

## **Darwinism in Contemporary Moral Philosophy and Social Theory**

### **1. Darwinism Characterized**

Philosophical Darwinism is a species of naturalism. Among philosophers, naturalism is widely treated as the view that contemporary scientific theory is the source of solutions to philosophical problems. Thus, naturalists look to the theory of natural selection as the primary source in coming to solve philosophical problems raised by human affairs. For it combines more strongly than any other theory relevance to human affairs and scientific warrant. Other theories, especially in physics and chemistry, are more strongly confirmed, especially because their more precise predictions can be tested in real time. But these theories have little to tell us about human conduct and institutions. On the other hand, actual and possible theories, in the social and behavioral sciences, may in the future have more tell us about humanity than Darwinian theory, but these theories do not as yet have anything like the degree of confirmation of Darwin's theory. Since Darwinism has important consequences for human affairs, the naturalist must look to Darwin's theory, above all others, in the search for philosophical understanding.

For present purposes, Darwinism is the thesis that the diversity, complexity and especially the adaptedness which organic phenomena manifest is solely the result of successive rounds of random variation and natural selection. Both the notions of 'selection' and 'random' need to be understood in special ways. Properly understood, Darwin's theory undermines the place of purposes in nature. Natural "selection" is a metaphor. There is no foresight in the way mutation and recombination produce variations on which the environment acts, filtering out those organisms which lack fitness minimal for survival long enough to reproduce themselves. One may hold that the theory of natural selection rids the world of purposes by showing that the apparent purposes manifest in adaptations are not real, as adaptation is the result of causes in which no purposes are represented. Or one may hold that Darwin's theory naturalizes purposes, showing how a naturalist can accept descriptions of nature in terms of purpose ("the heart beats in order to circulate the blood") as reflecting an evolutionary etiology (that brought about hearts). The first alternative banishes purpose from the universe; the second reduces it to mechanistic forces that naturalism countenances. Both threaten the "higher" purposes morality is traditionally supposed to serve. Most naturalists have long denied this threat, and have in fact held that Darwinism can illuminate and underwrite human values and moral commitments.

This chapter surveys contemporary strategies for proving a naturalistic understanding and vindication of morality, ethical norms, our conception of justice, and the cooperative human institutions which these norms and conceptions underlie.. We will see that while the prospects for vindication of moral claims as true or well-founded remain clouded, those for explaining the normative dimension of human affairs by appeal to Darwinism appear to be improving.

Moreover, the sort of evolutionary understanding of why human beings have been selected for being moral agents comes as close to a vindication of morality in human affairs as naturalism will allow.

## **2. Two Tasks for Darwinism in Ethics**

The ubiquitous human practice of making judgements of right and wrong, moral goodness and badness, imposing standards of fairness and justice, attributing moral duties and responsibility, and according autonomy, constitutes one of the most difficult challenges Naturalism faces. For the truth of statements expressing these judgements, standards and assumptions does not appear to be dependent on facts about the world accessible to scientific discovery. Indeed, these statements appear to report non-natural facts which cannot be accommodated in naturalism's metaphysics, nor are they amenable to evidential support by the employment of scientific methods that naturalism countenances. Naturalism has therefore called upon Darwinian considerations to reconcile our commitment to such normative judgements with a purely scientific world view.

There are broadly two "programs" which attempt to discharge this duty with which naturalism burdens the theory of natural selection: One of these programs seeks to underwrite either received moral judgements or some successor to them as true or correct in the light not of special normative truthmakers (this option being ruled out by naturalism) but in the light of the history of variation and selection through which they emerged. The second of these programs seeks to explain or explain away moral judgments as reflecting the operation of natural selection on hereditary variation in human activities. This second alternatives, naturalists will argue, is a new twist on the enterprise of analyzing the meaning of ethical claims which philosophers identify as metaethics. It is a new twist on traditional metaethics because it expresses naturalistic doubts about separating claims about meanings of ethical concepts from claims about the causes of ethical commitments expressed in these concepts. Thus, if naturalism can give an explanation of why we make the normative claims we do, it will claim to have provided as much of their meaning as can be provided. Call this project Darwinian metaethics. The first program is a compartment of substantive normative ethics, which identifies what is morally right and wrong, good and bad, just and unjust in terms of some evolutionary considerations. Call this project Darwinian morality

Both Darwinian metaethics and morality must take account of the peculiar fact about moral judgments that they are supposed to motivate us to do certain things, and to enjoin certain actions, not just as prudentially (in)advisable, in the light of our interests, but as right (or wrong) in themselves. This is a feature of normative claims which philosophers have dubbed "ethical

internalism”. If we accept that moral claims have this feature then they cannot, for instance, be merely injunctions of prudence, matters of merely instrumental ends-means rationality. On such an account of morality as instrumentally rational, if we do not accept the ends to which moral judgments report the means, we may disregard these judgements. But at least some moral judgements seem to make claims on us that are not merely instrumental, but categorical: “thou shalt not commit adultery”, not “If thou wish to avoid some bad end, or to attain some good one, thou should not commit adultery.” Darwinian metaethics may explain away the internalism of moral judgements as an illusion, though perhaps an adaptive illusion. Darwinian morality must harness it to evolutionary values as the motive for moral conduct. It will have to identify some naturalistically accepted normative grounds, some commitment to the ends or objectives such as species perpetuation or ecological preservation that make Darwinian morality internally motivating.

According to most philosophers the trouble with Darwinian morality has been well known for almost a century: As a philosophical project it rests on a mistake: the so-called “naturalistic fallacy”. In Principia Ethica<sup>i</sup> G.E. Moore offered the so-called “open question” argument against any identification of a normative property, like goodness, with a non-normative or “natural” property, like pleasure, or happiness or for that matter the survival of the individual or the species or for that matter the eco-system, planet or universe. Of any property, say an emotion such as love or a virtue such as heroism or a generalized feeling of pleasure, which is exemplified by someone, it may sensibly be asked whether the virtue or emotion or feeling is good . Accordingly, the identification of any such natural property with goodness cannot be correct. For if it were, the question “is Jones’ love for Smith, or for that matter of human-kind as a whole good?” would not be an “open question” to which a negative answer might be given. It would be a question like “Is Mr Jones’ mother a woman?” This question is not open to a negative answer. But all questions about whether some natural fact has a normative property are decidedly open questions, to which a negative answer may be intelligibly given. Accordingly Moore argued all attempts to naturalize the normative are fallacious. His open question argument defines the “naturalistic fallacy”. Its acceptance by philosophers has made Darwinian morality an unattractive option to most naturalists.<sup>ii</sup>

---

<sup>i</sup> Moore, G.E., *Principia Ethica*, London, Routledge, 1903, chapter one.

<sup>ii</sup> For further discussion see A. Rosenberg, “The biological justification of ethics: A best case scenario”, in *Social Policy and Philosophy*, 8:1 1990, pp. 86-101, reprinted in Rosenberg, A., *Darwinism in Philosophy, Social Science and Policy*, Cambridge, Cambridge University Press, 2000, pp. 118-136.

### 3. Darwinian Morality

The objection lodged against Darwinian morality may be illustrated by considering a philosophically sophisticated late 20<sup>th</sup> century version of this project: the attempt to establish certain normative principles as objective truths open to scientific discovery. The program took the name “moral realism” to echo the epistemological program of scientific realism, which argues that scientific theories about unobservable properties and entities should be treated as literally true descriptions of reality, and that the properties and entities to which they advert must exist in spite of the absence of direct empirical evidence for them. Similarly, latter day moral realism holds, we may know certain that certain favored moral properties—like goodness, in particular-- exist, and that some social arrangements have these moral properties, on the basis of scientific theory—in particular through considerations from a theory of the natural selection of moral norms. Peter Railton provides an excellent example of this school of Darwinism.<sup>iii</sup> Railton’s aim is to provide “descriptions and explanations of certain prominent features of the evolution of moral norms” (p.203) that will establish their naturalistic foundations. If Darwin’s name does not figure in his account it is because Railton recognizes that when it comes to the emergence of normatively right social institutions in the absence of ruling intentions to establish them, the only explanation can be Darwinian (see Railton, 1986, section III and IV, and especially footnote 21).<sup>iv</sup>

According to Railton the morally good reflects what it is rational to want, not from an individual point of view, but from “the social point of view” (p. 180). What is rational from the social point of view is what would be rationally approved of were the objective interests of all potentially affected individuals counted equally. Railton holds that social arrangements depart from rationality when they significantly dis-count the interests of particular groups. When this happens there is “potential for dissatisfaction and unrest” which reduces the viability, i.e the fitness, of these social arrangements and of the whole society so arranged. On Railton’s view, reduced viability of an arrangement—a norm, an institution, etc.--is reflected in “alienation, loss of morale, decline in the effectiveness of authority . . . potential for unrest,. . .a tendency towards religious or ideological doctrines, or towards certain forms of repressive apparatus,. . .” etc. (p. 192).

On the other hand, social arrangements which are more rational, i.e. tend more fully to be in the interests of all individuals in the society counted equally, will be selected for. That is, the societies bearing these traits will be more viable, presumably because arrangements that enhance

---

<sup>iii</sup> Peter Railton, “Moral Realism,” *Philosophical Review* 95 (1986): 163-207. Page references in this section of the chapter are to this to this paper.

<sup>iv</sup> See Railton, “Moral Realism”, section III and IV, and especially footnote 21.

equality of treatment are more adapted to the environments in which societies find themselves. This environment is not just the physical, geographical location of a society, it also includes societies with which it is in competition for scarce resources, and the society's environment also includes the fact that the individuals composing it have been selected for fitness- (and thus utility-) maximizing by natural selection. In the long run, just as biological natural selection winnows for those available traits that best "match" organisms' local environments, similarly, the struggle for survival among societies with varying moral traits will eventually winnow for those moral traits-- i.e. principles, norms, institutions-- that best match societies environment, and these, according to Railton, will invariably be ones that foster equality of various kinds. This will be so, since egalitarian arrangements most nearly fulfil individual people's objective--scientifically determinable--interests.

One objection to this approach is its commitment to natural selection of groups, whole societies, as opposed to individuals. What if in a society more viable than others because of its more extensively egalitarian norms, individuals arise who free-ride on and float these norms when they can. In this case, within group selection for immorality (i.e. inequality in treatment of others) may be stronger than between group selection for morality. In this case, evolution will not proceed in the direction of greater egalitarianism. Of this more in section 5 below. Meanwhile, Railton's account requires the truth of substantive claims that social arrangements which treat society's members in more nearly equal ways will be more adaptive under any conditions, for the society as a whole than those which entrain, enhance or preserve inequalities. Even if this claim were right, Railton's moral realism would still be subject to Moore's objection. There is no reason to think that the survival of any particular social group, individual, or Homo sapiens in general for that matter, is intrinsically good or morally required. There is in a naturalistic world view no scope for grounding such claims of intrinsic value.

Suppose it is retorted that Railton's thesis is analysis of what moral goodness consists in, not a justificatory endorsement of it. If this is true, moral realism does not accomplish what it has set out to do for Darwinian morality. For then Darwinism does not motivate any commitment to the moral principles it singles out as true. In effect, so understood, naturalisms like Railton's would deny or ignore the internal normativity of moral judgements and treat them as implicit claims about instrumental rationality, that is rules justified by the success of those (individuals or groups) who (or which) employ them in attaining their non-normative objectives (Railton, 1986, p. 200). Railton may well view his normative claims as merely instrumentally useful, and without internal moral force. He describes them as part of "the skeleton of a explanatory theory that uses the notion of what is ... rational from a social point of view...that parallels in an obvious way ... assessments of [instrumental] rationality ... in explanations of individual behaviors.") In fact,

Railton recommends we surrender “the idea that moral evaluations must have categorical force”[p. 204]. This denial of the internal normativity of moral judgments has the prospect of reducing Darwinian morality into some versions of Darwinian metaethics. For now it turns out that moral judgements are really just disguised claims about means-ends “instrumental rationality” to which we attribute some purely prudential normative force. Note that non-naturalistic forms of moral realism are not similarly threatened with such reduction to metaethics. For they claim that the normativity of moral judgment reflects some factual condition in the world which our moral detection apparatus enables us to identify. Thus, it has sometimes been claimed that we have direct intuition of the moral qualities of an act and these normative qualities motivate our approval or disapproval of the act in question. Naturalists of all stripes find such moral qualities either non-existent or unintelligible. It is certain there is no room for them in a naturalistic metaphysics.

The naturalists’ denial that a range of distinctive moral facts exist and make true moral judgements, together with the force of Moore’s diagnosis of a naturalistic fallacy, make Darwinian metaethics a far more attractive project for naturalists than Darwinian morality. Once we deny the existence of a separate range of moral facts to be learned by some sort of interaction either with nature or an with an abstract Platonic realm of values, metaethics becomes a matter of urgency. Metaethics is in large part the study of the nature and meaning of moral judgements. Without truth-makers for moral judgements, ethical claims may be threatened with meaninglessness. If they are meaningless we need at least an explanation of why we and all Homo sapiens make these apparent “judgements.” If they are not meaningless, but say, all false, we still need an explanation of why the error should persist time out of mind. If moral judgements are neither true nor false, but expressions of our emotions, we need an account of why this expression takes the form it does and why these expressions of our subjective states are coordinated in the way they are. And if moral judgements express the norms of conduct we embrace, we again need a theory to explain why we embrace these norms and not other ones. And in every case, an account needs to be provided of why we feel the commitment to an objective morality reflecting facts independent of us, and which motivate our conduct. About the only Naturalistic metaethical theory that can do any of these things is a Darwinian one.

#### **4. Darwinian Metaethics**

Most of the metaethical theories by Darwinian considerations belong to a species of metaethical theories collectively called “noncognitivist” owing to the fact that they share agreement that moral judgements are neither true nor false reports about the world—they have no propositional or “cognitive content”. Among the earliest noncognitivist theories was the “emotivist” doctrine advanced by A. J. Ayer and C. L. Stevenson and associated with Logical

Positivism.<sup>v</sup> This doctrine held that moral judgements expressed emotional states and attitudes of the utterer. Two virtues of this otherwise implausible theory are its ability to explain intransigent moral disagreement as expression of incompatible emotions, and its account of the apparent internalism of moral judgments: ethical judgments have motivational force derived from the emotional attitudes they express. But non-cognitivism will not account for the complex character of ethical reasoning characteristic of human life. More important, we often issue moral judgements on events distant in space/or time in such a cool and bloodless a way that they seem not to express emotions at all. Few latter day naturalists have been attracted by emotivism.

A more sophisticated version of non-cognitivist metaethics has been developed, with an eye to its place in a Darwinian framework, by Alan Gibbard.<sup>vi</sup> This widely discussed theory avoids many of the traditional objections to non-cognitivism, while making as strong a positive case for moral objectivity as naturalism will allow. Moreover, Gibbard's theory of the nature of moral judgments seeks to show at least how the emergence of morality might have reflected coordinated strategies that are adaptive for the individuals who employ them. As such Gibbard provides the philosophical foundation for an explosion of developments in evolutionary game theory and Darwinian political philosophy that we will explore below. Gibbard's theory is only one of a number of actual and possible Darwinian metaethics. The details of any such a theory will be important to philosophers anxious about the meaning of moral judgements. Biologists and others interested in the more general question of how moral judgements are possible within the Darwinian perspective will be more interested in how Gibbard develops the general strategy of a Darwinian metaethics. Before proceeding it is worth noting that the crucial difference between a moral realist like Railton's appeal to natural selection and Gibbard's is that the latter is not out to vindicate the norms which have in fact evolved as the morally right ones, only as the most adaptive ones.

“The key to human moral nature lies in coordination broadly considered” [p.26]

Organisms like Homo sapiens needed to coordinate their actions if they are to survive and flourish in competition with megafauna, and cooperative enterprises of proto-agriculture. The design-problem nature set for Homo sapiens of establishing and securing this coordination among them is accomplished in large measure by coordinated emotions (here Gibbard's noncognitivism shows its hand). Gibbard's objective is not to establish how institutions of morality or particular moral

---

<sup>v</sup>See Ayer, A.J., *Language, Truth and Logic*, London, Gollnaz, 1940, and Stevenson, C.L., *Ethics and Language*, New Haven, Yale University Press. 1944.

<sup>vi</sup> Gibbard, A., *Wise Choices, Apt Feelings*, Cambridge, Harvard University Press, 1992. Page references in this section of the chapter are to this to this paper.

judgements emerged or might have emerged as a result of random variation and natural selection, but rather to give an analysis of the meaning of moral judgements which, *inter alia*, make such a derivation possible.

A moral judgement is not the expression of an emotion, but a judgement of what sort of emotion or feeling it is rational to have; an emotion is a rational one to have if it is permissible in light of the norms one accepts. The capacity to accept norms depends on language, because language is required to coordinate several agents' norms in ways that are mutually fitness-enhancing. The environment of early man presumably selected for emotional propensities which enhanced coordination, and for linguistic potential that enable norms governing the display of these emotions to do so as well. Gibbard identifies resentment and anger, guilt and shame, as central moral feelings. Norms describing when it is appropriate to feel these emotions, are coordinated with one another so as to encourage or reestablish cooperative conduct among moral agents. Thus, what a person does is morally wrong if it makes sense, in the light of norms she accepts, for her to feel guilty about it, and it makes sense for others to feel resentment about her conduct in the light of their norms. A's Guilt meshes with B's anger, C's shame with D's disdain. If uncoordinated, these emotions can lead to escalating conflict; coordinated they make possible the acknowledgment of wrong-doing and reconciliation. What it makes sense to do, or to feel, in the light of norms a person accepts is what Gibbard defines as 'rational'. He rejects a purely instrumental account of rationality, both because of classical puzzle cases in the theory of decision, and more important, because 'rational' has an appraising or approving connotation (a reflection of the internalism of moral judgements), which analyses inspired by rational choice theory cannot capture. But to call an act or feeling rational is not to state a fact about it. It is to express ones acceptance of norms that permit the act or feeling.

Why is Gibbard's doctrine naturalistic and where is the special role for Darwinism in metaethics? The metaethics here is naturalistic because it requires no distinct range of independent moral facts to make true moral judgements. Our moral psychologies do not consist in systems which recognize and represent independent existing normative facts. Rather they are systems that coordinate what is in one agent's head with what is in other agents' heads. What is coordinated is the acceptance of norms in the light of which people's actions and emotions mesh to mutual advantage. The Darwinism emerges in the search for functions which these psychological mechanisms have, for a function, on Gibbard and most Darwinians' views, is as Larry Wright argued, <sup>vii</sup>is what emerges from an etiology of variation and selection.

The "design problem" which our hominid ancestors faced was how to establish and ensure

---

<sup>vii</sup> Wright, L., *Teleological Explanation*, Berkeley, University of California Press, 1976.

cooperation—acts of reciprocal altruism. Cooperation requires coordinated expectations of the sort that might emerge from a bargaining context. Gibbard suggests the emotions' function to enhance coordination must have been selected for by the same forces that made language adaptive for Homo sapiens. Gibbard's answer to the question of why did language emerge in the genus Homo has it that the capacity to be guided by words in action and emotion is indispensable to the acceptance of norms which produce cooperation. For such acceptance proceeds by discussion which tends towards consensus, consistency and similarity of motives.

But if moral judgement is not a matter of discerning truths but of expressing one's acceptance of norms that make sense of anger, resentment, guilt and shame, whence their apparent feeling of objectivity, of existence independent of us? Certainly not, Gibbard insists, from the existence of any Platonic range of moral facts or truths we can apprehend. The feeling of objectivity that accompanies these norms is a matter of how strongly we accept them. A norm is felt to be objective if one who holds it would consider it rational even if the holder did not accept the norm himself. And then there is a hierarchy of norms which agents accept. Judgements of objectivity will be a matter of derivation from these higher level norms. Gibbard is tempted by a parallel to the doctrine of secondary qualities. Color, it has long been argued by some empiricists, is a secondary property, that is a property of our experience, in us, and not in the objects we see as colored. But we mistakenly project this property on to objects in the world as an objective feature of it. Color is not a property of things out there in the world, but color attributions have considerable "objectivity". That is, a thing is red if and only if normal observers in normal conditions have red sensations when looking at it. Similarly, the objectivity of moral judgements is a matter of normal agents in normal circumstances accepting the same set of norms of anger, guilt, disdain and resentment.

So, in sum, a moral judgment is rational if it is in accordance with norms we embrace. These norms are ones selected for because they solve problems of cooperation, and their felt objectivity consists in their evolutionarily shaped ubiquity. The emotions that give these norms their internal motivational force are selected for because they coordinate and convey commitment to action in accordance with these norms.

Gibbard speculates that higher cognitive functions and language in Homo sapiens were selected for, owing to their role in the facilitation of social coordination. Language in particular enables agents to express norms and enhance their motivational power. The capacity to accept norms depends on language, and the discussions which language makes possible enhance mutual influence, consistency, and move people to act according to agreed-upon norms (whence their apparent internalist characters). The psychological state of accepting a norm, Gibbard holds, can at present only be identified as that psychological state which gives rise to the avowal of the norm

and to governance by it. Thus metaethics turns out to give empirical promissory notes about the origins of cognition and language that only biological anthropology and evolutionary psychology can cash in. It also requires demonstrations that the sort of cooperation which characterize morality is in fact adaptive.

It is worth noting that independent of Gibbard, developments in biological anthropology were in fact substantiating several factual presuppositions of his Darwinian metaethic. Only a sketch of these considerations can be given. To begin with, there is evidence that our hominid ancestors were originally solitary and highly competitive, not members of extended family troops with strong kinship relation. Cooperation can be expected to emerge among kin groups through the maximization of inclusive fitness (which calculates individual fitness as a function of total off-spring gene-copies an organism's genes leaves). But cooperation among originally solitary unrelated hominids requires communication of strategies. Independently, the shift from forest to savannah environments selects for the shift of vocalization from limbic to neocortical control (uncontrolled reflex vocalization in the vicinity of predators of the sort arboreal apes display is maladaptive on the savannah where there are no trees to climb). Whatever selected for the hominid shift to the savannah, selected also for neocortical control of vocalization that is necessary for language.<sup>viii</sup> The need for cooperation among unrelated individuals puts a further adaptive premium on language, as well as on the cognitive equipment required for recognizing cooperative strategies and non-cooperative ones. And this latter result is one evolutionary psychology has provided some evidence for.<sup>ix</sup> Finally, recent work on the theory of emotions provides further evidence that an adaptational account of anger especially as irrational precommitment to cooperative outcomes seems correct.<sup>x</sup>

### **5.Can Cooperation Evolve?**

Most of all, what this account of moral judgment requires is a great deal of detailed explanation of how natural selection could have brought about the norms of cooperation of which Gibbard claims our moral judgements express acceptance. Without the detail, such a Darwinian metaethic is little more than what S.J. Gould has stigmatized as a "just so" story. This need Gibbard's theory shares with any Darwinian metaethic.

It is just this sort of detail which evolutionary biology, game theory and political philosophy altogether provide, thus freeing a Darwinian metaethic from the charge of being

---

<sup>viii</sup> Maryanski, A, and Turner, J., *The Social Cage*, Palo Alto, Stanford University Press, 1992.

<sup>ix</sup> Barkow, R., Tooby, J., and Cosmides, L., *The Adapted Mind*, Oxford, Oxford University Press, 1992

<sup>x</sup> Griffiths, Paul, *What Emotions Really Are*, Chicago, University of Chicago Press, 1997.

merely a just so story. The substantiation of a naturalistic theory like Gibbard's, begins ironically with a major evolutionary problem. As E.O. Wilson wrote in *Sociobiology*, cooperation and altruism constitute "the central theoretical problem of sociobiology: how can altruism, which by definition reduces personal fitness, possibly evolve by natural selection." <sup>xi</sup> Natural selection relentlessly shapes organisms for individual fitness maximization: leaving the largest number of off-spring carrying the organism's genes. Call an act altruistic if it results in an increase in the fitness of another organism and a decrease in the fitness of the organism so acting. Now, the persistence of cooperation among organisms requires acts of reciprocated altruism so that the net-pay offs to mutual cooperators is greater than the rewards of mutual non-cooperation. Other things being equal, natural selection blocks the building up of altruism among randomly chosen organisms because altruistic acts offers opportunities to free-ride, to decline to reciprocate, and natural selection drives organisms to maximize fitness by taking every opportunity to free ride. Since altruism, and cooperation characterizes several infrahuman species, and all *Homo sapiens* societies, it appears that evolutionary theory has little to tell us about human conduct. This is what led Wilson to hold that the existence of altruism posed the gravest challenge to sociobiology.

Wrestling with this problem earlier in the 20<sup>th</sup> century some theorists, concluded that individual altruism is selected for because of the contribution it makes to the fitness of the group in which the individual finds itself. Group selection as an account of the evolution of altruism fell into great disfavor, however, for individual fitness maximization will swamp group selection. Suppose all members of a group are predisposed to cooperate, to engage in altruistic acts because their genes programmed them to act in this way. Suppose that through mutation, recombination, or immigration a new organism joins the group, lacking the gene for the propensity to cooperate. Instead it is genetically programmed to free-ride, cheat, slack off, shirk and take more than its share, whenever it can do so undetected. The free-rider has only to get away with free riding some of the time to have a higher fitness level than the rest of the group. Its off-spring will in turn bear the free-riding gene, and will take advantage of altruists as their immediate ancestor did. And so on, generation after generation, until genetically encoded reciprocal altruism has been extirpated from this group, which now of course has lower average fitness than it had when composed of altruists. In Maynard Smith's terms, genetically programmed altruism in a group is not an 'evolutionary stable strategy.' Wilson's problem of reconciling the ubiquity of human cooperation with natural selection remains.

However, if fitness is measured in term of the number of copies of itself a gene leaves, genetic selfishness must lead to one kind of organismal unselfishness. If an organism behaves

---

<sup>xi</sup> Wilson, E.O., *Sociobiology*, Cambridge, Harvard University Press, 1976.

altruistically towards its off-spring, enhancing their survival and reproductive opportunities, the result may be a decline in the altruistic parent's viability, but not a decline in its fitness or rather the fitness of its genes. This is kin-selection. But of course cooperation is far more widespread among Homo sapiens than selection for altruism towards kin. So, sociobiology is still faced with the problem that Wilson posed, of how altruism is possible. And Darwinian metaethics' claim that moral judgements are selected for coordinating behavior into cooperative exchanges remains ungrounded.

It was by exploring the economists puzzle of the prisoner's dilemma that evolutionary theorists were able to show how reciprocal altruism can be generated as the optimal strategy for fitness maximizing agents. Two agents, A and B are faced with mirror image choices of whether to cooperate with one another or to decline to do so, in other words, to defect. Payoffs to mutual defection are lower than payoffs to mutual cooperation, but defecting when the other party cooperates gives the highest pay-off. The Prisoner's dilemma is a dilemma because the rational strategy for each player--defection-- leads to an outcome neither prefers.

Something very much like the prisoner's dilemma situations occur frequently in real life. Every exchange across a store-counter, of money for goods represents what looks like such a problem: the customer would be best off if she grabbed the merchandise and left without paying, the sales-person would be best off if she could grab the money out of the customer's hands and with-hold the goods, the third best outcome for both is that the customer keeps the money while the sales-person keeps the good, and yet almost always, both attain the second most preferred outcome for both of exchanging good for money. The parties to this exchange are not irrational, so we need to explain why they attained the cooperative outcome, why the situation is not a prisoner's dilemma.

The reason is that the store-counter exchange problem is part of a larger game, the iterated or repeated prisoner's dilemma in which the two agents play the game again and again whenever the customer comes to the store. What is the best strategy in an iterated prisoner's dilemma? In computer simulations famously carried out by R. Axelrod, the optimal strategy in most iterated prisoners dilemma games of interest is one called "tit-for-tat": cooperate in game one, and then in each subsequent round do what the other player did in the previous round. In iterated prisoner's dilemmas among humans tit-for-tat is an effective strategy in part because it is clear--opponents don't need a great deal of cognitive skill to tell what strategy a player is using, it is nice--it starts out cooperatively, and it is forgiving--it retaliates only once for each attempt to free-ride on it.(It is important to bear in mind that tit-for-tat is an optimal strategy for maximizing the individual's pay-off (evolutionary or otherwise) only under certain conditions. See Axelrod, 1984.) When a group of players play tit-for-tat among themselves, the group and their strategy are not vulnerable

to invasion by players using an always-free-ride-never cooperate strategy. Players who do not cooperate will do better on the first round with each of the tit-for-tat-ers, but will do worse on each subsequent round, and in the long run will be eliminated. Tit-for-tat is an evolutionarily stable strategy: if it gets enough of a foothold in a group it will expand until it is the dominant strategy, and once it is established it cannot be overwhelmed by another strategy.

We can expect that nature's relentless exploration of the space of adaptive strategies in cooperative situations will uncover tit-for-tat, that long before the appearance of Homo sapiens, this strategy will have been written in the genes, and with it the genetic predispositions that make cooperation actual. By the time we get to human beings, these dispositions will include the cognitive ability to detect the strategies others use, enough language to coordinate them, and the emotions that mesh sufficiently to reinforce cooperation. In other words, Darwinian selection for fitness maximizers will have provided the biological details that a Darwinian metaethic such as Gibbard's requires.

It may even do more. Once we have recognition of partners, and memory about how they played in previous iterations, we may even have sufficient cognitive resources so that the one-shot prisoner's dilemma can be solved cooperatively. This, at any rate, appears to be the conclusion of Unto Others, Sober and Wilson's revisionist argument that group selection for cooperation is after all possible, and that the adaptational conditions under which it is actual, may well have obtained in hominid and human evolution.<sup>xii</sup> Their argument is disarmingly simple. Every one grants that kin-selection is not only possible but actual, as much evidence from infrahuman behavior demonstrates. It is also clear that kin-selection between one parent and one off-spring provides adaptational advantages to the two-membered "group" which they compose. In a one-shot prisoner's dilemma involving kin, both may be advantaged by cooperation regardless of the other's action, if the pay-off they are "designed" to maximize (reproductive fitness) satisfy the inequality,  $r > b/c$ , where  $r$  is the coefficient of relatedness ( $1/2$  in the case of off-spring and siblings,  $1$  in the case of identical twins,  $1/4$  in the case of cousins and nephews),  $b$  is the pay-off to mutual cooperation, and  $c$  is the cost of cooperation in the face of selfishness. If the group's fitness is a function of individual fitnesses, then groups of kin-related agents playing the cooperative (or "sucker's) strategy in a one-shot prisoner's dilemma will be fitter than groups composed of pairs of mutual free-riders playing the defector-strategy, or mixed groups of pairs of free-riders and suckers. The result generalizes to larger groups than pairs. But once players can recognize their degrees of relatedness, or for that matter what strategies they are genetically programmed to play in prisoner's dilemmas, they can preferentially aggregate into such fitter

---

<sup>xii</sup> Wilson, D.S., and Sober, E., *Unto Others*, Cambridge, Harvard University Press, 1998.

groups. When players seek out one on the basis of what strategy they play, the long term result is a “correlated equilibrium” of groups of cooperators only, the non-cooperating groups having been driven to extinction.

But recall the problem of invasion. Once these groups of cooperators get started, they are vulnerable to invasion or mutation that subverts from within, producing free-riders that take all other players in the group for suckers and increase in proportion from generation to generation until eventually selfishness becomes fixed in every erstwhile altruistic group. Wilson and Sober suggest that cooperating groups preserve themselves by means of secondary enforcement behaviors. Norms of cooperation are policed by norms of enforcement, and acting on these norms—shaming, reporting, confiscating-- are far less costly to the enforcing individuals than are the norms of cooperation they preserve from break-down. Wilson and Sober argue that unrelated human cooperative groups attain stable equilibria (ones that cannot be invaded) through the enforcement of social norms that lower the costs of cooperating and raise the costs of defecting.

## 6. Is Justice Selected For?

So, evolutionary game theory seems capable of solving Wilson’s problem of how to render human cooperation compatible with natural selection, and thus to help explain the emergence of the norms and emotions that underwrite them which Gibbard’s Darwinian metaethical project needs. But evolutionary game theory may even be able to go further and identify the content of some of these norms. In The Evolution of the Social Contract Brian Skyrms shows how a Darwinian process can result in the fixation among humans of the norm of justice as fair division. The key to this demonstration is again the evolution of a “correlated equilibrium” among like strategies through a mechanism of random variation and natural selection. Consider the problem of divide the cake: two players bid independently on the size of the cake they want. If the bids add up to more than the whole cake, neither gets any cake. Otherwise, they get what they bid. Most people bid  $\frac{1}{2}$ , of course. This outcome is an equilibrium: neither can do better, no matter what strategy the other employs. There are indefinitely many other Nash equilibria: for example, I bid 90 %, you bid 10%. But none of them is evolutionarily stable. A population whose members demand more than  $\frac{1}{2}$  or less than  $\frac{1}{2}$  of the cake will be invaded and swamped by pairs who demand  $\frac{1}{2}$ . Consider a bidding game in which random proportions of three strategies—bid  $\frac{1}{3}$ , bid  $\frac{2}{3}$ , bid  $\frac{1}{2}$  are represented to begin with. Skyrms has shown that in a computer simulation, in which strategies of lowest fitness are regularly removed, after 10,000 rounds, the fair division bid  $\frac{1}{2}$  is the sole remaining strategy 62 % of the time. When strategies correlate so that fair division plays against itself more frequently or with increasing frequency as the game proceeds, it almost always swamps any other strategy. Skyrms concludes,

“In a finite population, in a finite time, where there is some random element in evolution, some reasonable amount of divisibility of the good and some correlation, we can say that it is likely that something close to share and share alike should evolve in dividing the cake situations. This is, perhaps, a beginning of an explanation of the origin of our concept of justice.”<sup>xiii</sup>

Skyrms shows more than this: correlation among games enables selection of strategies for fitness to give rise to fair shares cooperation when cut-the-cake is played serially, instead of simultaneously (so that player one can demand more than  $\frac{1}{2}$ , forcing player 2 to choose between less than a fair share and nothing at all). Correlation among strategies in the defense of territories can give rise to private property as a cooperative solution to an adaptational problem. And finally, as we shall see, Skyrms sketches a way in which correlated strategies can give rise to meaning. One of Skyrms larger aims is to show that these happy Nash-equilibrium outcomes are attainable when the choice of individual strategies is governed by natural selection for optimal outcomes. None are attainable, when the choice of individual strategies is governed by considerations of economic rational choice seeking maximal pay-off.

But how can we be confident that correlation required for the evolution of cooperation arose? Here is the problem, illustrated by one of Skyrms’s results: In groups of kin-related individuals, for example vervet monkeys, signaling the presence of various threats--snake, leopard, eagle--can develop from correlated conventions about what noises consistently to make in the presence of different stimuli. Natural selection will prefer systems in which senders and receivers treat noises as bearing the same “news”. It will also select for altruistic employment of signals to warn kin, even at signaler’s expense. Note, this is a result that both Wilson and Sober, and Gibbard require. For norms of cooperation and enforcement require language; indeed, language is so important to the evolution of cooperation, that one might even argue that it emerged through selection for its impact on cooperation. But Skyrms’s model for the evolution of language presupposes strong correlation. In the case of vervets, it is provided by the kin-structure of aboral monkeys. Hominid evolution most probably proceeded however through solitary individuals dispersed from their kin and roaming a savanna alone.<sup>xiv</sup> The cooperation they needed to establish to survive could not presuppose kin-based correlation. There seems no other source of correlation. But without correlation there is no basis in evolutionary game theory to be confident that cooperation, or its semantic prerequisites will arise. There is thus more work to do in developing plausible models of the evolution of cooperation among fitness maximizers like us.

---

<sup>xiii</sup> Skyrms, B., *Evolution of the Social Contract*, Cambridge, Cambridge University Press, 1996, p. 21..

<sup>xiv</sup> See *Maryanski and Turner*, op. cit., note 6.

But what has been done in evolutionary game theory certainly has begun to provide the empirical foundations that a Darwinian metaethic requires for its claims about meaning and foundations of moral judgement. And in some attenuated sense, the result may even satisfy the hopes for a Darwinian morality. Without vindicating the internalism of moral judgements as reflecting objective demands on our conduct, the Darwinian metaethic approaches the goals that one tradition in ethics since Hobbes has set for itself: the task of showing that it is rational to be moral. Cooperation makes us each better off than we would be in a state of nature. But this outcome is not attainable as a bargain among rational agents; rather it is the result of natural selection operating over blind variation. This is almost, but not quite Darwinian morality.

## 7. Broader Implications of Darwinism for Social Theory

Well before the developments reported above, Darwinism was guiding a research program in the empirical social sciences that took the name of sociobiology. Laterally, some sociobiologists have substituted to name ‘evolutionary psychology’ for their program, in part to avoid the controversies which vexed sociobiology and in part to reflect a Darwinian commitment to individual selection, as opposed to group selection, as the force which shapes human behavior and social institutions. Sociobiology is controversial because it has been accused of adopting a Panglossian methodology that wrongly underwrites the status quo as inevitable and unchangeable. If social institutions—including the division of labor, both sexual and industrial, economic and racial inequality, vast power asymmetries, coercive violence, are the result of long term selection processes written into the genes, then they are no more subject to amelioration or change than eye-colour. And if this conclusion is derived from a method which simply finds some story about variation and selection that accords the status of an adaptation on extant institutions, without any empirical basis or even on the basis of a puerile misunderstanding of natural selection, then it will be no surprise that the research program is politically controversial. A sustained argument for this conclusion about Darwinism’s baleful influence in the social sciences is offered in Lewontin, Kamin, and Rose’s Not in Our Genes.<sup>xv</sup> 1984, especially chapter 9.]

Some work carried on under the banner of Darwinian sociobiology may certainly warrant such criticism.<sup>xvi</sup> But not all of it can be so criticized. Reviewing this work would take us too far afield, but at least some of the criticism of the research program of Darwinism in the social

---

<sup>xv</sup> Lewontin, R, Kamin, L., and Rose, S., *Not in Our Genes* , New York, Pantheon Books, 1984, especially chapter nine.

<sup>xvi</sup> See Kitcher, P., *Vaulting Ambition*, Cambridge, .MIT Press, 1989, for examples of such work and criticism of them.

sciences can be deflected by the developments in moral and political philosophy reported here. For if individual fitness maximization can result in the morality most of us share and in institutions of cooperation and justice, then it is not guilty of simply underwriting an unjust, non-egalitarian, sexist, racist status quo . There will have to be other factors at least in part responsible for these outcomes besides natural selection, and there will be environments—perhaps even attainable ones, in which natural selection will not inevitably lead to such nefarious outcomes.

Darwinian metaethics and evolutionary game theory have succeed, perhaps beyond the naturalist's hopes, in providing an account of how cooperative institutions are possible even where they not the result of conscious design or intention among any of their participants. Darwinism's success has also strongly encouraged other non-normative explanatory projects in the social sciences motivated by a search for stable equilibria that optimize some function without any participant intending or acting to attain such an outcome. In this respect Darwinism may in part vindicate the "invisible or "hidden" hand strategy of the approach of Adam Smith and his market-oriented followers in economics. Smith's laissez-faire economic theory implies that self-seeking in free markets will lead as if by a hidden hand to unintended outcomes which advantage all. It is now well known that this is not the case. Rational choice behavior among economic agents leads to non-optimum outcomes in many different circumstances: in the provision of public goods, or when large companies can make things more cheaply than small ones (what economists call "positive returns to scale"), or there is a small numbers of traders, asymmetries of information, when transaction costs are great, or there is a difference in the interests of principals and agents. These "market failures" have led critics of the market both to deny that economic arrangements reflect the operation of an invisible hand optimizing welfare or satisfaction, and to deny that social institutions are the result of what Hayek called "spontaneous order."<sup>xvii</sup> What evolutionary approaches show are that a)when behavior is the result of natural selection for outcomes that enhance fitness, instead of rational choice of outcomes that enhance individual welfare, market failures can be avoided and optimal outcomes may after all be attainable, and b) these outcomes result from the aggregation of individual behaviors, not the selection of some properties of the group (beyond those correlated pairs Sober and Wilson's group selection countenances). Of course, if the maximization of welfare is among the ways in which fitness is often maximized, then natural selection for individual fitness maximization will bring individual welfare maximization along with it, thus substantiating Smith's laissez faire conclusions if not his reasoning. Thus, successful Darwinian explanations in the social sciences will substantiate both

---

<sup>xvii</sup> Hayek, F., *Law, Liberty and Legislation*, v.1, Rules and Order, Chicago, University of Chicago Press, 1981.

methodological individualism and invisible or hidden hand perspectives. But there is another potentially more promising adaptation of Darwinism in social science.

If genes and packages of genes can replicate and be selected for in virtue of the adaptational phenotypes they confer on organisms, why cannot beliefs, desires, and other cognitive states be selected for as a consequence of the benefits thinking of them confers on cognitive agents. Following Dawkins, call these cognitive states “memes” (mental “genes”), whose varying individually rewarding effects in behavior (phenotypes) result in their being differentially copied (reproduced) into the cognitive systems of other agents. Here again, the attractions of memetic natural selection are its freedom from assumptions about the conscious rational choices of individuals to adopt particular ideas, values, fashions, etc, and the availability of an invisible hand mechanism that explains how they spread, become fixed in a population, and often become less widespread as environmental change (or even frequency-dependent selection) makes them less adaptive.

It would be wrong to suppose that Darwinism vindicates the notion, sometimes attributed to Smith and his followers, that social interactions, and economic ones, are largely competitive ones in which there are inevitably losers made extinct by the competition. As we have already seen, in some environments—i.e. under some pay-off distributions—individual selection makes cooperation the most adaptive strategy, not competition. That this is a possibility is something one might have inferred from Darwinian biology directly. For there are many natural cases in which selection fosters cooperation among organisms in different species, within the same species, whether closely related or not. And inference from Darwinism directly to a view of nature or society as “red in tooth and claw” is a mistake due to the neglect of the role of the environment which perhaps more often than not selects for competition and less often for cooperation. But Darwinism cannot deny the charge that non-competitive cooperation is in the end a strategy only locally adaptive, and adaptive for fundamentally “selfish genes” whose own fitness maximizing strategies organismal cooperation fosters.

## **8. Conclusion**

Darwinian morality has been a recurrent goal among naturalists. But, if the present orthodoxy among philosophers holds, it will remain an unreachable goal. Darwinian metaethics, on the other hand, seems to be carried forward on a rising tide of research into human affairs that twentieth and twenty first century research in game theory, biological anthropology, and evolutionary psychology. Several philosophers have made the most of the results, theories and findings which these disciplines have offered to provide an account of the nature and significance of morality. They have shown how it may be expected to have emerged among fitness maximizing animals, and how nature may have selected for the cooperative norms and the

emotions that express our commitment to them that give morality its universal content. The specificity and detail that these accounts seem already to have attained, should encourage opponents of naturalism to be modest in their claims about the long-term limits of a Darwinian understanding of human affairs.

Alex Rosenberg  
Duke University

## References

- Axelrod, R., *The Evolution of Cooperation*, Ann Arbor, University of Michigan Press, 1984
- Ayer, A.J., *Language, Truth and Logic*, London, Gollnaz, 1940.
- Barkow, R., Tooby, J., and Cosmides, L., *The Adapted Mind*, Oxford, Oxford University Press, 1992.
- Gibbard, A., *Wise Choices, Apt Feelings*, Cambridge, Harvard University Press, 1992.
- Griffiths, Paul, *What Emotions Really Are*, Chicago, University of Chicago Press, 1997.
- Hayek, F., *Law, Liberty and Legislation*, v.1, Rules and Order, Chicago, University of Chicago Press, 1981.
- Kitcher, P., *Vaulting Ambition*, Cambridge, .MIT Press, 1989.
- Lewontin, R, Kamin, L., and Rose, S., *Not in Our Genes* , New York, Pantheon Books, 1984
- Maryanski, A, and Turner, J., *The Social Cage*, Palo Alto, Stanford University Press, 1992.
- Moore, G.E., *Principia Ethica*, London, Routledge, 1903
- Peter Railton, "Moral Realism," *Philosophical Review*, 95 (1986): 163-207.
- Rosenberg, A., *Darwinism in Philosophy, Social Science and Policy*, Cambridge, Cambridge University Press, 2000.
- Skyrms, B., *Evolution of the Social Contract*, Cambridge, Cambridge University Press, 1996.
- Stevenson, C.L., *Ethics and Language*, New Haven, Yale University Press. 1944.
- Wilson, E.O., *Sociobiology*, Cambridge, Harvard University Press, 1976.
- Wilson, D.S., and Sober, E., *Unto Others*, Cambridge, Harvard University Press, 1998.
- Wright, L., *Teleological Explanation*, Berkeley, University of California Press, 1976.

## Notes