Darwinian Explanations of Morality: Accounting for the Normal but not the Normative

Jeffrey P. Schloss

The biologist, who is concerned with questions of physiology and evolutionary history, realizes that self-knowledge is constrained and shaped by the emotional control centers in the hypothalamus and limbic system of the brain. These centers flood our consciousness with all the emotions—love, hate, guilt, fear, and others—that are consulted by ethical philosophers who wish to intuit the standards of good and evil. What, we are then compelled to ask, make the hypothalamus and limbic system? They evolved by natural selection. That simple biological statement must be pursued to explain ethics and ethical philosophers, if not epistemology and epistemologists, at all depths . . . the time has come for ethics to be removed temporarily from the hands of philosophers and biologized.
–(E. O. Wilson, 1975)[1]

"The human brain is a product of evolution . . . We seem to be reaching a point at which science can wrest morality from the hands of philosophers,"
–(Franz de Waal, 1997)[2]

"Biology invades a field philosophers thought was safely theirs: Whence morality? That is a question which has troubled philosophers since their subject was invented. Two and a half millennia of debate have, however, failed to produce a satisfactory answer. So now it is time for someone else to have a go . . . Perhaps [biologists] and their colleagues can eventually do what philosophers have never managed, and explain moral behavior in an intellectually satisfying way."
–(Economist, 2008)[3]

"What shapes moral emotions in the first place? The answer has long been evolution . . . It challenges all sorts of traditions. It challenges the bookish

way philosophy is conceived by most people. It challenges the Talmudic tradition, with its hyper-rational scrutiny of texts . . ."

–(David Brooks, 2009)[4]

Introduction

Evolutionary theories of morality, notwithstanding the ambiguity of that phrase, have been viewed as an especially important subject of biological inquiry in recent years for at least three reasons. First, since the time of Charles Darwin's initial speculations on the connection between social instincts in animals and conscience in humans, morality has largely remained an unexplained conundrum for evolutionary theory, one which recently emerging theoretical tools and empirical findings have finally begun to illuminate. Second, in addition to being an unsolved problem, morality has also been looked to as a solution to another longstanding theoretical quandary in Darwinian theory: the widespread existence of human altruism or reproductively sacrificial cooperation. Third, the issue has been both prominent in the public representation of evolutionary theory and controversial within the scholarly community itself, initially because of those (from Huxley on) who were wary of the reduction of morality not just to material but specifically to Darwinian processes, and more recently, because a number of public exegetes of evolutionary theory have triumphalistically claimed that the reduction has essentially been accomplished. Many such claims are all the more provocative by virtue of failing to specify what kind of "reduction" has occurred: a) a descriptive account of how morally salient affective or cognitive *capacities* have evolutionarily arisen or are biologically mediated, b) a descriptive account of how dispositions toward particular moral *judgments* or *norms* have originated or (not the same, but frequently conflated) may be adaptive, c) a prescriptive, normative account of which moral judgments are in fact morally justifiable, and d) a meta-ethical analysis of how biological instantiation of morality influences our understanding of the grounding of moral truths, or if there even are such things as moral truths.[5]

This chapter affirms two perspectives on evolutionary accounts of morality that are often presented as strongly oppositional (including by some contributors to this volume). The first is that the fundamental behaviors, affective dispositions, cognitive capacities, and even many

concepts that are employed by or are unique to what we call human morality, are organismic phenotypes that have an evolutionary origin. More importantly, this origin is not opaque to scientific investigation, and recent theoretical and empirical work illuminates both the phylogenetic history and the selective forces that account for it. Such illumination is important both for evolutionary theory and for the moral enterprise itself.

The second is that crucial scientific questions remain to be solved. In fact, not only do we currently lack a fully adequate evolutionary account of morality, but the manifold accounts we do have are also disparate and are often represented by prominent exegetes as having resolved issues that are still in dispute. When E. O. Wilson famously posits that morality, like all behavior, is the means "by which human genetic material has been and will be kept in tact . . . morality has no other demonstrable ultimate function,"[6] he aptly expresses commitment to what he accepts as a Darwinian first principle or necessary truth. But it is, by no means, a conclusion that scholars of morality or evolutionary theorists take to have been empirically demonstrated to the point of excluding alternative proposals. Many evolutionary accounts of morality explicitly reject the notion that it serves to keep human genetic material intact.[7] Notwithstanding, Frans de Waal claims, "We seem to be reaching a point at which science can wrest morality from the hands of philosophers . . . The occasional disagreements within this budding field are far outweighed by the shared belief that evolution needs to be part of any satisfactory explanation of morality."[8] In ways I hope to demonstrate, this view both understates the serious and not occasional but ongoing disagreements within this field, and overstates the competency of science to address, much less resolve, important philosophical issues.

Still, the claim that evolution must be part of a satisfactory explanation of morality—though modest—is important. The problem is there is considerable ambiguity in what it even means to provide an evolutionary explanation of morality, and this ambiguity is often unrecognized by those positing different accounts. I want to begin by clarifying the range of meanings that attend this phrase. First off, providing an *evolutionary explanation* of any trait may involve several different things. It can involve phylogenetic analysis—using genomic, comparative physiological or psychological, and/or fossil data—to "explain" how the rudiments of a trait arose and were

successively refined or supplemented within the lineage of an organism (or across lineages in the case of convergent evolution). This enterprise may be entirely descriptive, with no adaptive or causal story proposed. Alternatively, evolutionary explanation can involve a causal proposal for the origins of a trait in terms of selection or other agents of evolutionary change. Finally, it can propose what current adaptive function a trait serves and may therefore be sustained by. It is important to note that each of these, especially the latter two, are not redundant but are often conflated. Demonstrating a function for a trait does not tell us how it evolved; providing an evolutionary account for a trait's origin does not tell us whether it presently serves any adaptive function.[9] And for any causal account, there are of course numerous options involving selection at individual or group levels, selection of genetic or cultural replicators, the trait as object or by-product of selection, and non-Darwinian (non-selective) agents of evolutionary change.

Similarly, what pass for explanations of *morality* often treat it, or at least popularly represent accounts of it, as if morality were a single phenotype or a unified trait for which "an" explanation could be proposed. But if there is any such thing as "morality," it is a constellation of interacting individual- and group-level traits that include, at a minimum, morally salient sentiments, concepts, behaviors, and social institutions. Moreover, for each of these, an explanation might seek to account either for the fundamental *capacity* (i.e., to feel, think, or behave in certain ways) or more ambitiously for central tendencies in *content* (i.e., what is the evolutionary cause or adaptive significance—if any—underlying central tendencies or cross-cultural universals in what humans feel, think, or do in morally salient situation *x*?).

Many assessments of evolutionary theories of morality conflate these manifold aspects of morality, by virtue of being organized around competing explanans—for example, spandrel accounts, individual or group selection, and cultural selection—for what is assumed to be a single explanandum.[10] In this chapter, I want to focus on three explananda that have often been compressed but are being increasingly distinguished in evolutionary literature: core behaviors widely, if not universally, imbued with moral significance, moral sentiments viewed as motivating or constraining behavior, and moral concepts that may emerge from and/or inform sentiments and behavior. Thus, morality does not involve a single explanadum but a 3:3:2 matrix of explananda:

	Moral Behaviors	Moral Sentiments	Moral Concepts
Phylogenetic History	Capacity / Central Tnd	Capacity / Judgments	Capacity / Norms
Evolutionary Origin	Capacity / Central Tnd	Capacity / Judgments	Capacity / Norms
Current Function	Capacity / Central Tnd	Capacity / Judgments	Capacity / Norms

"Moral" Behaviors

The most empirically clear but interpretively ambiguous focus of evolutionary studies of morality involves observable behaviors to which humans ascribe moral significance and from which some infer the existence and function of moral sensibility. This area of work examines several domains of behavior.

Animal Behavior

The first involves studies of animal behavior, especially but not exclusively emphasizing primates, with the goal of illuminating continuities between aspects of human and non-human prosociality that are taken to be morally salient. Frans de Waal and coworkers have argued that their justly famous observations of sharing, reconciliation, and consolation behaviors in chimpanzees constitute moral building blocks if not "proto morality"[11] (more recently described as genuine if primitive morality[12]). That human morality involves biological capacities shared by or rooted in the traits of other organisms is indisputable and important. However, de Waal goes beyond this, contending that there are no biologically meaningful discontinuities between humans and great apes: "It is the rare claim of human uniqueness that holds up for more than a decade . . . we have no basic wants or needs that cannot also be observed in our close relatives."[13] On the basis of posited continuity between human and animal prosociality, de Waal argues for a Humean, emotivist basis to ethics over and against a rationalist "veneer morality" that is not only unique to humans, but also pasted upon fundamentally selfish and morally recalcitrant social dispositions.[14] He surely is right to reject this vilified depiction of human (and animal) nature. But leaving aside questions of whether empirical findings offer definitive support for David Hume, whether rationality too is part of our nature, which enables the moral constraint and expansion of sociality

that sentiment cannot fund—there are several fundamental questions about the relationship of animal prosociality to morality.

First, it is not clear what is moral or even meaningfully proto-moral about prosocial behaviors in themselves. Reef fish groom other fish by picking parasites off their gums (as social mammals do in each others' fur); alligators, poison dart frogs, and even many arthropods care for their young; prairie voles form attachments to a single mate for life. The apt response to this is to note the hierarchy of prosocial complexity involving increasingly sophisticated conflict resolving and cooperation-enhancing behaviors, which appear to be underwritten by proto-moral emotions and intentions—specifically empathy.[15]

However, granting an empathic component to primate social behaviors does not necessarily betoken morality, even by emotivist standards. Empathy—the ability to intuit the feelings of another—is a morally neutral capacity that may be used for exploitation as well as aid. Specifically, *moral* salience entails employing empathy to direct behaviors toward the well-being of another. Moreover, even when it does result in aid rather than harm to another, this outcome may be a by-product of or means toward behavioral ends that are wholly indifferent to others' welfare—such as alliance building and dominance negotiations.

Second, it is surely reasonable to conclude that empathy, as found in primates and some other mammals, may be proto-moral in the sense that it is a necessary "building block to morality."[16] But for that matter, brains that mediate empathic capacities are also necessary, as are the oxygen-carrying pigments that supply energetic demands of metabolically intensive neural tissues. Is hemoglobin proto-moral? This is not to suggest a Gnostic view of morality that makes it entirely discontinuous with other animals by uncoupling it from creaturely embodiment. On the contrary, manifold biological capacities are clearly necessary for morality. But a fundamental question—significant both for understanding the moral enterprise and for explaining the human capacity to conduct it—is whether the various capacities that are present in other creatures, and that are necessary for morality, are also sufficient.

Primal Human Behaviors

Another area of behavioral study focuses on what might be considered "low hanging fruit" involving central tendencies, if not cultural universals, in morally valenced behaviors with clear adaptive significance like incest avoidance, parental care, and synergistic cooperation

(where the benefits of cooperation exceed those of defection). In many cases, such behaviors are associated with very powerful and demonstrably primal emotions, such as incest repugnance,[17] parental and mate attachment, and social bonding.[18] The proximal neurological mechanisms and ultimate evolutionary origin (both in terms of phylogenetic history and adaptive benefit) of these behavioral dispositions are fairly well understood. More recently, several elegant empirical studies of the role of aversive mechanisms in judgments about incest and other behaviors associated with moral purity have been represented as providing an evolutionary psychological explanation for morality itself.[19] Again, however, what these accounts do not illuminate is the moral or normative dimension to these behaviors. Given the adaptive consequences, the strong aversions, the universality in human populations, and (for incestuous inbreeding, for example) the prevalence of equivalent central tendencies in the behavior of non-human animals, why is morality a component of these (but not all aversive) behaviors, and in what way does the adaptive account provide an "explanation" of morality?

Game Theory and Cooperation

One of the most recent and fruitful areas of theoretical and experimental progress has applied the logic of evolutionary game theory to the domain of social behaviors involving cooperation, defection, and equitable distribution of risks and resources. While cooperation is clearly of adaptive benefit in many situations, in those situations where gains are additive rather than synergistic—that is, where cooperation does not yield gains that are qualitatively enhanced by collective action—it may be of even greater benefit to defect. This imposes a cost on cooperators such that the Nash equilibrium in dyadic interactions may be mutual defection (e.g., the well-known prisoner's dilemma) and in group interactions may involve a tragedy of the commons.

A concrete example of this involves rowing versus sculling games.[20] In rowing games, each participant has one oar. There is little incentive to cheat on investment, since there is a synergistic outcome of specialization, and defection imposes a cost on the defector—the boat goes off course! In sculling, where each participant uses two oars, one may defect without imposing a loss of directionality, and therefore, defectors share in the successful outcome without fully sharing the burden. This ushers in the commons problem of group laziness. Hume elegantly anticipated the moral solution to this problem in his *Enquiries*:

> Two men who pull on the oars of a boat do it by an agreement or convention, tho' they have never given promises to each other. Nor is the rule concerning the stability of possession the less deriv'd from human conventions, that it arises gradually, and acquires force by a slow progression, and by our repeated experience of the inconveniences of transgressing it. On the contrary, this experience assures us still more, that the sense of interest has become common to all our fellows, and gives us a confidence of the future regularity of their conduct: And 'tis only on the expectation of this, that our moderation and abstinence are founded. In like manner are languages gradually established by human conventions without any promise. In like manner do gold and silver become the common measures of exchange, and are esteem'd sufficient payment for what is of a hundred times their value.[21]

But while anticipating both the solution and the progressive process that may have given rise to it, Hume did not fully account for the dynamics underwriting its origin. Evolutionary game theory does so by rigorously describing the conditions under which cooperative investment or equitable division of profits can arise and be sustained as an evolutionary stable strategy.[22] In one of the best known (though subsequently much nuanced) examples of game theoretic analysis—Robert Axelrod's competition involving an iterated prisoner's dilemma—the most successful strategy was "tit-for-tat": I'll do to you what you did to me.[23] Stanford primatologist and neuroscientist Robert Sapolski has suggested that the Golden Rule and its widespread cultural variants are instantiations of this tit-for-tat reciprocity.

Actually, this is not the case, for the Golden Rule is not "do to you as you did to me," but "do to you as I would like you to do to me." As such, it conforms to a strategy called "forgiving tit-for-tat," which Martin Nowak has demonstrated dominates in iterated games with the more realistic inclusion of imperfect communication.[24] Where it is possible that an actor accidentally defects or where an actor cooperates but is mistakenly believed to have defected, it turns out to be more adaptive to give a second (and as generations of selection ensue, a third, a fourth . . .) opportunity. While this sounds moral (and indeed, it seems to be a good thing that the world works like this—but maybe not altogether good![25]), note that this strategy is converged upon by digital bits of code replicating in a computer. And in the biological world, we see it instantiated not only in humans but also in animals ranging from reef fish to foraging bees.

Here as well, it is not clear how evolutionary game theory explains *morality*. It is not that game theoretic accounts do not adequately

account for the behavioral strategies to which we attach moral significance; it is that it does not explain why we attach to them *moral* as opposed to prudential or instrumental significance. Indeed, it is not clear why valuation of any kind is involved, since we have found these evolutionary solutions to cooperative stalemates by determining which strategies can be demonstrated to be stable via computer programs that generate interactive strategies without values, feelings, cognitive internalization of rules, or anything other than the ability to replicate in ways that are sensitive to the existence and frequency of other strategies. Moreover, we observe natural selection to have solved analogous challenges to cooperative commitment in a series of major evolutionary transitions that don't involve morality—from gene to chromosome, from prokaryotic to eukaryotic cells, from single cells to multicellularity to social organisms.[26] Thus, the structure of at least some moral norms may reflect game theoretic solutions to cooperative barriers; but these solutions can be achieved without morality, and game theory itself does not provide an evolutionary explanation for its origin or function. Whatever morality is and however it came about, it clearly evidences the stamp of selection. But this is not the same thing as explaining either the origin or the adaptive function of morality itself.

Justin D'Arms points to this issue, claiming that such "evolutionary attempts to explain morality tend to say very little about what morality is." He argues that if evolutionary game theory is to

> "advance our understanding of morality . . . it must include an essential role for moral sanctions. Such an account might best begin with the moral sentiments, exploring how dispositions to feel them in response to some outcomes and not others were fitness enhancing for our ancestors."[27]

He is correct in identifying the fact that game theoretic approaches explain stable cooperative strategies without reference to *moral* values (or even conscious prudential assessment). However, as I will suggest in the next section, his understanding of morality as only involving "sanctions" is inadequate both for its lack of emphasis on approbation and for its failure to distinguish the specifically moral character of sanction (e.g., violation of dominance hierarchies in social mammals involves punishment and sanctions, but is not necessarily moral). Also, it is actually not accurate to say that evolutionary game theoretic accounts ignore sanctions and sentiments: there have been a number of recent game theoretic proposals for and empirical demonstrations of

both the importance of sanctions[28] and the internalization of punitive or sanction-avoiding sentiments.[29] However, again, whether these are *moral* in character is open to debate. There are also recent countervailing proposals for the limitations and even detrimental effects of punishment on cooperation.[30]

Thus, evolutionary game theoretic approaches have the dual benefit of accounting both for the stability of cooperative fidelity and for the emergence of sanctions against defectors. But aside from the fact that humans *do* ascribe moral salience to these phenomena, the evolutionary accounts themselves neither predict nor explain this. They merely offer an account of how certain cooperative strategies can be stabilized across a range of interacting and replicating entities—from genes to cells, to insentient social organisms, to social organisms ranging in sentience from insects to fish, to primitive mammals and primates, and to human moral agents.

Altruism

Game theory illuminates the kinds of moral standards that make evolutionary sense (though it does not tell us why these standards are moral in character). Altruism involves a behavior that seems to be highly if not quintessentially moral, but violates standards of Darwinian rationality.

Unfortunately, this important field of inquiry has been persistently beset by terminological ambiguity. In psychological and philosophical traditions, altruism typically refers to motivations (something like giving without expectation of return, or helping others with their benefit as a primary goal). In evolutionary theory, it usually (but not always) refers to consequences: helping another at some cost to the actor. Even with an emphasis on biological outcomes, there has been regrettable lack of precision in usage. In a range of biological literature, altruism has loosely referred to any helping behavior that is directed to non-progeny and is attended by some cost. Alternatively (and in its original employment by biologists), it has referred very specifically to a behavior that has a *net* cost, for which the cost metric is a reduction in the actor's *fitness*, while conferring a fitness benefit to others. This particular issue is the thorny problem of biological altruism, defined succinctly by E. O. Wilson as "self-destructive behavior performed for the benefit of others," which constitutes "the central theoretical problem of sociobiology."[31] The quandary is clear enough: natural selection promotes traits that enhance fitness, or at least, cannot sustain traits that

subvert it. Darwin himself observed that the existence of exclusively other-benefiting traits would "annihilate my theory, for such could not have been produced by natural selection."

The history of evolutionary theory over the last generation has entailed a series of immensely fruitful theoretical insights into these issues, referred to by some as a "second Darwinian Revolution."[32] Sequential developments, explaining a progressively wider domain of cooperative behaviors, range from kin selection (helping genetically related non-progeny)[33] to reciprocal altruism (helping non-kin likely to make a compensatory return),[34] to indirect reciprocity (conspicuously helping those who may not reciprocate, thereby generating reputationally mediated returns)[35], and to still-debated proposals for group selection (helping group members at net cost to the actor—relative to other group members—but at net benefit to the actor relative to other groups).[36]

The important thing about all of these proposals is that each one accounts for genuinely costly investment in non-progeny that appeared problematic for natural selection. *But equally important and widely misunderstood: they do this, not by explaining how "altruism" has evolved, but by explaining how these behaviors are not altruistic.*[37] This is because the investment or cost is understood to involve a net increase, not reduction, in fitness averaged across the situations in which the behaviors are deployed.

A recent and representative synopsis of the altruism issue in evolutionary theory affirms Wilson's thirty-year-old comment and recognizes its import for a biological theory of morality: "The problem of altruism, both biological and psychological, is at the center of grounding a theory of morality within biology."[38] And regrettably, though not atypically, it then reiterates the above approaches, none of which have clear connections to morality, and all of which stop short of solving the altruism question.

The problem, of course, is that any strictly or even largely Darwinian account of morality's origin and function cannot account for biological altruism (in the Wilsonian sense): selection for fitness-enhancing behaviors cannot explain the genesis or maintenance of fitness-subverting behaviors. And to whatever extent a cultural account is employed, while the fundamental cognitive or affective capacities involved may have been objects of selection, they must now function in a way that does not clearly reflect that fact and involves not mere irreducibility to genetically determined proclivities, but—it would seem—opposition to them[39] or at least a breaking of what Wilson has called "the genetic

leash."[40] In a typically provocative proclamation that clearly recognizes this, Richard Dawkins identifies the power of cultural information (moral "memes") as opening up possibilities for "cultivating and nurturing pure, disinterested altruism—something that has no place in nature, something that has never existed before in the whole history of the world . . . We, alone on earth, can rebel against the tyranny of the selfish replicators."[41] But this is far from an explanation of morality in terms of biological evolution; in fact, it is an acknowledgement that such an approach cannot succeed.

While proposals for the culturally mediated and to some degree genetically uncoupled nature of altruism seem inescapable, they have provoked a firestorm of controversy amongst evolutionary biologists. Geneticist Gunther Stent opines that suggesting genes give rise to the capacity to resist genes is a "biological absurdity."[42] Frans de Waal calls this a naïve and banal dualism,[43] the essence of a veneer morality that identifies the best of humanity as being contrary to the nature of humanity.

Whether a behavior to which we are not naturally disposed is "contrary" to our nature, or whether a gap between innate capacities and moral demand represents a dualism that is absurd or naïve,[44] are issues largely unattended by evolutionary biologists weighing in on these issues. But leaving aside the *ad hominems*, there remain several unsolved problems. I should mention that one of them is not the issue—raised by Hilary Putnam in this volume and by many others writing on this topic—that cultural change is Lamarckian rather than Darwinian and hence not amenable to the logic of selection and not rightly considered evolutionary. First off, the Lamarckian/Darwinian distinction does not rigidly apply to cultural versus genetic information. But second, even if it did, the more important thing is that both forms of information may be replicated and transmitted. And with this comes the possibility of differential replication, which is precisely what natural selection entails. Therefore, while such theories are not biological, they can be formulated in ways that, according to advocates, are not just evolutionary but fully Darwinian.

This raises two problems. First, the quest to employ Darwinian logic to cultural evolution has generated proposals for particulate units of cultural information most readily amenable to the "mutation selection" mechanism. But the particles ("memes") turn out to be ill-defined, and their transmission dynamics are unaccounted for.[45] Jerry Coyne opines that unlike our understanding of the relationship between phenotype

and genotype, which informs the dynamics of differential genetic transmission, selection acting on memes "is completely tautological, unable to explain why a meme spread except by asserting, post facto, that it had qualities enabling it to spread. One might as well say that relieves pain because of its pain-relieving properties."[46] Moreover, unlike genes, it is unclear that memes always require "spreading": in a suitable cognitive substrate—a mental "warm pond" so to speak—some memes may arise through the analogue of spontaneous generation or biogenesis, via processes of intuition or rational inference. There is, in fact, currently no specific evolutionary explanation for the origin or transmission of altruistic memes.[47]

Some of the problems entailed by the more reductive versions of memetics are avoided by recent proposals for the expansion of prosociality by culture.[48] But even these approaches entail a second and more difficult problem of providing a plausible account of the fulcrum by which altruistic moral ideas exert their causal leverage. On this point, de Waal seems justified in his critical characterization of the project as: "A position in search of a theory. It offers no explanation of why humans are 'nicer than is good for their selfish genes,' nor how such a feat might have been accomplished."[49] That a moral concept is irreducible to biology is not itself the problem, but for this to remain an *evolutionary* account (or any biological explanation at all), as opposed merely to being a concession of what doesn't suffice for an explanation (i.e., genes), we must understand why the organic substrate is so receptive to the meme or moral norm.

This is by no means impossible: we have numerous examples of cultural innovations that co-opt evolved reward pathways in ways that do not benefit fitness (e.g., opium ingesting, contraception, consumption of agriculturally produced sweets). The major proposal along these lines for altruistic notions is that they may co-opt proclivities related to parental care or familial attachment. Such proclivities help unconditionally without expectation of return, which E. O. Wilson calls "hard core altruism." The problem with this is that the data we have on the most significant and sustained forms of altruism (e.g., Holocaust rescuers) do not manifest similarities to attachment-mediated behaviors at all (although occasional cases of release by guards who befriended individuals do).[50] For this reason, Wilson and others look to "soft-core altruism" or socially mediated norms of cooperative reciprocity to expand the circle of care. But as the circle of cooperative care expands, the degree or conditions of sacrificial investment typically narrow. That

this is not altogether the case with human sociality has been designated "the culminating mystery of all biology."[51]

The bottom line is that altruistic behavior constitutes a puzzle not yet fully solved, and altruistic moral norms do not solve it. Moreover, if they did, then they would be the problem needing a solution. To the extent that morality is part of the explanation for why humans are the only creatures systematically to trade their own fitness for that of others, we do not have anything close to an evolutionary account of how such notions (whether or not we call them memes) originated, how they modify central tendencies of behavior that are prevalent in all other species and that we would expect on Darwinian grounds to constrain human behavior as well, and why they are successfully retained in human social systems.

Moral Sentiments

Each of the above, especially the game theoretic approaches involving the administration and receipt of sanctions, raises the issue of moral sentiments. It is not completely clear that what is called "moral sentiments" comprises a single "moral system." For example, Jonathan Haidt has argued that there are five major affective inputs into moral sensibility, which function relatively independently and which vary substantially between persons and cultures. Moreover, some sentiments may be employed by but are not unique to moral sensitivity (e.g., disgust, anger) and others may be distinctively moral (e.g., guilt, shame).

Interestingly, the first two more generic sentiments—disgust and anger—seem the easiest to construct an adaptationist story for. Repulsion over incest, or over having commerce (sexual or otherwise) with dead bodies makes imminent reproductive sense. What is not clear is why disgust alone, sans moral valence, is not sufficient to promote appropriate behavior. Perhaps it is, and the moral attribution is a byproduct or a pleiotropic[52] and motivationally inert overlay upon the sentiment. However, this would not explain why some forms of repulsion receive the attribution and others do not, nor would it explain recent empirical findings about the moral attribution.[53]

Moral "outrage" is amenable to the adaptive logic of protecting one's own prerogatives from aggression or defection, or protecting relatives or cooperation partners. Indeed, the sentiment itself may not differ much from what is inferred in animal behavior studies: retribution for violating or challenging dominance "rules," ostracization for violating norms of play,[54] harem or territorial defenses. Chris Boehm has

suggested that in human evolution, coalitions emerged that punished overt dominance or stealth defection.[55] While these punitive behaviors would be egalitarian promoting, the sentiments that fueled them might well have been co-opted from those underwriting the very dominance they challenged. But here too, what is not clear is why such feelings are given moral salience. C. S. Lewis has observed that when someone punches a thief to save his television, he is a regular guy; when he does this (or says he does this) to defend the right of private property, he is an ass—who is kidding himself but not others.

Evolutionary theory may offer both an insight and a caution here. The insight involves why humans might fund certain behaviors by construing them in elevated terms of honoring the "right" rather than merely protecting self-interest. Motivational (and metabolic) resources may be marshaled by optimistic forms of cognitive distortion, since the over-estimation of potential gain and underestimation of risk may amplify investment and help overcome commitment barriers. This is one theory for the idealization of the beloved in romantic attachment, or for the efficacy of the placebo effect, which by elevating the estimate of success, may conscript rather than reserve scarce metabolic resources to meet an immunological challenge.[56] It is has also been proposed to be important in marshaling resources for and committing to armed combat and other forms of conflict.[57] This line of thinking has been employed in a proposal for the cognitive dimensions of morality, involving internalized beliefs in cosmic sanctions (which I will emphasize in the last section), which may promote commitment by the sense that the cosmos is with you (and the likelihood of payoff is high) when behaving morally, and against you when not.[58] Thus, the invention of moral fictions (or the intuition of moral realities—this is not in itself a scientific question) that supplement nonmoral sentiments may (a) heighten motivation for costly or risky punitive behaviors and (b) more vigorously restrain impulsive behaviors that risk earning the actor censure by the group. Indeed, in his *Enquiries*, Hume says as much about the latter's contribution to restraining the knave's indiscretions.

The evolutionary caution is this: if dispositions toward feeling disgust at certain things are native and have arisen by natural selection, they must have—on average—been adaptive in the range of ancestral environments in which they took on moral salience. But they needn't have been adaptive in every situation; they needn't have been optimally adaptive in any situation; and they may presently be maladaptive in most or even every situation. Moreover, because such cognitive distortions

are by definition opaque to the actor, we may be unaware of both the distortion and the fact that it does not attend flourishing.

For example, there are many instances of disgust that are attended by moral judgment, but for which neither the disgust itself nor the concomitant judgment is morally justifiable or biologically adaptive. The repugnance at cultural out-groups, which may have had prior adaptive significance for immunological reasons, or the repugnance at interracial mating, which may relate to evolved preferences for symmetries that reflect population averages, deserves to be discarded.[59]

This is part of the reason for widespread criticism of Leon Kass's[60] prominent suggestion that repugnance, which "we intuit and feel, immediately and without argument" is an important component of moral judgment that signals "the violation of things that we rightfully hold dear." In this volume, Steven Pinker characterizes this as suggesting that "we should disregard reason."[61]

Strictly speaking, however, the latter is not the case. Kass does not propose we disregard reason (indeed, he is attempting to employ reason, however convincingly or unconvincingly, in his appeal to heed repugnance). What he is proposing is that there are some inputs into moral judgment that do not rely on *argument* to earn the right to vote, not unlike Jonathan Haidt proposes for moral sentiments, or Marc Hauser proposes for moral cognitions (both of whom, quite interestingly, Pinker endorses). If there is a problem with Kass, it is not that he grants the vote to a judgment that precedes an argument, but that he does not tell us how to count the vote once it is cast. Thus, we have no way to differentiate between repugnance that should be heeded and repugnance that should be overridden, or for the role of argument in making this distinction.

But Pinker's biologically grounded criticisms of Kass and the moral salience of repugnance do not accomplish this either. To be sure, in his recounting of history, the sheer number of times that repugnance has stood in the way of moral or cultural advance suggests the wisdom of skeptical disregard for its voice. But there are also occasions in which the quelling of repugnance has been morally catastrophic. I shall never forget the first and indelibly profound experience of moral repugnance—moral horror actually—when, at the age of what would have been a bar mitzvah, had my family of Holocaust refugees been observant, I was taken to see the film *Judgment at Nuremberg*, with the mounds of bull-dozed corpses in the camps. As is evident in myriad genocidal programs and orgies of atrocity, dismissing or reformulating

repugnance by derogating moral sentiment may entail grave moral risk.[62]

Evolutionary theory is not in a position to solve these issues much less "wrest morality from the hands of philosophers." However, it is in a position to contribute to discussion by illuminating the ways moral judgment may involve cognitive distortion, the ways it may reflect reified, affective vestiges that did not anticipate emerging social or technological innovations, and the ways in which fundamental pre-cognitive moral dispositions constitute affective algorithms for human flourishing. The philosophical question is how to assess these moral sentiments and the intuitions they attend; the biological question is how they both help form and are transformed by culturally labile cognitive understanding.

Conscience, Guilt, and Self-Approval

While aversive sentiments (disgust, or repulsion from doing harm) and indignation (punitive sentiment or offense at social defection) may be imbued by humans with moral significance, they are not intrinsically moral in character. In fact, they (along with other affective capacities like that of empathy) may exist in animals that, to use a distinction suggested by Harry Frankfurt[63] and cited by Philip Kitcher and Christine Korsgard[64] in their criticisms of attributing proto-morality to non-humans, are "wantons" rather than "persons," that is, creatures whose behavior seems to be governed by whatever happens to be the most powerful affective impulse at the moment.[65] Darwin himself noted this and concluded that "A moral being is one who is capable of comparing his past and future actions or motives, and of approving or disapproving of them. We have no reason to suppose that any of the lower animals have this capacity."[66] Similar to Darwin, Jerome Kagan, in an emphatic critique of de Waal, argues that specifically, moral sentiments require the ability to envision future consequences of actions upon others, to choose freely between alternative courses of action, to reflect on past choices and recognize distress such choices may have caused, and to feel guilt or self-approval in response to this evaluation. Although he acknowledges that these involve biologically rooted capacities, he holds them to be unique to human beings, "biologically prepared biases [that] render the human experience incommensurable with that of any other species."[67]

While asserting that these capacities are unique to humans is not central or even directly relevant to asserting that they are fundamental

to morality, the fact that they do appear to be distinctively human has influenced construal of what is called moral and ensuing accounts of how morality may have evolved. For example, in the face of the discontinuity between humans and animals that this distinction seems to entail, Darwin ended up postulating conflicting accounts of conscience. In the above and other passages, he suggests conscience involves social instincts (established by group selection) along with the development of "intellectual powers" seen presently only in humans (though by no means necessarily restricted to them in future evolution). Elsewhere, though, he describes the "regret" of conscience as merely involving the feeling that ensues upon choosing, from amongst two desires, the one that is most intense but also most fleeting, which leaves the more enduring desire unsatisfied and experienced as lingering regret. Economists and evolutionary psychologists refer to this as future discounting, the existence and consequences of which are widely evident in human and non-humans.[68]

Although it is not clear in Darwin's formulation how remorse over unsated desires is a moral sentiment, it is this latter view that characterizes many evolutionary accounts of conscience. In his seminal work, *The Biology of Moral Systems*, Richard Alexander[69] develops an evolutionary account for the role of conscience in solving future discounting problems that is both uniquely human and (he holds) distinctively moral. He argues that human beings are the only social primate living in group sizes that are too large for direct reciprocity, which requires personal history with exchange partners in order to determine whom to cooperate and not cooperate with. Cooperation at the scale of human social exchange must be facilitated not only by direct knowledge through past interactions but also by indirect knowledge through moral reputation as enabled by language: "for direct reciprocity you need a face, for indirect reciprocity you need a name."[70]

Indirect reciprocity (IR) ends up not only enlarging the domain of cooperation to include strangers but also expanding the degree of cooperation to include situations in which an actors' behavior may not be reciprocated by the beneficiary, but can nevertheless be compensated for by benefits to reputation. "Moral" standards are the culturally variable but cooperation-salient rules by which reputational capital is assessed. And conscience, according to Alexander, functions as a "reputation alarm" that goes off when you are behaving in a way that involves present reward but is likely to entail future reputational losses. Hume seems to have given a similar account of this risk in his

description of "knaves betrayed by their own maxim; and while they purpose to cheat with moderation and secrecy, a tempting incident occurs, nature is frail, and they give into the snare; whence they can never extricate themselves, without a total loss of reputation, and the forfeiture of all trust and confidence with mankind."[71]

Alexander's proposal does not just involve the adaptive benefit of forgoing present benefits that might injure future reputation, but also posits the future fitness *benefits* of forgoing present benefits, where such abstention might actually advance reputation. As such, it is part of a currently growing emphasis on "costly signals" in evolutionary theory, which began with Darwin's ideas about sexual selection and have more recently been expanded as the "handicap principle."[72] Just as the weight and conspicuous color of a peacock's tail entail costs in the benefit of attracting a mate, so public acts of charity or magnanimity might involve forgoing the pleasures of presently retaining a resource, for reputationally mediated future gains of that or other resources. (Indeed, the logic is not only analogous to that of sexual selection, but some have proposed that virtuous displays function as sexual signals.[73] This principle has also been proposed for why humans invest so much in seemingly "useless" enterprises like art.[74])

Since Alexander's proposal, there has been a flowering of both game theoretic models specifying the conditions of punishment, reward, and reputational transmissibility under which IR is viable, as well as persuasive observational and experimental demonstrations of its efficacy in stabilizing cooperation. There have even been recent observations of IR functioning in non-human animals, involving bystander observation rather than oral transmission of cooperative fidelity. Chimps appear to bias their interactions toward individuals they have seen exhibit generosity before direct interactions; even grouper fish will avoid interacting with a cleaner wrasse if they have observed another grouper flinching when the wrasse "cheats" by taking a piece of gum material rather than the ectoparasites they are "supposed" to clean.[75] This is secondary IR; humans are the only species known to evidence tertiary or higher-order IR. (We stay away from careless dental hygienists, if our friends or friends' friends report they are not to be trusted!)

However, there are two clear shortcomings to IR's adequacy as an explanation of moral sentiment. First, it is not clear how the proposed "reputation alarm" of conscience represents *moral* sentiment, that is, sentiments tied to assessment of a self's rectitude, as opposed to being merely an affective instrumental calculus of likely material gain

or a fear of retributive aggression or social exclusion.[76] In the same passage of the *Enquiry* cited above, Hume makes this distinction in affirming "the invaluable enjoyment of a character . . . the peaceful reflection on one's own conduct . . . [such] pleasures, indeed, are really without price; both because they are below all price in their attainment, and above it in their enjoyment." Second, many moral norms (and the sentiments that seem to underwrite them) specifically urge, "give not as the hypocrites give, with trumpets before men, but give in secret." Indeed, first-person reports and frequently observed behaviors—from anonymous donations to whistleblowers to Holocaust rescuers—clearly demonstrate that reputational benefit is not all, and perhaps not even part of what there is to conscience. Not only is following conscience routinely done privately, but it also often compels, not constrains, behaviors that violate rather than obey prevailing social norms—and this results in extraordinary reputational loss, even forfeiture of life.[77]

In a critique of both Alexander's theory of morality and Herbert Simon's notion of altruism being due to a combination of "docility" and "bounded rationality,"[78] economist Robert Frank has developed an alternative and very influential evolutionary account of how moral sentiments (which he refers to as passions) are adaptive, precisely because they are *not* coupled to adaptive intentions, and are not even readily calculable to have adaptive outcomes.[79] But neither do they reflect stupidity or laziness. Rather, they are imminently reasonable, but not strictly calculable, in the following way: Because rationally self-interested reputation maximizers are just the kinds of people that others desire to avoid as cooperation partners, Frank argues that such strategies are likely to be maladaptive in a species such as human beings, who have a well-developed theory of mind and sophisticated abilities to infer the motivations of others. Thus, prosocial behaviors that cannot conceivably have a positive reputational impact—even something as simple but adaptively enigmatic as tipping in an out-of-town restaurant—if done routinely, both reflect and habituate behavioral dispositions that may function as reliable nonverbal signals of character.[80] This approach illuminates the difference between "costly signals"—whose cost may be calculated and therefore hypocritically born for the sake of gain—and "hard-to-fake" signals, unconscious displays that emerge from and genuinely reflect behavioral dispositions that transcend an actor's desire for personal gain. These signals are posited to have adaptive benefit, the value of which can be "calculated"

by natural selection, but not by the individual: Moral virtue may generate socially mediated fitness rewards, but only to the extent that it is pursued as its own reward.

Frank's adaptationist proposal applies to all sorts of sentiment that is felt to be morally funded. For example, not only may what most of us regard as the moral virtue of anonymous generosity have fitness benefits by conveying cooperative disposition, but also the feeling of honorific duty to exact revenge (for those who pursue it, often felt as a moral obligation) may have adaptive benefit. Precisely to the extent that revenge is "irrationally" pursued at a cost that exceeds the value of attaining restitution or just compensation—back to Lewis's distinction of protecting your TV versus protecting the right of private property—it may function as a deterrent to those who would otherwise feel free to mess with you in circumstances where the costs of your exacting retribution are so high that no reasonable person would bear them.[81]

According to Frank, some people really are defending private property rights and not just their own property, and it is reasonable on evolutionary grounds to do so, even at cost to their own property. His is the major evolutionary account of motivations and sentiments (including and especially moral sentiments) that involves the human capacity to value things that—unlike valuing the well-being of family or members of our cooperative matrix or our reputations[82]—have no clear connection to adaptive benefit. In game theoretic terms, it accounts for human psychological utilities that appear to be divergent from the evolutionary utility of fitness.[83] But importantly, they are not.

There is much to commend in Frank's account. It makes evolutionary sense of a creature that can not only recognize but celebrate the truth in Hunter Thomson's witty paraphrase of Hume (and other moral thinkers and wisdom teachings): "Honesty does pay, but he who is honest for that reason, is not honest." It avoids de Waal's apt criticisms of "veneer morality" by providing an evolutionary rationale to moral sentiments. Yet, it expands the objects of motivational affections to things that people feel obligation to—legal concepts, national identity, art, and scientific truths—that are cultural inventions, not in our genes and not all of which are even pan-cultural.

While this proposal may be part of an account of moral sentiments, there are two arguments against its adequacy. First, Alexander, Robert Trivers (whose seminal work on reciprocity laid the groundwork for Alexander) and others have argued that if hard-to-fake signals do

have fitness benefits, then the very best strategy would be to have the signal, but not have the behavioral disposition that the signal typically attends: that is, while being genuinely virtuous may be better than being a moral hypocrite, being morally self-deceived would be better still.[84] They point to a substantial social psychological literature that demonstrates systematically biased cognitive distortions about motives and virtues, and argue that—back again to the Lewis image—actual behavior indicates that moral sentiments are routinely attached to televisions when people say (and apparently believe) they are attached to higher principles.[85] Of course, either may be true, and an important aspect of the moral life involves distinguishing between genuine other-regard and self-interest masquerading as the former under the cloak of a moral principle.[86]

The second argument against the adequacy of the sentiment-as-signals account of morality is that just as Alexander's theory is challenged by cases of moral sentiment motivating behavior that has no reputational benefit, so Frank's solution to Alexander's deficiency is challenged by examples of sentiment-motivating behaviors that have negative consequences for fitness, uncompensated by the benefits of signals. Tipping in an out-of-town restaurant or anonymously giving to the poor may generate signals with positive fitness effects that more than compensate for the relinquishment of monetary resources; being fed to the lions along with your whole family does not. We are back to the lack of a plausible evolutionary account for Wilsonian altruism and the feelings of compassion or moral obligation that motivate it.

Moral Beliefs: Categorical Norms and Cosmic Sanctions

There are varying evolutionary accounts for the biological underpinnings of moral cognition. For example, Jonathan Haidt posits that moral beliefs arise spontaneously from specific non-rational sentiments that are evolutionary adaptations. The beliefs themselves, though, may do little work in moral behavior. Alternatively, Marc Hauser proposes an innate moral grammar that shapes ethical reasoning according to the contours of fitness. Whatever the case, the fact that we have moral beliefs and that these beliefs entail notions of normative demand that is taken to transcend our own desires, is central to most people's experience of moral life and most reflections on its nature. However, in contrast to psychological, neurological, and philosophical work on ethics, this cognitive aspect of morality is an issue that *evolutionary* theory has dealt the least with.

Moral Norms

There are two features of moral cognition requiring explanation: belief in particular moral norms, and the fundamental belief that there is a normative moral reality at all. The dominant sociobiological and evolutionary psychological approach focuses on the former. It views moral norms as "epigenetic rules" that may not be fully reducible to but do emerge from genetic proclivities to adaptive behaviors: for example, incest avoidance, parental care, and repaying cooperative investments. But if these behaviors have adaptive value and we are genetically disposed to them, then why have moral norms? A common answer is that since human behavior is labile—which is itself an adaptation, but an adaptation that if unconstrained can result in disbenefits—we need "back-up mechanisms" to restrict the range of behaviors. Parents do desert or abuse their children, incest does occur, and sex and close contact with corpses do occur. Our innate repugnancies may not always be effectively constraining. In the film, *Lawrence of Arabia*, an American reporter extols the virtues of Lawrence to Prince Fasel: "Prince, Lawrence is so merciful!" The Prince replies: "For Major Lawrence, mercy is a passion; for me, mercy is a convention. Judge for yourself which is the more reliable of the two." By the end of the movie, Lawrence is shouting, "Take no prisoners!"

But there are two problems with accounting for moral norms as epigenetically encoded backup mechanisms. First, as is the case with evolutionary accounts of moral sentiments, there is no proposed explanation for the specifically *moral* dimension of these rules. The Prince identified the problem to be solved and specified a solution: "convention." There is no clear adaptive reason or evolutionary rationale for why cognitively affirmed behavioral standards must carry the weight of morality rather than that of social rules or convention (which may be emphatically transmitted and stringently enforced—think of shined shoes in the military). One way to put this involves the distinction between normed and normative behaviors that may be underwritten, respectively, by structuring versus moral algorithms. If there is adaptive value to both variability and central tendency in important behaviors, but if native dispositions allow too much "drift" in such behaviors and therefore require cognitive backup: Why do such backups delimit the normal by asserting the morally normative?

Second, and more significantly, it is difficult to see how this proposal actually works. If moral norms back up moral (or even nonmoral)

sentiments, and the former are sculpted by the same biological dispositions that sculpt the latter, then they are not backups. Of course, one way cognitive beliefs may effectively backup emotional dispositions is if they are arrived at independently. If I have a natural desire for fruit, and if I also have come to understand that fruit is necessary to make up for my primate-specific inability to synthesize vitamin C, then I am much more likely to make sure I eat fruit, even when it may be inconvenient to obtain. But the advocates of epigenesis are not suggesting anything like this for morality. Humans have not come to the belief that incest is wrong because we learned about inbreeding depression. Nor, according to sociobiological accounts, have we devised Golden Rule morality because we came to understand the nature of commitment barriers and the principles by which they can be overcome. Rather, we have come to these beliefs because we have innate cognitive structures that, while perhaps not determining such beliefs, bias us toward something in their neighborhood. On In the nativist view—the view by which moral judgment is "constrained and shaped by the hypothalamus"[87]—it is therefore not clear how morality constitutes a backup.[88]

Moral Demand

The other major issue in evolutionary theories of moral belief involves the origin of what seems to make a moral norm a *moral* norm, that is, why do humans believe in categorical imperatives that demand conformity, regardless of desire and regardless of what seems will be the cost of obedience? One emerging account relates to notions mentioned earlier involving placebo effect, overconfidence, and commitment of resources. The general approach is referred to as "error management theory," and it goes like this. Every social decision (indeed every organismic "decision," including how to spend your next metabolic calorie) involves risks and benefits associated with uncertain outcomes. If I had perfect knowledge of the outcome, I would choose the behavior (cheat or not cheat) or choose the metabolic investment (repair my tissues, or have babies fast) that would pay the most in fitness. But I don't have such perfect knowledge. If the risk of an error in my calculations of outcome that inform a set of choices is symmetric, then I should be expected to do these calculations as accurately as I possibly can, and then make my choice in the context of random errors of assessment. But if the risks or costs of error are not symmetric, my cognitive dispositions will be biased in favor of not making a particular kind of error. For example, the time cost of stepping away when I errantly think the

bush next to me rattles because a snake is in it is considerably less than the cost of not stepping aside when I think it's due to the wind, but it is actually a snake.

A version of evolutionary game theory—building on Hume's insights mentioned earlier—posits that belief in moral reality is a means of error management. According to this view, it is not enough for conscience to function as Alexander's "reputation alarm." This is because in such a case the set-point for an alarm going off is the very best assessment of when reputation is indeed likely to be compromised, but the costs of being wrong are unacceptable for a species so completely dependent on inclusion in the social matrix. Thus, at least in the prevailing proposals, conscience has generated cognitive beliefs that are biased toward over estimates of getting caught. In fact, beliefs that are common to moral systems—a punishing or approving God who sees the heart and all deeds, rewards or punishments in karma or reincarnation, a morally structured cosmos that naturalistically confers rewards and punishments, or just a construal of human telos for which genuine virtue is necessary to happiness—all entail notions of essentially infinite improbability of getting away with wrongdoing. Such beliefs are posited to constrain the fitness-reducing defections that are likely to occur in the face of epistemic frailty underlying behavior[89] and to motivate behavioral valor that ensues from believing that the right (or God) is on your side.[90]

Error management proposals for moral beliefs have as such been the subject of several critical reviews,[91] but I will make two brief comments. First, scientific findings have not demonstrated that belief in a moral structure to life entailing intrinsic and/or reliable consequences for flourishing of moral choice, involves a "cognitive bias" that exaggerates the likelihood of such consequences. The metaphysical underpinnings of such a belief cannot be addressed scientifically; and the relationship between various measures of flourishing and moral behavior is an empirical but unresolved question. Second, we do not yet have enough empirical data about the actual social risks of getting caught, the benefits of defection, the frequency of opportunities to defect, and the likelihood of getting caught—all these data would be necessary to calculate whether and under what conditions dispositions to such beliefs would be evolutionarily stable. Models are very clear that Machiavellian intelligence—cheat when you can, cooperate when you must—is highly adaptive in some circumstances, and that unconditional cooperation is virtually never adaptive. Moreover, few people exhibit

the latter and many exhibit the former. Thus, if humans have evolved to believe in categorical moral demand and anticipate inevitable punishment for norm violations and rewards for moral virtue—solely as an adaptive response to biased risks in behavioral choices—then there must also have evolved cognitive biases against holding these beliefs in a way that consistently shapes behavior.

Meta-ethical Claims

Much of the recent philosophical literature on evolutionary ethics involves arguments and counter arguments over whether and how ethical norms may be justified by, derived from, or informed by the "facts" of morality's origin or humanity's nature as illuminated by evolution. In this section, I want to comment briefly on one aspect of this discussion: prominent claims in the scholarly, interdisciplinary, or public intellectual literature, which argue (or merely assert) that evolutionary explanations of morality reveal moral beliefs to be fundamentally illusory in one or more ways.

Morality and God

First, and briefly, a number of prominent evolutionary psychologists (and creationists) claim that scientific findings subvert the idea that ethics may be grounded in God. In his essay, "The Moral Instinct"[92] (reprinted in this volume), Steven Pinker says, "The scientific outlook has taught us that some parts of our subjective experience are products of our biological makeup and have no objective counterpart in the world." Setting aside the question of whether or in what way this is true, and if it is, whether science has taught us this, Pinker then suggests that this finding of science leads to the question of whether or not morality "is just a collective illusion." One move might be to believe that ethical reality is grounded in something beyond the world, for example, in God. But Pinker points out—citing the Euthyphro dialogue—"Plato made short work of it 2,400 years ago." To be fair, Pinker is not claiming that science settles the issue (Hauser and Peter Singer do, below). He is claiming that science raises the issue, which Plato settled, 2,400 years before it was raised by science. But the reason for cultural receptivity to science's raising the question is that there have been 2,400 years of unresolved philosophical discussion since Plato, including claims that moral propositions to which God assents are, contra Euthyphro, neither a matter of divine whim nor a matter of moral principles to which God submits and to which we therefore ought to submit independent

of God, but are matters grounded in God's moral nature.[93] I am not presuming to comment on this issue, other than to say it is not clear how biology does anything to illuminate it further.

Marc Hauser, whose influential proposal for a universal moral grammar I cited earlier as a critique of Jonathan Haidt's emotivism, has published several essays with Peter Singer, which argue more explicitly that evolutionary accounts of moral cognition do not just illuminate, but settle the question of God's role in human ethical judgment.

> One view is that a divine creator handed us the universal bits [of morality] at the moment of creation. The alternative, consistent with the facts of biology and geology, is that we have evolved, over millions of years, a moral faculty that generates intuitions about right and wrong. For the first time, research in the cognitive sciences . . . has made it possible to resolve the ancient dispute about the origin and nature of morality.[94]

How is this dispute resolved? Hauser's work analyzing computer-gathered responses to moral dilemmas has resulted in the significant but not altogether surprising discovery that people all over the world, regardless of cultural or religious background, share certain kinds of dispositions toward similar moral judgments. For example, they prefer not to push people into the path of trains, even if it will save others. Thus, moral judgment is not rooted in God's activity, because "If morality is God's word, atheists should judge these cases differently from religious people."[95] The explicit claim is that after centuries of debate, we have determined that morality does not come from God, because people the world over, whether or not they believe in God, prefer not to push people onto the tracks in front of trains.

There are two problems with this assertion, one obvious and the other less obvious. The obvious one is that the reported results do not justify the conclusion: they do not, in fact, constitute a challenge, or even a surprise, to those who believe ethics are grounded in God's nature and initiative. For one thing, the dichotomy between evolution and divine instantiation of morality that the authors initially pose is a false one: God could well have instantiated the "bits" of moral information into the structure of nature "at the moment of creation," and this structure could have informed the evolution of "a moral faculty that generates intuitions about right and wrong." (Indeed, under these circumstances such intuitions could even have the status of knowledge.) But let's imagine the dichotomy is legitimate, as they pose and

as understood by some who believe that the fundamental "elements of morality" (a term used by Hauser and Singer in a variant of their essay) did not evolve, but are supernaturally instantiated by God into the first or into each human "at the moment of creation" via a process of ensoulment or divine illumination. Even in this case of clearly opposing naturalistic and supernaturalistic proposals for the origin of moral judgment, the cited empirical findings do nothing to adjudicate between the two. In either situation, human beings would be expected to evidence commonalities of moral judgment—as we have known exist long before computerized administrations of trolley car dilemmas and as are affirmed by ancient theological concepts of common grace and philosophical notions of a shared moral telos for humanity.

But the second problem is this: the moral judgments, or what many would view as the important, morally salient components of these judgments, are actually not uniform, and their variability is not unrelated to religious or cultural background. Hauser has found that when faced with a train speeding down the tracks on the way to killing five people, most people would *push a button* that sends the train down an alternative track where it kills one person. But they would not *push a person* into the way of the train, even if the body would impede the train and thus save the other five. Across cultures, the overwhelming majority of people would indirectly cause the death of their "neighbor" to save others, but would not by direct physical contact cause the death of their neighbor for similar ends. However, a fundamental moral question not addressed in this work is: "who is my neighbor?" What if the person to be pushed is (or people to be saved are) viewed as an out-group member: Arab, or black, or Jew, or homosexual, or atheist, or fundamentalist Christian, or someone who has personally wronged the actor?

The answer to the profoundly important question of who falls within the domain of moral concern is neither part of a universal moral grammar nor determined by innate moral sentiment. It is a function of culturally variable moral teaching and is in fact known to be influenced by a variety of social factors and also by religious commitment.[96] If there is a God who reveals moral truths (especially if they are not fully deposited at "the moment of creation" but revealed or learned progressively as many theistic traditions affirm), then perhaps Hauser and Singer are right to expect to see, in such a case, differences between various religious and non-religious traditions. But this is precisely what we do see. Some traditions do not and some do evidence an "expanding circle"

(to use Singer's own phrase) of moral concern, progressively restricting revenge (from seven-fold, to eye for eye, to no revenge) and enlarging care from kinsman, to neighbor, to nation and household of faith, and to enemy. The empathic sentiments underlying and the grammar that communicates the nature of moral concern may be universal; the construal of who falls within the circle of concern is not. Each may be amenable to the logic of evolutionary analysis;[97] but such analysis does not have straightforward implications for meta-ethics.

Moral Relativism

> "Human beings function better if they are deceived by their genes into thinking that there is a disinterested objective morality binding upon them, which all should obey . . . ethics as we understand it is an illusion fobbed off on us by our genes to get us to cooperate."
>
> – Michael Ruse and E. O. Wilson.[98]

> "morality is an illusion foisted upon us by evolution . . . [yet] we may be able usefully to employ a moral discourse, warts and all, without believing in it."
>
> – Richard Joyce.[99]

> . . . the science of the moral sense can advance it, by allowing us to see through the illusions that evolution and culture have saddled us with . . .
>
> – Steven Pinker[100]

Two claims that have been widely made about the implications of an evolutionary explanation for morality are that (a) the sense of moral absolutes is illusory, since any moral norms are relative to the fitness-enhancing requirements of our particular biology and (b) even if there are moral truths, we have no justification for believing we can know them.

Following Darwin, Michael Ruse develops arguments for the illusory character of our belief in specific moral absolutes, which goes like this: Imagine you were a moral creature with an altogether different ecology than humans, perhaps akin to that of the black widow spider. Then, you would consider it moral to kill and eat your mate in order to obtain sufficient nurture for yourself and for the care of young.[101]

There are two problems with this. The first one is that the "moral spider" counterfactual may in principle be as impossible for biology as a square circle is for geometry. This is because certain very specific biological characteristics and social ecologies may be necessary for

morality to emerge at all. Darwin speculated that we would get morality in any creature with high intelligence and well-developed social instincts, including parental and filial affections. This insight negates Ruse's essentially dualist suggestion that we might get it in a creature with the black widow's biology—no sociality, no parental attachment, and little intelligence. But Darwin's initial insight understates the biological requirements that subsequent to his work, we have discovered are linked with the requisite levels of sociality, parental and filial bonds, and intelligence we find in humans. These include life expectancy, mortality rate, fertility rate, body size, and the relationship between degree of infant dependence, parental care, lifelong pair bonding, group hunting, and even bipedal gate (which modified the pelvis resulting in increased dependence, care, and pair bonds). All of these factors synergistically contribute to intelligence, sociality, and reproductive bonds as an adaptive suite. Thus, it is not at all clear that evolution could produce "morality" in an animal with a radically different ecology. It certainly could not produce it in a spider, or (contrary to the suggestion of Simon Conway Morris) a reptile. If it did, the reptile would have to be homoeothermic, viviparous, and intensively young-nurturing, highly social, pair bonding, and have low fecundity—in short, not be a reptile.

The other problem is this: the truth of specific moral norms that apply to humans is not subverted by considering the counterfactual that under different circumstances there might be different specific norms. For example, imagining a radically different natural history, where the first human was created directly from dust and the second human from a rib of the first, or a human being generated by in vitro fertilization and raised in a collective with no father and mother and therefore no obligation to honor them, would not challenge the moral justification of honoring parents or ancestors, nor would it underwrite the notion that moral norms were illusory. Moral standards may apply with legitimate normative force to a presently universal but temporary situation, or they may entail conditional formulations of an unconditional general principle.

Moral Skepticism

Alex Rosenberg and Richard Joyce argue on similar grounds that an evolutionary account of morality does not actually entail that there are no moral truths, but it does mean that we have no warrant for holding particular beliefs to be true. Michael Ruse points out that if believing

a particular moral proposition enhanced fitness, then people would believe it even in a world where there were no moral reality. Rosenberg formalizes it like this:

> "To turn the Darwinian explanation [of morality] into an 'explaining away' the Nihilist need only add the uncontroversial scientific principle that if our best theory of why people believe P does not require that P is true, then there are no grounds to believe P is true."[102]

Evolution may indeed raise questions about the reliability of belief-forming cognitive mechanisms,[103] but as developed here, there are at least two problems. First, what would seem to be needed to argue in this direction is not just that we have a best theory, but that we have an adequate theory to explain belief P in terms that are indifferent to the truth of P, or at least a plausible theory. But what if our best theory is conceptually inadequate (as I have argued evolutionary accounts of morality are)? Or what if it is adequate in principle, but the likelihood of its being true in light of the empirical data, or other things we believe, is vanishingly small? Or what if it is the "best" theory in virtue of being slightly better than individual alternatives (though not better than their disjunction), or in virtue of being the only theory at all that conforms to the epistemic standards of naturalistic science? None of these situations would seem to raise serious worries about grounds for believing P to be true.

Second, say we have an adequate and well-supported theory for why someone believes P—for example, the belief P that the earth is flat—that does not entail the truth of P. Rather, it entails the truth of other things (the presence of light, of eyes, of spatial cognition) that under particular circumstances (living in a rain forest, without access to visual horizons, and prior to the advent of geometry) reliably lead to belief P, whether or not it is true. This does not mean that under such circumstances, there are "no grounds to believe P is true." It is not the case that there are no grounds to believe anything that, with further evidence, turns out not to be less credible than alternative proposals.

Richard Joyce formulates the argument with a counterfactual. He suggests that we imagine a belief pill that is known to cause anyone who takes it to believe that Napoleon lost at Waterloo. If you were fed this pill in infant formula at birth, then although Napoleon did lose at Waterloo and there is abundant evidence to demonstrate this, you

would have no grounds for confidence that your belief is true. Joyce maintains that natural selection is essentially a "belief pill" for certain moral beliefs about cooperation, which might actually be morally true, but which, having taken the pill, we have no grounds to believe: "knowledge that your belief is the product of a belief pill renders the belief unjustified . . ."[104]

One problem of course is that natural selection could be viewed as a pill for inducing belief in *anything*—including the deliverances of science itself—so long as the belief or the inclination to form beliefs of particular kind generates, on average, fitness-enhancing behaviors. Joyce recognizes this challenge and suggests that ". . . we have no grasp of how any innate human faculties pertaining to 'scientific inquiry' might have been selected for independently of their producing judgments that at least have some positive connection to the truth. Thus the 'evolutionary debunking of morality' does not in this manner debunk itself."[105] I suppose technically this qualifies as an argument from ignorance: perhaps we don't know how natural selection produces judgments that are not positively connected with truth. But we don't have a clear idea of how it would reliably produce judgments that are. What we do know is this: natural selection must produce behaviors that are isomorphically related to reality and to the extent that beliefs produce behaviors (this is arguable), then beliefs must *reliably covary* with truth. But they need not "positively" covary. With any given state of affairs, representational beliefs could depart significantly, and in different directions than and to different degrees than the relationship between beliefs and reality in other conditions. All that is necessary is that the system mediating beliefs and behavioral motivation be able to decode the function that specifies the relationship between beliefs and truth. Beliefs may entail massive deviations from the truth that are completely opaque to the believing agent but deviate in ways that—though consciously unrecognized—are sufficiently regular to inform adaptive behavior.

Rosenberg has recognized this and argues that we can sustain our confidence evolution that produces true beliefs if we attenuate notions of what constitutes "true belief" to include any conceptual representation of reality that reliably changes when reality changes. This could be the case even if conceptual schema are arbitrary, differing conceptualizations are logically incompatible, and no given representation tracks reality with complete fidelity. However, by this standard, it is not clear that the belief pill constitutes a threat to moral justification, so long as

the effect of the pill on beliefs is not random with respect to but is in some way connected to an aspect of reality. Since the "pill" is natural selection, which, by definition, entails the concordant structuring of phenotypes to the natural world, then any beliefs caused by selection track the world with some fidelity and are by this standard "true."

Lastly, we do not, in any case, have an account of natural selection that is equivalent to a pill for certain moral beliefs. For one thing, we don't even have a selectionist explanation of any kind, for why people have beliefs of the morally fundamental kind that motivate radical altruism. But even if we did, evolutionary accounts of moral beliefs do not explain them "all the way down," in a way that is indifferent to whether or not they are true. Some kinds of beliefs may require that the belief be true as a precondition for having the belief: for example, any explanation of my belief, P, that I was born, requires as part of the explanation that P be true. Others kinds of belief may contingently follow from the truth of the belief. For example, if on the basis of credible first person reports, and analysis of pharmaceutical ingredients in the pill, and ethnobotanical studies demonstrating these ingredients were found only on the island of St. Helena in the early nineteenth century, it is known that Napoleon invented, while in exile, a pill causing the belief that he lost Waterloo, then having consumed the pill does not defeat my belief about the outcome of Waterloo.

Joyce actually acknowledges this point, but dismisses claims that there are plausible reasons for believing moral truths to be involved in the causal chain leading to those beliefs. Yet, to explain particular moral beliefs—for example, a belief that it is morally obligatory to forgive others as I have been forgiven by God—in a way that does not entail the truth of that belief, would involve having an adequate explanation not only for minds' characteristic inclination to hold the belief, but also for the origin of minds that hold it, and laws and local conditions that facilitated the development of minds from monads, and a universe with characteristics that permitted the rise of life, and something rather than nothing at all—which does not require there to be or plausibly suggest there is a God or a moral structure to the cosmos that influenced its unfolding. Whether or not we have such an account, and if so, whether it constitutes "our best theory" is an issue that many view as settled. However—and I make this observation as an agnostic on this question—those who profess the greatest surety that it *has* been settled, do not seem to agree on the resolution.

Acknowledgments

I would like to thank Robert Audi, Celia Deane-Drummond, Michael Murray, Susan Neiman, Alvin Plantinga, and Michael Rea for their thoughtful comments on this manuscript, as well as colleagues in the Notre Dame Center for Philosophy of Religion and Harvard University Program in Evolutionary Dynamics.

Notes

1. Wilson, E. O., *Sociobiology: The New Synthesis* (Harvard, 1975), 3 and 562.
2. De Waal, F., *Good Natured: The Origins of Right and Wrong in Humans and Other Animals* (Harvard, 1997), 218.
3. *Economist* 386, no. 8568 (2/23/2008): 98–98.
4. Brooks, D. "The End of Philosophy" *New York Times*, April 7, 2009; Section A; Column 0, 29.
5. For discussions of these distinctions, see for example: Schloss, "Evolutionary Theories of Morality: Surveying the Landscape," In *Evolution and Ethics: Human Morality in Biological and Religious Perspective*. P Clayton and J Schloss eds, (Eerdmans, 2004); Kitcher, P, "Four Ways of Biologizing Ethics," in *In Mendel's Mirror: Philosophical Reflections on Biology* (Oxford, 2003); Fitzpatrick, W. "Morality and Evolutionary Biology" *Stanford Encyclopedia of Philosophy*, 2008.
6. Wilson, E. O. *On Human Nature* (Harvard University Press, 1978), 167.
7. Genetic lag proposals hold that morality *may have served* such a function in ancestral environments, but it no longer does. Spandrel accounts view morality as having no adaptive function at all, being a byproduct of other traits that do serve reproduction (Ayala, F., "The difference of being human: Ethical behavior as an evolutionary byproduct," In *Biology, Ethics, and the Origins of Life*, ed. H. Rolston III (Boston, 1995), 113–36: Jones & Bartlett; Gould & Lewontin, "The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme," *Proceedings of the Royal Society of London* 205 (1979): 581–98). Gene-culture coevolutionary proposals construe "the kernel of internalized morality" as entailing cultural transmission that may co-opt and even oppose genetic replication so that the morally infected are "led to subordinate their genetic fitness (and their self-interest in general) to the fitness and interests of others" (Lopreato, *Human nature and biocultural evolution* (Boston: Allen & Unwin, 1984), 234; Durham, *Coevolution: Genes, culture, and human diversity* (Stanford, CA: Stanford University Press, 1992)).
8. De Waal, F., *Good Natured: The Origins of Right and Wrong in Humans and Other Animals* (Harvard University Press, 1997), 218.
9. Gould, Stephen J., *The Structure of Evolutionary Theory* (Cambridge MA: Harvard University Press, 2002).
10. Schloss, op cit.; also Kitcher, op cit.; Joyce, R., *The Evolution of Morality* (MIT, 2006); Hauser, M., *Moral Minds* (Harper, 2006); Katz, L., ed. *Evolutionary Origins of Morality* (Imprint Academic, 2000).
11. De Waal, F., *Good Natured: the Animal Origins of Right and Wrong* (Harvard, 1997); Flack, J., and F. de Waal, "'Any Animal Whatever': Darwinian

Building Blocks of Morality in Monkeys and Apes," In Katz, op cit., 1–30.

12. de Waal, F., *Primates and Philosophers: How Morality Evolved* (Princeton, 2006).
13. de Waal, Frans, "Obviously, Says the Monkey," In Does Evolution Explain Human Nature? (Templeton Press, 2009), Available online, http://www.templeton.org/evolution/ Also *Primates and Philosophers*, op cit.
14. "Whereas Veneer Theory, with its emphasis on human uniqueness, would predict that moral problem solving is assigned to evolutionarily recent additions to our brain, such as the prefrontal cortex . . . neuroscience seems to be lending support to human morality as evolutionarily anchored in mammalian sociality. We celebrate rationality, but when push comes to shove we assign it little weight. This is especially true in the moral domain.' de Waal, 2006, op cit., 55–56.
15. This claim, along with the above suggestion that we can not only infer, but somehow *observe* there to be no differences between human and animal "basic wants and needs," has generated criticisms of anthropomorphic imputation of human intentions and desires upon the animal behavioral data, for example, Wright, R., "The Uses of Anthropomorphism" in *Primates and Philosophers* (2006), de Waal, op cit.
16. de Waal, 2006, op cit.
17. Lieberman, D., J. Tooby, and L. Cosmides, "Does Morality have a Biological Basis? An Empirical Test of the Factors Governing Moral Sentiments Relating to Incest," *Proc. Biol. Sci.* 270, no. 1517 (2003): 819–26. Also Borg, J. D. Lieberman, and K. Kiehl, "Infection, Incest, and Iniquity: Investigating the Neural Correlates of Disgust and Morality," *Journal of Cognitive Neuroscience* 20, no. 9 (2008): 1529–46.
18. Insel, T. and Young, L., "The Neurobiology of Attachment," *Nature Reviews, Neuroscience* 2 (2001): 129–36.
19. Lieberman, op cit; Haidt, J., "The New Synthesis in Moral Psychology," *Science* 316 (2007): 998–1002; Wheatley, T., and Haidt, J., "Hypnotically Induced Disgust Makes Moral Judgments More Severe," *Psychological Science* 16 (2005): 780–84.
20. Maynard Smith, J., and Szathmary, E., *The Major Transitions of Evolution* (Oxford University Press, 1995).
21. Hume, D., *Enquiries Concerning Human Understanding and Concerning the Principles of Morals* (Oxford: Clarendon Press, 1975).
22. Binmore, K., *Game Theory and the Social Contract* (MIT Press, 1994); Skyrms, B., *Evolution and the Social Contract* (Cambridge University Press, 1996).
23. Axelrod, R., The Evolution of Cooperation (Basic Books, New York, 1984).
24. Nowak, M., and Sigmund, K., "Chaos and the Evolution of Cooperation," *Proc. Natl. Acad. Sci. USA* 90 (1993): 5091–94. Also, Nowak and Sigmund, "A Strategy of Win-Stay Lose-Shift that Outperforms Tit-For-Tat in the Prisoner's Dilemma Game," *Nature* 364 (1993): 56–58.
25. Across successive rounds of play, the dominant strategy evolves from generous TFT ("forgiving" one defection) to progressively more generous, to finally, when all actors are very generous, unconditional cooperation. But at this moment, the population is completely vulnerable to infection by

the strategy of wholesale defection, and it reverts to standard TFT as the dominant strategy. The cycle then begins anew.

26. Maynard Smith, and Szathmary, above; Michod, R., *Evolutionary Transitions in Fitness and Individuality* (Princeton University Press, 2000).
27. Justin D. Arms., "When Evolutionary Game Theory Explains Morality, What Does it Explain?" in *Evolutionary Origins of Morality*, ed. L. Katz (Imprint Academic, 2000), 296–99.
28. Boyd, R., and Richerson P., "Punishment Allows the Evolution of Cooperation (or anything else) in Sizable Groups," *Ethol Sociobiol* 13 (1992): 171–95; Rockenback B., and Milinski, M., "The Efficient Interaction of Indirect Reciprocity and Costly Punishment," *Nature* 444 (2006): 718–23.
29. de Quervain et al., "The Neural Basis of Altruistic Punishment," *Science* 305 (2007): 1254–58.
30. Ohtsuki, Iwasa, and Nowak, "Indirect Reciprocity Provides Only a Narrow Margin of Efficiency for Costly Punishment," *Nature* 457 (2009): 79–82; Fehr and Rockenbach, "Detrimental Effects of Sanctions on Human Altruism," *Nature* 422 (2003): 137–40.
31. E. O. Wilson, *Sociobiology: The New Synthesis* (Harvard University Press, 1975), 578.
32. Alexander, R., *The Biology of Moral Systems* (Aldine de Gruter, 1987); Holcomb, H., *Sociobiology, sex, and science* (SUNY Press, 1993).
33. Hamilton, W., "The Genetical Evolution of Social Behavior," *Journal of Theoretical Biology* 7 (1964): 1–16.
34. Trivers, R., "The Evolution of Reciprocal Altruism," *Quarterly Review of Biology* 46 (1971): 35–39.
35. Alexander, R., *The Biology of Moral Systems* (Chicago, IL: Aldine de Gruyter, 1987).
36. Sober, E., and Wilson, D. S., *Unto Others: The Evolution and Psychology of Unselfish Behavior* (Cambridge, MA: Harvard University Press, 1998).
37. With respect to fitness consequences, not psychological motives.
38. Zinser, J., "Altruism," in *Evolution: The First Four Billion Years*, eds. M. Ruse, and J. Travis (Harvard University Press, 2009), 407–10.
39. "Opposition" is one of five modes of gene-culture interaction proposed by William Durham, *Coevolution: Genes, Culture, And Human Diversity* (Stanford University Press, 1991).
40. Henry Plotkin concludes the following about humans: "the only explanation is that culture entails causal mechanisms that are somehow decoupled, not necessarily completely, but some partial decoupling is necessary, from the causal mechanisms of our biological evolution. It certainly cannot be that culture is tightly held on some biological 'leash'." Plotkin, H., *Evolution In Mind: An Introduction To Evolutionary Psychology* (Harvard University Press, 1997), 231.
41. Dawkins, R., *The Selfish Gene* (Oxford University Press, 2006), 200–201.
42. Stent, G., *Morality as a Biological Phenomenon* (University of California Press, 1981).
43. de Waal, *Good Natured*, op cit.
44. Hare, J., "Is There an Evolutionary Foundation for Human Morality?" in Clayton and Schloss, op cit, 2004, 187–203. Also, *The Moral Gap: Kantian Ethics, Human Limits, and God's Assistance* (Oxford, 1997).

45. This problem is true of the most reductive versions of memetics. For considerably more rigorous and nuanced proposals of coevolution, see, for example, Richerson and Boyd, *Not by Genes Alone: How Culture Transformed Human Evolution* (University of Chicago Press, 2006).
46. Coyne, "The Self-Centered Meme," *Nature* 398 (1999): 767–68.
47. There are numerous game-theoretic proposals for the spread and dominance of various cooperative strategies transmitted by generic replicators—genetic or cultural. But not altruistic (replication-subverting) strategies.
48. Bell, Richerson, and McElreath, "Culture Rather than Genes Provides Greater Scope for the Evolution of Large-Scale Human Prosociality," *PNAS* 106, no. 42 (2009): 17671–74.
49. de Waal, 2006, op cit. 23.
50. Monroe, K., *The Heart Of Altruism: Perceptions Of A Common Humanity* (Princeton University Press, 1996); Oliner, S., *Altruistic Personality: Rescuers of Jews in Nazi Europe* (1992).
51. Wilson, 1975 op cit. 382.
52. Ayala, op cit.
53. Lieberman, op cit.
54. Marc Bekoff, and John Byers, eds., *Animal Play: Evolutionary, Comparative, and Ecological Perspectives* (Cambridge University Press, 1998).
55. Boehm, C., *Hierarchy in the Forst: The Evolution of Egalitarian Behavior* (Harvard University Press, 2001).
56. McKay D., and D Dennett, "The Evolution of Misbelief," *Behavioral and Brain Sciences* 32 (2009): 493–510.
57. Johnson, D., *Overconfidence and War: The Havoc and Glory of Positive Illusions* (Harvard University Press, 2004). Also Johnson, D., and J. Fowler, "The Evolution of Over Confidence," *Quantitative Biology* (2009), http://arxiv.org/abs/0909.4043v1.
58. Schloss, J., and Murray, M., "Evolutionary Theories of Supernatural Punishment: A Critical Review," *Religion, Brain, and Behavior* 1, no. 1 (2011): 4–27.
59. Of course it is not clear that either of these adaptationist stories underwriting the supposed innateness of out-group repugnance is even true.
60. Kass, L., "The Wisdom of Repugnance," *New Republic* 216, no. 22 (1997): 17–26.
61. E.g., Steven Pinker, "The Moral Instinct," *New York Times*, 2008; or Art Caplan's characterization of Kass as "Dr. Yuck."
62. The derogation of moral sentiment seems to have played an important role in the eugenic programs of the last century. Even in America, Madison Grant pre-emptively dismissed resistance to his proposal for the "elimination of defective infants" and "obliteration of the unfit" as being founded on "a sentimental belief in the sanctity of human life" (Grant, M., *The Passing of the Great Race* [Charles Scribners, 1918], 49). Nor does appeal to reason or moral principles in itself solve this risk. As recent discussions of ethical universalism versus particularism suggest, though we may assess particular judgments in light of universals, we may also assess proposals for universals by seeing where they lead, and if they lead to conflict with a deeply held particular or to what we believe to be a "violation of things we rightfully hold dear"—if a moral principle allows for the Holocaust—so much the worse for the principle.

63. Frankfurt, Harry G., "Freedom of the Will and the Concept of a Person," In *The Importance of What We Care About* (New York: Cambridge University Press, 1988), 11–25.
64. Kitcher, P., "Ethics and Evolution: How to Get There from Here," In de Waal, 2006, op cit.
65. Kitcher, P., 2006, op cit.
66. Darwin, C., *Descent of Man and Selection in Relation to Sex* (Princeton, 1981), 88–89.
67. Kagan, J., "The Uniquely Human in Human Nature," *Daedalus*, Fall (2004): 77–88. "The concern with right and wrong, the control of guilt, and the desire to feel virtuous are, like the appearance of milk in mammalian mothers, a unique event that was discontinuous with what was prior . . . The continual desire to regard the self as good is a unique feature of *Homo sapiens*. Although it has a firm foundation in the human genome, it is not an obvious derivative of the competences of apes and monkeys."
68. Wilson, M., and M. Daly., "Do Pretty Women Inspire Men to Discount the Future?" *Proc Royal Soc Lon B* 271, no. 4 (2004): S177–S179.
69. Alexander, R., 1987. Op cit.
70. Harvard biologist, David Haig, quoted in Karl Sigmund, *The Calculus of Selfishness* (Princeton University Press, 2010), 10.
71. Hume, D., *An Enquiry Concerning the Principles of Morals* (Oxford University Press / Kessinger Publishing, 2004(1751)), 83.
72. Zahavi, A., *The Handicap Principle: A Missing Piece of Darwin's Puzzle* (Oxford, 1999).
73. Miller, G., "Sexual Selection and Moral Virtues," *The Quarterly Review of Biology* 82, no. 2 (2007): 97–125. Also, Miller, G., "Kindness, Fidelity, and Other Sexually Selected Virtues," In *The Evolution of Morality: Adaptations and Innateness*, ed. Walter Sinnott-Armstrong (MIT Press, 2007), 209–44.
74. Konner, M., *Why the Reckless Survive . . . and Other Secrets of Human Nature* (Penguin, 1991); Miller, G., *The Mating Mind: How Sexual Selection Shaped the Evolution of Human Nature* (Anchor, 2001).
Ridley, M., *The Red Queen: Sex and the Evolution of Human Nature* (Harper, 2003).
75. Subiaul, et al., "Do chimpanzees learn reputation by observation?" *Animal Cognition* 11 (2008): 611–23; Shary and D'Souza, "Cooperation in communication networks: indirect reciprocity in interactions between cleaner fish and client reef fish," *Animal Communication Networks* 22 (2005): 321–530.
76. Schloss, 2004 op cit. Joyce, R., *The Myth of Morality* (Cambridge, 2001). "They give good instrumental reasons for acting in a cooperative manner . . . On the face of it, they don't appear to give any reason to cultivate *moral beliefs*. ..On the contrary, such reasoning promises to lay the foundation for a cooperative society that has done away with moral thinking altogether . . ." 210.
77. See Monroe, 1996 and Oliner, 1992 (op cit.) for extensive descriptions of such behaviors. I realize that citing them may conflate the issues of moral sentiments and moral beliefs in conscience, though: a) in both of these studies, moral heroism by virtually all subjects was described as entailing

no analysis or even reflective beliefs, but an overwhelming feeling of what must be done and b) whatever the role of sentiment and belief, the data indicate that "conscience as reputation maximizer" is inadequate.

78. Simon, H., "A Mechanism for Social Selection and Successful Altruism," *Science* 250, no. 4988 (1990): 1665–1668.
79. Frank, R., *Passions Within Reason: The Strategic Role of the Emotions* (Norton, 1988).
80. Recent empirical studies have shown that human beings are indeed very proficient at recognizing defectors from facial cues and at recognizing generous cooperators. E.g., Eckman, P., *Darwin and Facial Expression: A Century of Research* (Malor, 2006).
81. Frank's proposal, by the way, provides an alternative to the costly signaling view of art. It affirms what we all know to be true: that it is pursued with motives that go beyond, and under situations where the investment of time or money greatly exceeds, the benefits of social prestige. Yet, it still may yield prestige, most emphatically where it is perceived to have been pursued "for arts' sake."
82. Enumerating family, friends, and reputation as having a transparent connection to fitness is not to say that valuing them is transparently selfish. These things may be viewed as genuine goods, whose cultivation may be rightly pursued as a moral end. The point is there is no reason that they need to be pursued in this way, to end up generating fitness returns.
83. Frank's proposal that utilities do not transparently collapse into each other is not merely speculative. In addition to the observations of moral heroism and vengeful feuding cited above, a number of recent experimental studies have demonstrated that humans pursue goals like fairness or justice at cost to themselves (see recent seminal papers by experimental economist, e.g., Ernst Fehr and colleagues, "Altruistic Punishment in Humans," *Nature* 415 [2002]: 137–40). However, these studies have also attracted criticism for flawed methodology (e.g., actual stakes too low) and the inability to reject alternative interpretations such as aggression against dominants (e.g., Fowler et al., *Nature* 433 [2002]: E1) or pursuit of reputational gain.
84. E.g., Trivers, R., "Deceit and Self-Deception," Forthcoming in *Mind the Gap*, eds. Kappaler and Silk (Springer, 2010), 373–93. Also "The Elements of a Scientific Theory of Self-Deception," *Annals New York Academy of Sciences* (2000): 114–31.
85. The idea of "self-deception"—that somehow a self can lie to itself—is a difficult issue that evolutionary biologists seem to assert, without having thought much about what it means or how it could occur. A more cautious, if still somewhat ambiguous, approach used in some recent cognitive neuroscience work involves the notion of incomplete information transfer between brain regions or biased communication between cognitive modules.
86. The thirty-year sociobiological tradition (consistent with the older Hobbesian tradition) has an evident commitment to the latter as the primary explanation: in Michael Ghiselin's famous mantra of sociobiology, "scratch an altruist, watch a hypocrite bleed." Ghiselin, M. T., *The Economy Of Nature And The Evolution Of Sex* (Berkeley: University of California Press, 1974), 247.
87. Wilson, E. O. as cited in chapter head.

88. There is an alternative explanation for beliefs that appear structured by fitness-related challenges, which entails neither innate dispositions nor development. Rather, some beliefs may be generated by random processes of cultural innovation and then are selectively transmitted and retained, based on contributions to fitness. The fact that many societies have converged on diets that include plants with balanced amino acids (e.g., corn and rice, beans and rice) may be an example of this phenomenon. Beliefs arising by this process would "back-up" native dispositions or structure behaviors where no dispositions existed.
89. Johnson, D. D. P., and Kruger, O., "The good of wrath: supernatural punishment and the evolution of cooperation," *Political Theology* 5, no. 2 (2004): 159–76.; Johnson, Dominic and Bering, Jesse, "Hand of God, mind of man: punishment and cognition in the evolution of cooperation," *Evolutionary Psychology* 4 (2006): 219–33.
90. Johnson, D., and J. Fowler, "The Evolution of Over Confidence," *Quantitative Biology* (2009). http://arxiv.org/abs/0909.4043v1.
91. Schloss, J., and M. Murray, "Evolutionary Accounts of Belief in Supernatural Punishment: A Critical Review," *Religion, Brain, and Behavior* 1, no. 1 (2011): 46–66. Also 7 responses and authors reply in same volume: 66–95.
92. *New York Times*, January 13, 2008.
93. E.g., Robert M. Adams, "A Modified Divine Command Theory of Ethical Wrongness," In *The Virtue of Faith and Other Essays in Philosophical Theology* (New York: Cambridge University Press, 1987). Also, "Moral Arguments for God."
94. Hauser, M., and P. Singer, "Godless Morality," Project Syndicate, 2006. http://www.project-syndicate.org/commentary/hausersinger1; Also, "Morality without religion," *Free Inquiry*, December 2005.
95. Ibid.
96. Though influenced by religious commitment, the direction of influence is complicated. Social psychological studies indicate that religious fundamentalism is correlated with racial prejudice; studies of Holocaust rescuers indicate that rescuing was correlated with religious belief.
97. There are various proposals for this expanding circle of moral concern in light of evolutionary dynamics (e.g., Robert Wright, *Non-Zero: The Logic of Human Destiny* (Vintage Press, 2001)); Elliot Sober, and David Sloan Wilson, *Unto Others: The Biology and Psychology of Unselfish Behavior* (Harvard University Press, 1999)). And several recent meta-analyses of evolutionary history suggest a trend of increasing cooperative interdependence characterizes major transitions in evolution (John Maynard Smith, and Eors Szathmary, *The Major Transitions of Evolution* (Oxford University Press, 1998); Richard Michod, *Evolutionary Dynamics*: *Evolutionary Transitions in Fitness and Individuality* (Princeton University Press, 2000)).
98. "Evolution of Ethics," *New Scientist* 17 (1989): 51; Also Ruse and Wilson, "Moral Philosophy as Applied Science," *Philosophy* 61 (1986): 173–92; Ruse, *Taking Darwin Seriously*, (Blackwell, 1986).
99. "Darwinian Ethics and Error," *Biology and Philosophy* 15, no. 5 (1989): 713–32, 713. Also *The Myth of Morality* (Cambridge University Press, 2001); *The Evolution of Morality* (MIT Press, 2006).
100. Pinker, S., 2008, op cit.

101. Darwin expresses a very similar worry that what we believe to be moral is based on contingencies of our ecology and natural history. Therefore, morality is relative, though he does not take the Rusean next step of concluding that belief in the truth of particular moral norms is illusory.
102. Sommers, T., and A. Rosenberg, "Darwin's Nihilistic Idea: Evolution and the Meaninglessness of Life," *Biology and Philosophy* 18 (2003): 653–68, 667.
103. Plantinga, A., and M. Tooley, *Knowledge of God* (Blackwell Publishing, 2008); Schloss, J., and M. Murray, "Evolution and True Beliefs: You Can't Always Get What You Want," *Brain and Behavioral Sciences* 32, no. 6 (2009): 533–34
104. Joyce, 2006, op cit. 180.
105. Ibid, 183.