

13 This defect in the heuristic of personification has helped make hypotheses of group selection seem less worth considering than they really are. Suppose hunters must share their kill equally with everyone else in the group. If the hunter's share exceeds the cost of hunting, many biologists would conclude that it is in the self-interest of an individual to hunt and that the trait will therefore evolve by individual selection. This ignores the fact that free-riders do better than hunters in the same group. It takes a group selection process for this type of "self-interest" to evolve. See Wilson and Sober (1994) for further discussion.

14 Bradie (1994) provides a useful review of work in evolutionary epistemology that elaborates the idea that change in opinion can be modeled as a selection process.

References

Axelrod, R. (1984). *The Evolution of Cooperation.* New York: Basic Books.

Bicchieri, C. (1993). *Rationality and Coordination.* Cambridge: Cambridge University Press.

Bradie, M. (1994). Epistemology from an evolutionary point of view. In E. Sober (ed.), *Conceptual Issues in Evolutionary Biology* (Cambridge, MA: MIT Press).

Davidson, D. (1984). *Inquiries into Truth and Interpretation.* Oxford: Oxford University Press.

Eells, E. (1982). *Rational Decision and Causation.* Cambridge: Cambridge University Press.

Kahnemann, D., A. Tversky, and P. Slovic (1982). *Judgment under Uncertainty: Heuristics and Biases.* Cambridge: Cambridge University Press.

Luce, D., and H. Raiffa (1957). *Games and Decision.* New York: Wiley.

Maynard Smith, J. (1982). *Evolution and the Theory of Games.* Cambridge: Cambridge University Press.
Popper, K. (1972). *Objective Knowledge.* Oxford: Clarendon Press.
Skyrms, B. (1994). Darwin meets *The Logic of Decision*: Correlation in evolutionary game theory. *Philosophy of Science,* 61: 503–28.
Sober, E. (1985). Methodological behaviorism, evolution, and game theory. In J. Fetzer (ed.), *Sociobiology and Epistemology* (Dordrecht: Reidel).
——— (1992). Stable cooperation in iterated prisoners' dilemmas. *Economics and Philosophy,* 8: 127–39.
——— (1993). *Philosophy of Biology.* Boulder, CO: Westview Press.
——— (1994a). The adaptive advantage of learning and *a priori* prejudice. In *From a Biological Point of View: Essays in Evolutionary Philosophy* (Cambridge: Cambridge University Press).
——— (1994b). Did Evolution Make Us Psychological Egoists? In *From a Biological Point of View: Essays in Evolutionary Philosophy* (Cambridge: Cambridge University Press).
Wilson, D. (1991). On the relationship between evolutionary and psychological definitions of altruism and egoism. *Biology and Philosophy,* 7: 61–68.
Wilson, D., and Sober, E. (1994). Reintroducing group selection to the human behavioral sciences. *Behavioral and Brain Sciences,* 17: 585–654.

19

Evolutionary Models of Co-operative Mechanisms: Artificial Morality and Genetic Programming

Peter A. Danielson

1. Introduction

Social dilemmas, modeled by the Prisoner's Dilemma shown below, contrast rationality and morality. The moral appeal of C is obvious, as joint co-operation is mutually beneficial. But *D* (defection) is the rational choice because each does better by defecting.

	C	*D*
C	2,2	0,3
D	3,0	1,1

The contrast is not *complete* because agents that safely achieve joint co-operation do better than rational agents, and so steal some of the intuitive pragmatic appeal that the theory of rational choice attempts to formalize. On the other hand, the contrast is not *sharp* because while we know quite precisely what rational agents should do, the recommendations of instrumental moral theory are less well worked out. Both points lead us in the same direction; we would like to know if there are agents that can capture the benefits of mutual co-operation in these difficult social situations. *Artificial Morality* (Danielson 1992) argued for the existence of instrumentally robust moral agents by designing some. This Introduction sketches that argument, indicates some of its problems, and suggests why an evolutionary elaboration of the model might solve these problems. The remainder of the paper introduces Evolutionary Artificial Morality.

Artificial Morality

The simplest case in which a morally constrained agent might do better than a rational agent is the Extended version of the Prisoner's