Reflections on the Evolution of Morality

Christine M. Korsgaard

Harvard University[1]

> All instincts that do not discharge themselves outwardly turn inward – this is what I call the *internalization* of man: thus it was that man first developed what was later called his "soul." The entire inner world, originally as thin as if it were stretched between two membranes, expanded and extended itself, acquiring depth, breadth, and height, in the same measure as outward discharge was inhibited.
>
> - Nietzsche

1. Introduction

In recent years there has been a fair amount of speculation about the evolution of morality, among scientists and philosophers alike. From both points of view, the question how our moral nature might have evolved is interesting because morality is one of the traditional candidates for a distinctively human attribute, something that makes us different from the other animals. From a scientific point of view, it matters whether there are any such attributes because of the special burden they seem to place on the theory of evolution. Beginning with Darwin's own efforts in *The Descent of Man*, defenders of the theory of evolution have tried to show either that there are no genuinely distinctive human attributes – that is, that any differences between human beings and the other animals are a matter of degree – or that apparently distinctive human attributes can be explained in terms of the

---

[1] Notes on this version are incomplete. Recommended supplemental reading: "The Activity of Reason," APA Proceedings November 09, pp. 30-38. In a way, these papers are companion pieces, or at least their final sections are.

interaction between other attributes that *are* matters of degree. Darwin's own account of the evolution of morality, which I will be discussing later, is of this second kind.[2]

From a philosophical point of view, of course, understanding the ways we are different from the other animals is one way of understanding ourselves. And although it is a little obscure exactly how it works, one of the traditional modes of philosophical understanding, especially of morality, is the origin story: think, for instance, of the accounts of morality that we find in Hobbes, or Nietzsche, or Rousseau. All of these thinkers try to throw light on what it means to be human by telling us stories about how moral motives, emotions, or even obligations themselves might have emerged from events or processes that are envisioned as historical. So it is natural to think that an evolutionary account of morality might somehow throw light on the phenomenon itself.[3]

I am tempted by this possibility, but, just for that reason, I am dissatisfied with some recent biological accounts of the evolution of morality. In Section One, I will explain why I think there is a problem with these accounts. Basically, the problem is that it is unclear how they can explain the emergence of what I call "normative self-government": the capacity to be motivated to do something by the thought that you ought to do it. In Section Two, I will explore some solutions to that problem that have emerged from the sentimentalist tradition of moral philosophy, including Darwin's own solution, which drew on that tradition. And I

---

[2] *The Descent of Man*, Princeton edition.

[3] Philosophers at present do not go in much for origin stories. Analytic philosophy these days has become a crisp no-nonsense discipline, aligning itself with the sciences rather than with literature, and rejecting any modes of understanding whose methodological credentials are obscure. Since philosophy is a discipline of self-understanding, we are of course right to try to understand our own methods where we can. But crisp no-nonsense attitudes often express nothing more than a lack of imagination, and a desire to shut down perplexity as soon as possible. Philosophy should be wary of curbing its own resources.

will explain why I think those solutions don't work. My own account of morality is in a sense intended to address the problem, but in Section Three I will explain why it might seem to leave the difficulty in place. Finally, in Section Four, I will draw on an earlier tradition of theorizing about the evolution of morality, to suggest a possible origin story of my own.

1. Moral Content and Normative Self-Government

Many of the traditional candidates for the distinctively human attribute seem to have given way to recent discoveries or rediscoveries about the other animals. Animal scientists have established that many of the other animals acquire much of their know-how through learning rather than innate instinct, that some of the other animals use and manufacture tools, that some of them have local cultural traditions concerning what to eat, how to prepare it, and how to medicate themselves, and so on, and that a few can be taught some of the basic elements of language. So it is not surprising that scientists have also gone looking for the rudiments of morality in our non-human ancestors, and have claimed to find such rudiments in the evidence of tendencies to altruism, cooperation, empathy, or reconciliatory behavior that can be observed among some of the social animals.

The research supporting these kinds of claims has met with a degree of controversy that is a little puzzling. It is not surprising that those who reject the theory of evolution should dispute them; but it may seem surprising that scientists themselves, who presumably accept it, should still sometimes hotly contend for the uniquely human character of some of these attributes. Those who teach the other animals to communicate linguistically, for example, may be met with the claim that what the animal learns is not really language until the syntax reaches a certain level of complexity. By raising the standards for what counts as

having a certain attribute, we can perhaps preserve its distinctiveness, but what is the point of the exercise? It is not uncommon for those who wish defend our continuity with the other animals to speculate that there is some lingering piece of pride or ego at work in these controversies, something that makes human beings *want* to believe that we are unique among the animals.

I am sympathetic to the worry, and yet, I must also confess that I *am* inclined to believe that something I call "reason," one of whose manifestations is something I call "morality," is a distinctively human attribute, and one that might explain a lot of what seems to be so different about human beings.

But it is important to be clear about what I mean by "reason" here, and about its implications for the question of evolution. Frans De Waal, in *Primates and Philosophers*, distinguishes two schools of thought about morality. According to one of them, he tells us, morality is "a cultural innovation achieved by our species alone," where this is supposed to imply that "our ancestors became moral by choice." The other, his own theory, "views morality as a direct outgrowth of the social instincts we share with the other animals."[4] He associates the two views loosely with the rationalist and sentimentalist traditions in moral philosophy, and suggests that according to proponents of the rationalist view, morality is not something about which it is appropriate to tell an evolutionary story at all.

In fact I know of no philosophical view according to which human beings "became moral by choice," as De Waal puts it. But we might take De Waal's description of the rationalist position as a rough characterization of the sort of neo-Hobbesian or contractarian

---

[4] *Primates and Philosophers*, p. 6.

view according to which morality is founded on something like a social contract, entered into for reasons of self-interest. Such views take it for granted that "reason" is the standard of doing what is in your own best interests, and argue that morality is "rational" in the sense that it promotes those interests. When I talk about morality being a manifestation of reason, I am not talking about that sort of thing, but rather about views according to which moral laws are themselves principles of reason – such as rational intuitionist views, or Kant's view that the categorical imperative is a principle of reason. And I do think that "reason," in the sense that supports those theories, is something that must have evolved. So when I suggest that morality is a manifestation of reason, I do not mean to suggest that there is no evolutionary story to tell about its origins. But I do mean to register one source of my dissatisfaction with some of the current attempts to trace the evolution of morality, which is that I think that *what* they are trying to explain – which is characteristically altruism, cooperation, sharing, and so forth – is not quite the thing that needs to be explained.

Morality, as treated in these kinds of accounts, is defined by its characteristic content, which has something to do with, say, social relationships which take the interests of others into account. Marc Bekoff and Jessica Pierce, in their book *Wild Justice*, say, for example, "We define morality as a suite of other-regarding behaviors that cultivate and regulate complex interactions within social groups."[5] And De Waal, in *Primates and Philosophers*, claims that the essence of human morality is taking "the interests of the entire community into

---

[5] *Wild Justice*, p. 7.

account."[6] In the discussion following the lectures that make up that book, at which I was present, De Waal remarked that he regarded morality as "a system of conflict resolution."

But to someone working in the tradition of Plato, Aristotle, and Kant – or for that matter, as we will see, of Hume and Adam Smith – the characterization of morality as "a system of conflict resolution" or of a tendency to good social behavior is bound to sound a little thin. These philosophers, or so I want to protest, had something rather grander in mind.[7] They were talking about what they took to be our unique human capacity to take responsibility for ourselves, to give shape and form to our own identities or characters, and to make laws for our own conduct. They were talking not just about a relation in which we stand to others, but about a relation in which we stand to ourselves, which it does not seem very tempting to attribute to any of the other animals. Morality so regarded is one manifestation of the human capacity for what I am calling "normative self-government." That is, we have the capacity to assess the potential grounds of our beliefs and actions, to ask whether they constitute good reasons, and to regulate our beliefs and actions accordingly. In the theoretical realm, the capacity for normative self-government is expressed in the deliberate construction of systems of belief, employing consciously held standards of good evidence and valid argument. In the practical realm, it is expressed most obviously in the capacity to act from what we familiarly call "a sense of obligation," grounded in consciously

---

[6] *Primates and Philosophers*, p. 58.

[7] "Grander" may make it sound as if I am claiming that human beings are in some way superior to the other animals. But I'm not. When I say that human beings are the only moral animals, I mean that we are the only animals who are subject to moral standards – who can be either morally good or bad. I do not think that having that property is itself a virtue. I hope to explain more exactly why in "Kant's Formula of Humanity Revisited," indefinitely forthcoming, or anyway, somewhere.

held principles of good or right action. I think that *that*, the human capacity for normative self-government, and not just good social behavior, is the thing whose evolution needs to be explained.

Of course, everyone involved in these discussions grants that morality is not *merely* a tendency to good social behavior. If altruistic and cooperative behavior were the essence of morality, the ants and bees would be our moral heroes, and no one supposes that they are. And everyone also agrees that what these thinkers call "*human* morality" plainly involves something over and above altruistic or cooperative dispositions: some cognitive element such as the ability to follow explicit rules; or the self-conscious use of moral concepts; or the related capacity to make and be motivated by moral judgments. But explaining how that capacity arose is not usually part of the biologist's enterprise. In my commentary on De Waal in *Primates and Philosophers*, I claimed that the essence of morality rests in normative self-government rather than in altruism or cooperation. Bekoff and Pierce, commenting in part on those remarks, say that they regard such matters as being motivated by conscious moral judgments as "relatively late evolutionary additions to the suite of moral behaviors."[8] De Waal himself, in his response to the commentaries, suggested that the human capacity for "internal dialogue" "lifts moral behavior to a level of abstraction and self-reflection unheard of before our species entered the evolutionary scene."[9] I don't know exactly what these authors have in mind, but such remarks may suggest the idea that what is distinctive about "human morality" is the result of *adding* some kind of advanced intellectual faculties onto sociable instincts or desires. But exactly which advanced intellectual faculties are supposed to

---

[8] *Wild Justice*, pp. 139-40.

[9] *Primates and Philosophers*, p. 175.

be involved and how adding them to social instincts is supposed to produce a normatively self-governing animal is left rather vague. So something more needs to be said.

2. Darwin and the Sentimentalist Tradition

Unlike many of his more recent followers, Darwin did attempt to fill in this gap. Darwin took a keen interest in the sentimentalist tradition of moral philosophy that gave rise to the utilitarian theory that was dominant in his day. No doubt this was partly because of the time and place in which he lived, but I think it is also because philosophers in the sentimentalist tradition had tried to give an answer to the question how the sense of obligation might be something that human beings acquired. David Hume gives us one picture of how that might be. Leaving aside a complication about what Hume calls the "artificial" virtues, Hume thinks that moral standards are the result of our approving and disapproving of motives that we already, naturally, have.[10] Approval and disapproval are themselves sentiments, but they require advanced intellectual faculties for two reasons. First, as Hume himself emphasized, they arise only when we look at things from an impartial perspective that we must use reasoning to achieve.[11] Second, they require what contemporary ethologists call "theory of mind" – an awareness that people and animals have

---

[10] In "Natural Motives and the Motive of Duty: Hume and Kant on Our Duties to Others," I argue that Hume's account of the operation of the motive of duty in the case of the artificial virtues differs only slightly from Kant's account of its operation. It is little more than a matter of whether the moral sense operates through the mediation of self-disapproval, or directly as a kind of will. In a sense the argument of this paper is the same.

[11] See especially Hume, David, *Enquiry Concerning the Principles of Morals*, p. 173.

mental states, including motives, since those are the main objects of approval and disapproval.[12]

In Hume's account approval and disapproval are not in themselves motives – they are sentiments we feel about motives, our own and other people's. But Hume has a pretty good story about how it is possible for us to be motivated by the standards we form as a result of our approvals and disapprovals – how it is possible for us to be motivated by thoughts about what we ought to do.[13] Approval and disapproval are, according to Hume, forms of love and hate – a kind of disinterested love and hate that we feel when we view things from an impartial standpoint, not governed by our own self-interest.[14] These feelings of distinterested love and hate arise because we sympathize with the victims and beneficiaries of an agent's conduct, and love or hate that agent accordingly. So to know that you yourself are an object of the disapproval of others is see yourself as an object of their hatred. And since our natural sympathy with other people induces us to enter into what we suppose to be their feelings, it induces us to turn this hatred against ourselves. This motivates us to conform to moral standards: we wish to be lovable in the eyes of others,

---

[12] Hume affirms this at T.3.2.1,477, but in fact his practice does not conform to it; he also praises, e.g. qualities of character such as courage and industry which are not in themselves motives.

[13] Or rather, to put the point more strictly, he has a story to tell about how the standards we form as a result of our approval and disapproval *become* standards that tell us what we "ought" to do. I prefer to put the point that way, because I am an "internalist" about the moral "ought." "Ought" is a word used to express a *practical* judgment, so I do not think anything could count as a judgment about what you "ought to do" that is not capable of motivating you to do it. Practical normative force does not reduce to motivational force, but must always include it. Hume himself seems to accept something along these lines, for he famously criticized his rationalist opponents for being unable to explain how moral considerations, if they were grounded in reason – a faculty he regarded as inactive and inert - could possibly motivate us.

[14] Tref

because we wish to be lovable in our own.[15] Just to make sure I haven't confused you here, let me emphasize that sympathy plays a double role in Hume's theory: impartial sympathy with the victims and beneficiaries of action determines *what* we approve and disapprove of; sympathy with the approval and disapproval of imagined moral judges then motivates us act in ways we ourselves approve of, so that we can be lovable in our own eyes. Of course, one might complain that this theory does not really imply that, strictly speaking, we are motivated to do what we ought to do simply by the judgment, or by what goes into making the judgment, that we ought to do it. Rather, it implies that we are motivated to do what we ought to do because that is a way of avoiding self-hatred.[16] That there is a problem shows up in this fact: the same mechanism that motivates us to do what we ourselves approve of would motivate us to avoid the disapproval of others even if we thought that their disapproval was ill-founded. Sympathy, as Hume understands it, tends to make us hate ourselves if we think others either do or would hate us regardless of the causes of their hate.

Adam Smith modified this story in several ways, two of which are important for our purposes. Hume thought of approval and disapproval as forms of love and hate based on sympathy with the victims and beneficiaries of the conduct of the person who is morally judged. Smith, on the other hand, thought of approval itself as a form of sympathy with the

---

[15] Tref and see my "The General Point of View: Love and Moral Approval in Hume's Ethics"

[16] In fact Hume is explicit about this. "But may not the sense of morality or duty produce an action?...I answer, It may: …When any virtuous motive or principle is common in human nature, a person who feels his heart devoid of that principle, may hate himself upon that account, and may perform the action without the motive, from a certain sense of duty, in order to acquire by practice, that virtuous principle, or at least to disguise to h himself, as much as possible, his want of it." (T 3.1.1, 479) The role of sympathy with imagined moral judges in this is brought out more clearly at E2 276. See "Natural Motives and the Motive of Duty: Hume and Kant on Our Duties to Others"

person judged.[17] To disapprove of someone is to be out of sympathy with him. The other important modification is that Smith added a notion of what he called "propriety."[18] Hume thought that our approval and disapproval of motives is aroused by reflections on their utility and agreeableness. We approve of beneficence, say, because it is useful to those to whom the beneficent person offers assistance, and we sympathize with them. Smith argued that we also approve and disapprove of motives and the emotions on which they are based because of their suitability or proportionality to the objects that arouse them. We disapprove of the enraged person, say, because his anger seems out of proportion to the little annoyance that caused it, and this makes us unable to sympathize with him. Smith believed that strong emotional responses generally seem disproportionate to those who are not in the grip of them, and therefore that the tendency of our natural desire to be in sympathy with others is to moderate and control our violent responses. The person judged tones his responses down in order to win the sympathy of others; at the same time, the person making the judgment tries to imagine the situation more vividly in order to enter more fully into the feelings of the person who is judged. The eventual 'compromise' position reached – a level of response that puts the person judged and the person judging in sympathy with each other – is the "proper" response. These judgments of "propriety" give us the notion of a response being "worthy" of its object: we may say that the cause of someone's rage is "not worth" so strong a feeling.

Importantly, we can make judgments of "propriety" about the sentiments of approval and disapproval themselves. So when we do something wrong, we may judge that

---

[17] TMS ref

[18] TMS ref

it would be *proper* for others to disapprove or blame us if they knew. And when we judge that it would be proper for others to blame us, we are judging not merely that others would blame us if they knew, but that we are *blameworthy*. This appears to solve the problem in Hume's theory: we are only motivated to avoid conduct that we deem genuinely worthy of blame.[19] Smith thought of such judgments as being rendered by what he called "the man within," or the "impartial spectator," a kind of internalized representative of the other, but one whose view of our motives is unimpeded and therefore reliable.[20] When we are motivated to avoid the disapproval of the man within, it is as if we are in danger of falling out of sympathy with ourselves.

We know from his notebooks that Darwin studied this tradition of moral philosophy, and it seems clear that he was influenced by it when he came to produce his own account of the evolution of morality.[21] Darwin argued that the evolution of morality could be explained through the interaction of two powers, advanced intellectual faculties and social instincts. As he says:

> "Any animal whatever, endowed with well-marked social instincts, would inevitably acquire a moral sense or conscience as soon as its intellectual powers had become as well developed, or nearly as well-developed, as in man."[22]

The developed intellectual powers in question, as we will see, turn out to be memory and "theory of mind" - an awareness of our own motives.

---

[19] Worth noting here: Kant was a great admirer of Smith.

[20] TMS refs

[21] Notebooks ref

[22] *Descent of Man*, pp. 71-2.

Darwin's story turns on the difference between two kinds of instincts. There are social instincts, whose influence tends to be felt constantly by a social animal, and there are the instincts associated with the appetites, whose felt influence is only occurrent but, when it does occur, stronger than that of the social instincts. It is an important feature of the appetites and the instincts associated with them that, once they are satisfied, it is hard to recapture the sense of their force and urgency. So it is often the case that, once we have satisfied an appetite, what we have done seems to us not to have been worth it, especially if we have done it at the cost of satisfying some other desire or disobeying the call of some other instinct. Once our intellectual faculties have developed to the point where we can remember and reflect upon our motives and actions, this difference between the two kinds of instincts has an important effect. Darwin explains it this way:

> Thus, as man cannot prevent old impressions continually passing through his mind, he will be compelled to compare the weaker impressions of, for instance, past hunger, or of vengeance satisfied or danger avoided at the cost of other men, with the instinct of sympathy and good-will to his fellows, which is still present and ever in some degree active in his mind. He will then feel in his imagination that a stronger instinct has yielded to one which now seems comparatively weak; and then that sense of dissatisfaction will inevitably be felt with which man is endowed, like every other animal, in order that his instincts may be obeyed.[23]

---

[23] *Descent*, p. 90

According to Darwin this dissatisfaction is regret or remorse, and its painful character ultimately teaches us to control our appetites when they conflict with our social instincts. In addition, Darwin brings in, as it were direct from Hume, the consideration that even if a man does not regret his bad conduct for its own sake, "he will be conscious that if his conduct were known to his fellows, it would meet with their disapprobation, and few are so destitute of sympathy as not to feel discomfort when this is realized."[24]

Of course one might be inclined to protest – as I did against Hume – that this is not really doing what you ought to do because you ought to do it. We learn to conform to moral principles in order to avoid the uncomfortable feeling of "regret" or "remorse." Smith, as we saw, tried to remedy the problem by adding a normative element to the negative emotion, the self-disapproval, itself: it is not the sense that we *will* be blamed or that we *would* be blamed if others knew of our bad conduct, but the sense that our conduct would be *worthy* of blame, that motivates us to avoid it. Darwin, I believe, is trying to capture this feature of Smith's theory in his own account by emphasizing the difference in the ways in which the two kinds of instincts affect us: like Smith, he thinks that when we are not immediately in the grip of an appetite, it is hard to recapture the sense of urgency we have when we *are* in its grip. So when we think about it later, it seems to us as if it is not *worth* satisfying our appetites at the cost of the interests of others, and that looks like a normative thought.

Nevertheless, Darwin's account does give rise to a problem similar to the one I noticed in Hume's. In Hume's theory, the problem is that the disapproval of others would motivate us even if it were not properly grounded in standards of right and wrong. In

---

[24] *Descent*, p. 92

Darwin's theory, the parallel problem is that the difference between constantly and occurently felt instincts would eventually teach us to conform to the constantly felt ones even if those were not the social instincts.[25] Why, in Darwin's theory, are constantly felt instincts the right ones to act on? *Any* instincts that were constant and steady in their influence would become authoritative over *any* instincts whose influence was occurrent, regardless of the content of those instincts. Darwin is unable to appropriate Smith's idea successfully, because of a problem in Smith's theory itself: Smith never really tells us why exactly the motives and responses with which others can sympathize are supposed to be the right ones to act on, or even why we should tend to think that they are. In the same way, Darwin has no story about why constantly felt instincts should be the right ones to act on.

Of course, Darwin, unlike Smith, was not trying to produce a general normative theory. In fact he was assuming a vaguely utilitarian framework, although he suggests it is not the greatest happiness of the community, but rather something he calls the greatest "good or welfare" of the community, at which moral conduct aims.[26] His account of this "good or welfare" has a distinctly biological ring. He says: "The term, general good, may be defined as the means by which the greatest possible number of individuals can be reared in full vigor and health, with all their faculties perfect, under the conditions to which they are exposed."[27] It might seem easy enough to marry such an account of morality to an account of its evolution, but even if we accepted the moral view in question, the problem would still exist. It is not *because* the social instincts are constant and steady in their influence that it is

---

[25] Darwin actually says it: "The imperious word *ought* seems merely to imply the consciousness of the existence of a persistent instinct…" *Descent*, p. 92.

[26] *Descent*, pp. 97-98.

[27] *Descent*, p. 98

wrong to ignore the interests of others in pursuit of the satisfaction of your own appetites. Of course, if you think that *all* that morality *is* is the way in which the social instincts express themselves in intellectually advanced animals, this point may elude you. But if you think there is more to the idea that an action is wrong than that it is unsociable, then the relation between Darwin's motivational story and the normative one is, after all, too accidental: our capacity for moral motivation is a mechanism that just happens to favor the kind of conduct that Darwin considers moral.

These theories, born of the empiricist tradition of associationist psychology, try to explain the origin of normative self-government by showing how some sort of pain gets attached to conduct independently identified as wrongful. One might complain that this doesn't give us a creature who is normatively self-governed; this still only gives us a creature who is governed by the desire to avoid pain. But it would be uncharitable to take them to be suggesting that the creature's *goal* is simply avoid pain, for that is not the only role that pain can play in the explanation of action. We should take them to be explaining, in associationistic fashion, how avoiding wrongdoing itself becomes a goal. So instead I will put my point this way. I think that these theories come very close to explaining moral motivation in the right way. If they were true, they would succeed in explaining the existence of creatures who inevitably find wrongdoing painful. And although just now I said that the conduct is "independently" identified as wrongful, I did not mean that the causes of the conduct's painfulness and the reasons for its wrongness are totally unrelated. In Hume's view, the fact that you disapprove of an action is what makes it wrong, and it is also what makes it painful for you to do it. Nevertheless, its being wrong is not what makes it painful

for you to do it – your desire to be lovable is what does that.[28] And although Darwin doesn't tell us exactly how he arrives at his normative account of the good, I think we may say, in a similar way, that in Darwin's theory, the fact that conduct is against our social instincts is both what makes it wrong and what makes it painful. Nevertheless, it is not painful because it is wrong, but because of the way the social instincts express themselves, constantly rather than occurrently. But a normatively self-governed being is one who is motivated to avoid wrongful conduct because it is wrong; the motivation must be produced by the wrongness itself, not merely attached to it, even if it is non-accidentally attached to it. The reasons why actions are right or wrong must be the reasons why we do or avoid them. So it looks as if nothing short of what Kant called "pure practical reason" can possibly do the job.

Actually, I don't really mean to make such a strong claim, anyway on this occasion. My point is rather that whatever it is that makes some actions required and some wrong must also be the *source* of our motivation for doing and avoiding them accordingly. And what makes some actions required or wrong is not merely their content: that they are

---

[28] And what makes you unlovable is not the wrongness of your conduct, but its content: that it is disagreeable or disadvantageous. To this extent Hume's theory shares a problem with the brand of naturalistic moral realism that claims that we know moral properties exist because they do play a role in explanation: say, the laborers revolted because they were treated unjustly. No: the laborers revolted because they didn't have enough to live on. Their not having enough to live on was unjust, and was why they revolted, but they didn't revolt because it was unjust; they revolted because their families were hungry. That would have caused them to revolt even if it were not unjust. (I owe the point to Chris Furlong.) The parallel point about Darwin is a little hard to formulate, but it goes like this: even if it were essential to the nature of social instincts that they be expressed constantly rather than occurrently – even if a constant expression and social content had to go together – it would be the case that it was the constant-expressedness of the social instincts, rather than the wrongness of violating them, that motivated us not to violate them.

altruistic, or cooperative, or sociable, or whatever, but rather whatever it is that confers normativity on that content, whatever it is that *makes* it right to act cooperatively or altruistically or whatever. Kant does give one answer to that question – what makes an action right or wrong is determined by whether its maxim has the form of a law, and he claims that the moral motive – respect for law itself – is directly responsive to that consideration. But the more general point is that whatever confers a normative status on our actions – whatever makes them right or wrong – must also be what motivates us to do or avoid them accordingly, without any intervening mechanism.

This may seem to imply that we cannot explain the evolution of morality until we have the correct moral theory – until we know what it is that actually makes our actions right or wrong. Among other things, that would mean giving up any hope that thinking about the evolution of morality could throw any light on morality itself. But I do not take the implication of what I have just said to be quite that strong. Rather, I take the implication to be that no account of the evolution of morality can be complete unless it includes an account of why we assign normative properties – rightness or wrongness – to our actions in the first place: that is, to say, of why we think of our actions as the sort of thing that must be morally or rationally justified. And for this we need to know what the problem is to which justification, or the assignment of a normative status, is a response. For an animal who is motivated to do or avoid certain actions depending on whether or not they can be morally justified must see himself as faced with the problem of justifying his actions in the first place, and must be motivated to do what he judges to be right by the fact that it solves that problem. And most of the evolutionary theories on the table these days tell us little or

nothing about what that problem is or why it arose. The other animals do not need to justify their actions. Why do we?

3. Self-Consciousness and the Problem of Justification

My own views are in part an attempt to address the question I just raised. In this section, however, I will explain why it might seem as if they leave the situation pretty much in the same place as the sentimentalist views do. I will start by being a little more specific about what I think "reason" is. A non-human animal, I believe, is guided through her environment by means of a representation of that environment that incorporates both perceptual information and appropriate desiderative or aversive responses. What I mean is that, for the other animals, perceptual representation and desire and aversion are not strictly separate. The animal finds herself in a world that consists of things that are directly perceived *as* food or prey, *as* danger or predator, *as* potential mate, *as* child: that is to say, as things that are to-be-eaten, to-be-avoided, to-be-mated-with, to-be-cared-for, and so on. In this sense, we might say that an animal's perception has *teleological content*: the objects she perceives are marked out as being "for" certain things or as calling for certain responses. I believe this because I think it is hard to see how perception could have been of any use to the relatively unintelligent animals in which it first evolved if something like this were not the case. Perception could not merely provide a simple animal with theoretical information on the basis of which the animal had to figure out what to do, so it must be that it tells the animal what to do. If you feel tempted to say that it is "instinct" that tells the animal what to do, I will reply that I am imagining that this is form that instinct takes. But then it is important to add that the contrast that I want here is not between "instinctive" and

"learned." An animal might learn from experience that certain things are to-be-avoided, but if the form that the learning takes is that she now simply sees them that way, as to-be-avoided, her actions are still "instinctive" in the sense I have in mind.

Rational actions, as opposed to ones that are instinctive in this sense, involve a certain form of self-consciousness: namely, consciousness of yourself as a subject – the subject of certain thoughts, desires, experiences and so forth. I will explain why in a moment. But first let us ask: are human beings the only animals that are self-conscious in this sense? Some scientists believe that this form of self-consciousness is revealed by the ethologist's mirror test. In the mirror test, a scientist paints, say, a red spot on an animal's body and then puts her in front of a mirror. Given certain experimental controls, if the animal eventually reaches for the spot and tries to rub it off, or looks away from the mirror towards that location on her body, we can take that as evidence that the animal recognizes herself in the mirror, and is curious about what has happened to her. Apes, dolphins, and elephants have passed the mirror test, in some cases moving on to use the mirror to examine parts of their bodies that they can't normally see – apparently with great interest. Other animals never recognize themselves, and instead keep offering to fight with the image in the mirror, or to engage in some other form of social behavior with it.

It is a little difficult to articulate exactly why the mirror test is supposed to reveal an awareness of oneself as a subject. The animal grasps the relation between the image in the mirror and her own body. In so doing, she seems to show that she grasps the relationship between *herself* and her own body. For she grasps the relationship between two things, a certain physical body and – well, what? – we can say "and herself" – but what exactly is the "herself" that she identifies with that body? Perhaps the idea is that what she identifies as

herself is the self that is the subject of her own experiences, for instance the one who sees the spot in the mirror, of whose existence she must then have some awareness.

Interestingly, however, even if this is right, and shows that the animal knows herself as the subject of her *experiences*, it does not yet show that the animal must be aware of herself as the subject of her *attitudes* – that is, of her beliefs, emotions, and desires. And this suggests a possible division within this form of self-consciousness. An animal might be aware of her experiences and of herself as the subject of those experiences, and yet her attitudes might still be invisible to her, because they are a lens *through* which she sees the world, rather than being parts of the world that she sees.[29] In that case, she would still function in the way I have called "instinctive." The experiences that she was aware of having would still be experiences of things as "to-be-eaten" "to-be-fled" "to-be-cared-for" and so on; and her responses to those things would still be governed by the teleological content of her experiences.

But as rational beings we are aware of our attitudes. We know of ourselves that we want certain things, fear certain things, love certain things, believe certain things, and so on. And we are also aware of the potential influence of our attitudes on what we will decide to do. We are aware of the *potential grounds* of our actions – of the ways in which our attitudes incline us to respond. And once you are aware of the influence of a potential ground of action, you are in a position to decide whether to allow yourself to be influenced in that way

---

[29] It's easier to understand what I mean here when you are thinking about practical, evaluative attitudes. It sounds odd to think of beliefs as a lens through which we see the world. But they are, in the sense that an animal could be moved by one belief to take up another without having any awareness of making an inference. Unlike a person, a non-human animal can think "X" without commitment to "I believe X" or "X is true," because he (probably) has no commitments of that sort.

or not. As I have put it elsewhere, you now have a certain reflective distance from the impulse that is influencing you, and you are in a position to ask yourself, "but *should* I be influenced in that way?" You are now in a position to raise a *normative* question, a question about whether the action you find yourself inclined to perform is *justified*.[30]

Or so I have said in the past. And we might at first suppose that if something along these lines is right, it is easy to explain the evolution of the point of view from which normative problems arise. It is just a matter of a gradual increase in the scope of "theory of mind" – our grasp of our inner world expanding from knowledge of ourselves as the subject of experiences to knowledge of ourselves as the subject of certain attitudes towards those experiences. But there are several problems with leaving it at that. The first problem is that even if self-consciousness about the grounds of our beliefs and actions makes it *possible* to raise normative questions, in the sense that it makes room for them, that fact by itself does not explain exactly why these questions arise for us or what kind of questions they are. The second problem, which many of my own readers have pointed out to me in the past, is that it is not perfectly clear why just being conscious of the grounds of your beliefs and actions should be sufficient to put you, as it were, in normative control. It seems perfectly possible that we could be aware of a force operating on us mentally, but still be helpless in the face of it.

So at this point it might look as if my own account needs as a supplement just the sort of theory that the sentimentalists offered – we are aware of what goes on in our own minds, and in particular of the motivational forces at work upon us, but now something

---

[30] Korsgaard, *The Sources of Normativity* (Cambridge: Cambridge University Press, 1996), §§3.2.1-3.2.3, pp. 92-8.

must motivate us to take control of those forces and redirect them in accordance with normative standards. But I have already argued that such an account cannot work. The trouble with this picture, I now believe, may be that it gets things the wrong way around. We are not able to take control of our mental attitudes because we are aware of them. Rather, I will suggest, our possession of self-conscious mental attitudes is a *product* of our efforts to take control over what goes on in our minds.[31]

4. The Origins of Rationality

Before I explain what I have in mind, I want to remind you of an older line of thought about the evolution of morality, proposed in slightly different ways by Nietzsche, in the *Genealogy of Morals*, and Freud, in works like *Civilization and its Discontents* and *Totem and Taboo*. Both were concerned in particular about the origin of guilt, and both suggested that guilt originated when an animal who was not allowed to give expression to his aggressive instincts turned those aggressive instincts against himself. Suffering from guilt is a way of hurting yourself, done for the sheer satisfaction of hurting *someone* when you need to hurt someone and are not allowed to do it. In Nietzsche's theory guilt is continuous with the self-mutiliating behavior often observed in animals kept in cages, and for that matter in unhappy human children. The originator of what Nietzsche called "the bad conscience" was "the man who, from lack of external enemies and resistances … impatiently lacerated, persecuted, gnawed at, assaulted, and maltreated himself; this animal that rubbed itself raw

[31] This formulation isn't quite right but will be refined below.

on the bars of its cage as one tried to tame it…."[32] The details of why we had to turn our aggressive instincts inward are deliberately vague, and don't much matter. In Nietzsche's story, stronger people, blond beasts from the north, impose social forms on weaker people, for purposes of their own; and it is these social forms that inhibit the expression of the aggressive instincts; in Freud's it is of course the omnipotent father who inhibits the expression of aggression in his child.

Freud and Nietzsche wrote of turning aggression against your own instincts, and punishing yourself for having them, but it seems to me that there is another possibility here, closely related to that but not quite the same. Nowadays, scientists believe that versions of the dominance hierarchy are pervasive among social animals. When one animal dominates another, the subordinate animal gives way to the dominant one in competitive situations, as when they both want access to a certain bit of food or a mate. Dominance is sometimes established by aggression, and sometimes maintained that way, but not always: in some animals dominance hierarchies can be inherited and apparently go unchallenged for longish stretches of time. It appears that its evolutionary function of dominance may be to *reduce* the frequency of aggressive encounters in animal life. I think that that dominance is interesting in this context, because dominance looks a lot like something that we think of as essentially normative: it looks like *authority*. A dominated animal does something that he does not want to do, or foregoes something that he would like to have, because he acknowledges something like the *standing* of another animal. It is not mere fear of the consequences – if

---

[32] *Genealogy* p. 85

you successfully dominate your dog, for example, he isn't afraid of you.[33] He just recognizes that you are in charge, and he is supposed to do what you tell him to.

I'm not interested in defending the details of these theories.[34] What I want from them is the suggestion that the origin of morality might rest in the internalization of mechanisms of dominance and social control: that is, the suggestion that we began to become rational animals when we began, as individuals, to exert a kind of dominance over ourselves – to inhibit our own instinctive responses. I'm not going to speculate about how exactly it happened, or why. Nietzsche and Freud make their stories sound like cataclysmic events in the lives of individuals; somehow that has to be translated into evolutionary terms.[35] The important point for me now is that in Nietzsche's story, the internalization of the aggressive instincts is explicitly linked with a kind of deepening of consciousness itself. He writes:

> All instincts that do not discharge themselves outwardly turn inward – this is what I call the *internalization* of man: thus it was that man first developed what was later called his "soul." The entire inner world, originally as thin as if it were stretched between two membranes, expanded and extended itself,

---

[33] Grandin: it's not usually dominance with dogs, its parenthood. But dominance does happen.

[34] One thing I find attractive in these theories is that they lack the "happy talk" character of some of the biological theories, in which morality is all about being nice, sociable, sharing… there is a dark side to the life lived in judgment on the self, and these more psychological theories aim to capture that.

[35] It is tempting to speculate that the evolution of individuals capable of a distinctive form of self-control is a route to making complex forms of social life possible that is in a sense opposite to the one taken by the social insects. Instead of getting rid of all our anti-social impulses and becoming mere cogs in a larger social machine, we learn to control them.

> acquiring depth, breadth, and height, in the same measure as outward discharge was inhibited.[36]

What I want to suggest, following Nietzsche's lead, is that the consequence of this internalization was a new form of self-consciousness, which set us altogether new kinds of problems of its own: normative problems.

So let me rephrase the suggestion with which I started. I suggested that normative self-government is not the result of our awareness of our own mental attitudes; rather our awareness of our own mental attitudes is the result of the control we began to assume over ourselves and our own responses. That way of putting it is right in a way, but it doesn't quite capture what I take to be the radical nature of Nietzsche's suggestion. It makes it sound as if our minds are stocked with a full panoply of mental attitudes, and what internalization does is turn on the lights so we can see them, and that is not what he says: what he says is that our minds acquired depth, breadth, and so on – a dimension they lacked before, not one they had in the dark. So I take the suggestion to be that at least some of our mental attitudes are the *products* of the internalization: that our beliefs, desires, emotions, and so on, are the result of the new form of consciousness that emerged.[37]

I know that what I am saying sounds mysterious – how could a form of consciousness produce its own objects in that way? Although, for that matter, of course there is a way in which forms of consciousness do produce their own objects – just think of sensory qualities. And the way I am describing it also may make it sound as if I think animal

---

[36] GM 84-85

[37] Similar views in *Self-Constitution*: we become agents by taking control over our own movements; also, mental states as produced by mental activity, by a dividing off of previously unified elements in consciousness.

minds are empty of mental attitudes, that they must lack mental states. And I don't think that. What I have in mind is rather that things we identify as our own attitudes – our "beliefs" "desires" and to some extent our "emotions" are the products of the breakdown of the teleological consciousness that I have claimed must characterize the nonhuman mind. They are the result of our beginning to factor out and identify the ways in which we ourselves contribute to, and so are responsible for, the way the world is for us.

Adam Smith can help us out here. He suggested that we would never think of our own minds if we were never exposed to other people. Contrary to what the privileged access view of the mind might lead you to suppose, we first spot mental attitudes in other people. From my own, untutored, point of view, *I* am not angry: I am simply the victim of an outrage, and that's a plain fact about the world. That is the teleological view of the world at work in me: the situation confronting me is one I perceive as to-be-defeated, or something like that. But when I see you in that situation, when I'm not in it myself, I see that *you* are getting angry. There is a distancing use of mental attitude language: was he in danger? well, he *believed* that he was; well, he was certainly *frightened.* A gap between the way world seems to me and the way it seems to you appears to me at first as a *distortion* in the way it seems to you; so I conclude that something about *you* must be distorting the way it seems to you. If I am a dominant animal, perhaps I see this as an occasion to inhibit your response.

But when I begin to see occasion to inhibit my own responses, then I also begin to regard myself in the way that in Smith's story, I was regarding you. The identification of something as an attitude at work in me is a recognition that I am, or something about me is, making some sort of contribution to the way the world is for me. If being aware of a mental

attitude, or more properly of the workings of your own mind, is essentially being aware of your own contribution to the way the world is for you, then as Kant said our mental attitudes are always accompanied by an "I." I think, I want, I intend. And from this recognition that our own mental activity is implicated in the way the world is for us arises a new relation in which we stand to the world. When we begin to recognize the ways that conceptualizing, evaluating, and responding to the world are things that our minds *do* – that is, things that we do – then we begin to do them in a whole new way, namely self-consciously. And then we are confronted with a new problem and a whole new set of questions, questions about what (if anything) counts as doing these things *correctly*. Is this a good ground for belief? Is this a good reason to act? Those are the questions of justification, questions that, so far as we can tell, only human beings ask. And when we begin to find answers to those questions, then the use of mental attitude language about ourselves no longer carries the implication of *distortion*: instead it carries the implication of *normative commitment*: "yes, this is what I believe" "yes, this is the right thing to do." To believe and act on the basis of such thoughts is to be a normatively self-governed animal.

Conclusion

I have suggested that the internalization of mechanisms of dominance and social control – the attempt to inhibit our own instinctive responses – was the first step in a process that led to a kind of general takeover, or attempted takeover, of our own mental lives. Mental states with an essentially normative dimension are the product of this takeover, factored out from the teleological consciousness when we identify our own contribution to the way the world is for us. The recognition that our own mental activity contributes to the

way the world is for us leads us to attempt to regulate that contribution, to get it right, and that leads to the formation of consciously held standards for constructing our own conception of the world and consciously held standards for determining our own actions. Those are the standards of reason, which we then take to govern these activities. That is how we become normatively self-governing animals.

But now I must conclude by bringing this all back home to morality. For perhaps you may feel that I have only reversed the problem I started out from: I've got normative self-government on the table, but lost characteristic moral content. After all why, according to this theory, should the kinds of conduct we ordinarily call "moral" represent the correct solution to the problem of justifying our actions? In particular, why should altruism, cooperation, and fairness, be part of that solution? In the absence of a particular theory of justification, which obviously I can't give here, it is difficult to be specific, but let me end by making a couple of suggestions about how we might get what we ordinarily think of as moral content back on the table. Both suggestions turn on this fact: that the problem of justification arises for an animal for whom the teleological view of the world has broken down.

The first point is this. Once we have reflective distance from our grounds of our attitudes, and can ask whether we should act on them or not, we need a way of answering that question. To ask whether you should indeed flee from something you perceive as to-be-fled, for instance, is, *in the first instance*, to ask whether it is really a threat, whether it can really harm you. I say "in the first instance" because at this stage we have not yet arrived at the fully practical question. At this stage the practical question is still mainly instrumental, taking it for granted that, say, objects that really can do us harm are to-be-avoided, and only asking

which objects those are. When we only think or reason instrumentally, we are still seeing the world through the lens of our own desires and interests, and to that extent we are still seeing the world teleologically. There is a further question to be asked about when danger is *worth* facing or harm *worth* incurring and when it is not, not just instrumentally, but for its own sake. That is not just a question about how best to satisfy our interests, but a question about what our interests ought to be – in fact, it is essentially the very question whose answer Smith and Darwin tried to build into their theories. So the breakdown of the teleological worldview of the non-rational animal means that we can longer take it for granted that we should measure the world by our own interests, but instead must form an independent standard of what is worth doing for the sake of what.

The second point is this: an essential part of overcoming the teleological worldview of the animal is recognizing that things don't exist in relation to me. The world does not does not after all consist of my predators, my prey, my offspring, but of rather of beings with an independent existence of their own, who happen to stand in those relationships to me. Getting that fact firmly into view is essential to achieving a rational *theoretical* conception of the world, a conception of a world that exists independently of me and my practical interests. But it is also – intuitively speaking – essential to achieving the conception of the world that we nowadays recognize as practically rational –that is to say, as moral. That women do not exist to bear men's children and keep their houses, that strong young men are not fodder for older people's cannons, that people of color were not born to work in white people's fields, and the poor and ignorant do not exist that the rich may have servants, and that other animals are not there for human beings to eat and work for us and submit to our experiments – that all of these beings do not exist *for us*, and with reference simply to our

interests, but have an independent existence and interests of their own – grasping these facts is essential to forming a theoretical conception of a world that exists independently of us as well as a practical conception of the world we must relate to.

So justification is not merely about how we can best satisfy our own interests, but is about what is worth doing for its own sake. And it must be responsive to the fact that there are many other beings, who do not exist just for us, or in relation to us, but independently of us, with interests of their own. If someday we can put those two thoughts together in just the right way, perhaps one day we ourselves *will* become animals in whom morality has finally evolved.