Forthcoming in *Philosophical Studies* (January 2006).

SHARON STREET

A DARWINIAN DILEMMA FOR REALIST THEORIES OF VALUE[*]

1. INTRODUCTION

Contemporary realist theories of value claim to be compatible with natural science. In this paper, I call this claim into question by arguing that Darwinian considerations pose a dilemma for these theories. The main thrust of my argument is this. Evolutionary forces have played a tremendous role in shaping the content of human evaluative attitudes. The challenge for realist theories of value is to explain the relation between these evolutionary influences on our evaluative attitudes, on the one hand, and the independent evaluative truths that realism posits, on the other. Realism, I argue, can give no satisfactory account of this relation. On the one hand, the realist may claim that there is *no* relation between evolutionary influences on our evaluative attitudes and independent evaluative truths. But this claim leads to the implausible skeptical result that most of our evaluative judgments are off track due to the distorting pressure of Darwinian forces. The realist's other option is to claim that there *is* a relation between evolutionary influences and independent evaluative truths, namely that natural selection favored ancestors who were able to grasp those truths. But this account, I argue, is unacceptable on scientific grounds. Either way, then, realist theories of value prove unable to accommodate the fact that Darwinian forces have deeply influenced the content of human values. After responding to three objections, the third of which leads me to argue against a realist understanding of the disvalue of pain, I conclude by sketching how antirealism is able to sidestep the dilemma I have presented. Antirealist theories of value are able to offer an alternative account of the relation between evolutionary forces and evaluative facts—an account

[*] For their comments on earlier versions of this paper, I am indebted to Melissa Barry, Paul Boghossian, Hartry Field, Patricia Kitcher, Philip Kitcher, Christine M. Korsgaard, Thomas Nagel, Derek Parfit, T. M. Scanlon, Nishi Shah, Michael Strevens, and audiences at Amherst College, Columbia University, and Duke University. I am also indebted to Paul Bloomfield and Earl Conee for their comments on this paper at the 2005 Pacific Division APA.

that allows us to reconcile our understanding of evaluative truth with our understanding of the many non-rational causes that have played a role in shaping our evaluative judgments.

2. THE TARGET OF THE ARGUMENT: REALIST THEORIES OF VALUE

The defining claim of realism about value, as I will be understanding it, is that there are at least some evaluative facts or truths that hold independently of all our evaluative attitudes.[1] *Evaluative facts or truths* I understand as facts or truths of the form that *X* is a normative reason to *Y*, that one should or ought to *X*, that *X* is good, valuable, or worthwhile, that *X* is morally right or wrong, and so on.[2] *Evaluative attitudes* I understand to include states such as desires, attitudes of approval and disapproval, unreflective evaluative tendencies such as the tendency to experience *X* as counting in favor of or demanding *Y*, and consciously or unconsciously held evaluative judgments, such as judgments about what is a reason for what, about what one should or ought to do, about what is good, valuable, or worthwhile, about what is morally right or wrong, and so on.

It is important to note that it is not enough to be a realist to claim that the truth of an evaluative judgment holds independently of one's making *that particular* evaluative judgment. Antirealists can agree with that much. Consider, for example, a constructivist view according to which the truth of '*X* is a reason for agent *A* to *Y*' is a function of whether that judgment would be among *A*'s evaluative judgments in reflective equilibrium. This view is antirealist because it understands truths about what reasons a person has as depending on her evaluative attitudes (in particular, on what those attitudes would be in reflective equilibrium). Yet on this view, it is quite possible for someone to have a reason independently of whether she thinks she does, for whether she has a reason is not a function of whether she (presently)

[1] More broadly, realism about value may be understood as the view that there are *mind-independent* evaluative facts or truths. I focus on independence from our *evaluative attitudes* because it is independence from this type of mental state that is the main point of contention between realists and antirealists about value.

[2] My target in this paper is realism about *practical reasons*, or reasons for action, as opposed to epistemic reasons, or reasons for belief. While I actually think the Darwinian Dilemma can be extended to apply against realism about epistemic reasons, that topic is more than I'll be able to pursue here. Throughout the paper, I use the word 'reason' in the sense of a normative reason—in other words, in the sense of a consideration that *counts in favor of*, or *justifies*, some action.

judges she has it, but rather a function of whether that judgment would be among her evaluative judgments in reflective equilibrium. Antirealists can therefore agree with realists that the truth of a given evaluative judgment holds independently of whether one makes that particular judgment. Where antirealists part ways with realists is in denying that there are evaluative truths which hold independently of *the whole set* of evaluative judgments we make or might make upon reflection, or independently of *the whole set* of other evaluative attitudes we hold or might hold upon reflection.

The kind of independence from our evaluative attitudes that realists endorse is what Russ Shafer-Landau has called *stance-independence*.[3] To illustrate: realists of course agree that the evaluative truth that 'Hitler was morally depraved' depends in part on *Hitler's* evaluative attitudes in the sense that if Hitler had valued peace and universal human rights instead of dictatorial power and genocide, then it would have been false instead of true that he was morally depraved. But given that Hitler *did* value dictatorial power and genocide, value realists think that it is true, independent of all of our (and any of Hitler's other) evaluative attitudes, that Hitler was morally depraved. According to realists, the truth that Hitler was morally depraved holds independently of any stance that we (or Hitler) might take toward that truth, whether now or upon reflection.

There are different brands of realism about value. What unites them is the view that there are evaluative facts or truths that hold independently of all our evaluative attitudes (now keeping in mind the qualification about stance-independence). What separates different kinds of realists from one another is how they construe the nature of these facts or truths. According to what I will call *non-naturalist* versions of value realism, evaluative facts or truths are not reducible to any kind of natural fact, and are not the kinds of things that play a role in causal explanations; instead, they are irreducibly normative facts or truths.[4] This brand of realism has been gaining increasing numbers of adherents in recent years, and it lies squarely within the target of the Darwinian Dilemma.

---

[3] See Shafer-Landau (2003), 15. Shafer-Landau borrows the term from Ronald Milo.

[4] Important statements of this view include Nagel (1986), especially chapter 8; Dworkin (1996); and Shafer-Landau (2003). T. M. Scanlon's view on the nature of reasons in chapter 1 of Scanlon (1998) is

In contrast to non-naturalist versions of value realism, the position I will call *value naturalism* holds that evaluative facts are identical with or constituted by (certain) natural facts, and that evaluative facts *are* the kinds of things that play a role in causal explanations.[5] According to such views, much as water is identical with H_2O, so evaluative properties are identical with certain natural properties, though we may or may not ever be able to provide a reduction telling exactly *which* natural properties evaluative properties are identical with (different naturalists taking different views on the possibility of such a reduction[6]). Whereas non-naturalist versions of value realism lie straightforwardly within my target in this paper, it is a more complicated matter whether versions of value naturalism lie within my target. Answering this question requires making a distinction (in section 7) between versions of value naturalism which count as genuinely realist on my understanding and versions which don't; my argument will be that the former, but not the latter, are vulnerable to the Darwinian Dilemma. Before introducing these complexities, however, it is important to get the fundamental dilemma for realism on the table.[7]

---

also plausibly read along these lines. Many of these authors (though not Shafer-Landau) might resist the label 'non-naturalist,' due to its potential connotations of mysterious "extra" properties in the world, but so long as we keep in mind these authors' insistence that their view involves positing no such properties, the label is useful enough to be adopted.

[5] Key statements of this view include Sturgeon (1985); Railton (1986); Boyd (1988); and Brink (1989). As Brink notes in Brink (1989), value naturalism can be construed as claiming either that evaluative facts are *identical* with natural facts or that evaluative facts are *constituted* by natural facts; Brink argues that value naturalism should be construed as making the constitutive claim (see section 6.5 and 176-177). For brevity's sake, I gloss over this distinction in what follows and talk simply in terms of identity, not constitution.

[6] Railton, for example, thinks that such reductions will be forthcoming, and sketches what they might look like in Railton (1986), whereas non-reductionist naturalists such as Sturgeon, Boyd, and Brink think that such reductions may not be, and need not be, forthcoming.

[7] In addition to naturalist and non-naturalist versions of realism, there is one other very different brand of realism that should be mentioned, namely the quasi-realism of Simon Blackburn and Allan Gibbard. (See Blackburn (1984), (1993), and (1998); and Gibbard (1990) and (2003).) Their views occupy an uneasy position with regard to the realism/antirealism debate as I am understanding it. There is not space to address their positions here, so for the purposes of this paper I set quasi-realism entirely to one side, and focus exclusively on "non-quasi" brands of realism.

3. A CAVEAT

In his 1990 book *Wise Choices, Apt Feelings*, Allan Gibbard notes that his arguments "should be read as having a conditional form: If the psychological facts are roughly as I speculate, here is what might be said philosophically."[8] I attach a similar caveat to my argument in this paper: If the evolutionary facts are roughly as I speculate, here is what might be said philosophically. I try to rest my arguments on the least controversial, most well-founded evolutionary speculations possible. But they are speculations nonetheless, and they, like some of Gibbard's theorizing in *Wise Choices, Apt Feelings*, fall within a difficult and relatively new subfield of evolutionary biology known as evolutionary psychology.[9] According to this subfield, human cognitive traits are (in some cases) just as susceptible to Darwinian explanation as human physical traits are (in some cases). For example, a cognitive trait such as the widespread human tendency to value the survival of one's offspring may, according to evolutionary psychology, be just as susceptible to evolutionary explanation as physical traits such as our bipedalism or our having opposable thumbs. There are many pitfalls that such evolutionary theorizing must avoid, the most important of which is the mistake of assuming that every observable trait (whether cognitive or physical) is an adaptation resulting from natural selection, as opposed to the result of any number of other complex (non-selective or only partially selective) processes that could have produced it.[10] It is more than I can do here to describe such pitfalls in depth or to defend at length the evolutionary claims that my argument will be based on. Instead, it must suffice to emphasize the hypothetical nature of my arguments, and to say that while I am skeptical of the *details* of the evolutionary picture I offer, I think its *outlines* are certain enough to make it well worth exploring the philosophical implications.[11]

---

[8] Gibbard (1990), 30.

[9] For introductions to the field of evolutionary psychology, see Barkow et al. (1992) and Buss (1999).

[10] See Gould and Lewontin (1979). For a more recent overview, see Pigliucci and Kaplan (2000).

[11] Cf. Gibbard (1990), 30.

4. THE FIRST PREMISE: THE INFLUENCE OF EVOLUTIONARY FORCES ON THE CONTENT OF OUR EVALUATIVE JUDGMENTS

In its first approximation, the opening premise of the Darwinian Dilemma argument is this: The forces of natural selection have had a tremendous influence on the content of human evaluative judgments. This is by no means to deny that all kinds of other forces have also shaped the content of our evaluative judgments. No doubt there have been numerous other influences: some of them were perhaps evolutionary factors *other* than natural selection—for example, genetic drift;[12] and many other forces were not evolutionary at all, but rather social, cultural, historical, or of some other kind. And then there is the crucial and *sui generis* influence of rational reflection that must also be taken into account—a point I return to in the next section. I am discounting none of these other influences. My claim is simply that one enormous factor in shaping the content of human values has been the forces of natural selection, such that our system of evaluative judgments is thoroughly saturated with evolutionary influence. In this section, I make a brief case in support of this view, starting with a highly simplified and idealized evolutionary picture, then discussing two important complications, and ending with a more refined statement of the first premise.

To begin, note the potentially phenomenal costs and benefits, as measured in the Darwinian currency of reproductive success, of accepting some evaluative judgments rather than others. It is clear, for instance, how fatal to reproductive success it would be to judge that the fact that something would endanger one's survival is a reason to do it, or that the fact that someone is kin is a reason to harm that individual. A creature who accepted such evaluative judgments would run itself off of cliffs, seek out its predators, and assail its offspring, resulting in the speedy elimination of it and its evaluative tendencies from the world.[13] In contrast, it is clear how beneficial (in terms of reproductive success) it would be to

[12] Genetic drift is the random fluctuations of gene frequencies within a population (see Avers (1989)). Later in this section, I argue that natural selection's influence on our evaluative judgments is best understood as having been *indirect*. A similar point would apply to the influence of other evolutionary forces such as genetic drift.

[13] This assumes that other things are equal—for example, that the effects of these evaluative judgments on the creature's behavior are not cancelled out by other evaluative judgments that the creature makes. The

judge that the fact that something would promote one's survival is a reason in favor of it, or that the fact that something would assist one's offspring is a reason to do it. Different evaluative tendencies, then, can have extremely different effects on a creature's chances of survival and reproduction. In light of this, it is only reasonable to expect there to have been, over the course of our evolutionary history, relentless selective pressure on the content of our evaluative judgments, or rather (as I discuss below) "proto" versions thereof. In particular, we can expect there to have been overwhelming pressure in the direction of making those evaluative judgments which tended to promote reproductive success (such as the judgment that one's life is valuable), and against making those evaluative judgments which tended to decrease reproductive success (such as the judgment that one should attack one's offspring).

The hypothesis that this is indeed very roughly what happened is borne out by the patterns of evaluative judgment that we observe in human beings today. There is, of course, a seemingly unlimited diversity to the evaluative judgments that human beings affirm. Yet even as we note this diversity, we also see deep and striking patterns, across both time and cultures, in many of the most basic evaluative judgments that human beings tend to make. Consider, as a brief sampling, the following judgments about reasons:

(1) The fact that something would promote one's survival is a reason in favor of it.
(2) The fact that something would promote the interests of a family member is a reason to do it.
(3) We have greater obligations to help our own children than we do to help complete strangers.
(4) The fact that someone has treated one well is a reason to treat that person well in return.
(5) The fact that someone is altruistic is a reason to admire, praise, and reward him or her.
(6) The fact that someone has done one deliberate harm is a reason to shun that person or seek his or her punishment.

What explains the widespread human acceptance of such judgments? There are so many other possible judgments about reasons we could make—so why these? Why, for instance, do we view the death of our offspring as a horror, rather than as something to be sought after? Why do we think that altruism with no

---

statement in the text also assumes that a creature is motivated to act in accordance with its evaluative judgments, other things being equal. I do not offer an explicit defense of this internalist assumption in this paper, but I take it to be supported by the plausibility of the overall picture that emerges, and by the hypothesis, argued for in section 6, that the function of evaluative judgments from an evolutionary point of view is not to "track" independent evaluative truths, but rather to *get us to respond* to our circumstances in ways that are adaptive.

hope of personal reward is the highest form of virtue, rather than something to be loathed and eliminated? Evolutionary biology offers powerful answers to these questions, very roughly of the form that *these* sorts of judgments about reasons tended to promote survival and reproduction much more effectively than the alternative judgments. The details of how survival and reproduction were promoted will vary depending on the evaluative tendency in question. In the case of judgment (1), for instance, the rough explanation is obvious: creatures who possessed this general evaluative tendency tended to do more to promote their survival than those who, say, had a tendency to view the fact that something would promote their survival as counting *against* it, and so the former tended to survive and reproduce in greater numbers. The explanation of evaluative tendencies in the direction of judgments such as (2) and (3) will be somewhat more complicated, drawing on the evolutionary theory of kin selection.[14] The explanation in the case of evaluative tendencies in the direction of judgments (4), (5), and (6), meanwhile, will appeal to the biological theory of reciprocal altruism.[15]

For the sake of contrast, consider the following possible evaluative judgments:

(1') The fact that something would promote one's survival is a reason against it.
(2') The fact that something would promote the interests of a family member is a reason not to do it.
(3') We have greater obligations to help complete strangers than we do to help our own children.
(4') The fact that someone has treated one well is a reason to do that individual harm in return.
(5') The fact that someone is altruistic is a reason to dislike, condemn, and punish him or her.
(6') The fact that someone has done one deliberate harm is a reason to seek out that person's company and reward him or her.

If judgments like these—ones that would, other things being equal, so clearly decrease rather than increase the reproductive success of those who made them—predominated among our most deeply and

[14] On this theory, see Hamilton (1963) and (1964); chapter 2 of Sober and Wilson (1998); and chapters 7 and 8 of Buss (1999).

[15] On the theory of reciprocal altruism, see Trivers (1971); Axelrod (1984); chapter 2 of Sober and Wilson (1998); and chapter 9 of Buss (1999). It is important to note what the explanandum is in these sorts of evolutionary explanations. The explanandum is *not* particular attitudes held by particular individuals—for example, your or my or George W. Bush's judgment that the fact that something would help a family member is a reason to do it. Such individual-level facts are not appropriate objects of evolutionary explanation. What *are* appropriate objects of evolutionary explanation are population-level facts about patterns of variation in a given trait across a population—and the widespread presence of certain basic evaluative tendencies in the human population are such objects. For further discussion of how population-level and not individual-level facts are appropriate objects of explanations in terms of natural selection, see chapter 5 of Sober (1984).

widely held evaluative judgments across both time and cultures, then this would constitute powerful evidence that the content of our evaluative judgments had *not* been greatly influenced by Darwinian selective pressures. But these are not the evaluative judgments we tend to see; instead, among our most deeply and widely held judgments, we observe many like those on the first list—many with exactly the sort of content one would expect if the content of our evaluative judgments had been heavily influenced by selective pressures. In this way, the observed patterns in the actual content of human evaluative judgments provide evidence in favor of the view that natural selection has had a tremendous influence on that content.

A further piece of evidence in favor of this view is the striking continuity that we observe between many of our own widely held evaluative judgments and the more basic evaluative tendencies of other animals, especially those most closely related to us. It does not seem much of a stretch, for example, to say that chimpanzees, in some primitive, non-linguistic sort of fashion, experience certain things in the world as *calling for* or *counting in favor of* certain reactions on their part. Moreover, the *content* of these evaluative experiences seems to overlap significantly with the content of many of our own evaluative tendencies. Like us, individual chimpanzees seem to experience—at some basic motivational level—actions that would promote their survival or help their offspring as in some way "called for." More strikingly, and again at some basic motivational level, chimpanzees seem to experience the fact that another chimpanzee has helped them, whether by sharing food, grooming them, or supporting their position within the group hierarchy, as "counting in favor of" assisting that other individual in similar ways.[16] While more work is needed to make such claims precise and subject them to thorough scientific testing, they have a strong basic plausibility, such that the conspicuous continuities between the basic evaluative tendencies of our close animal relatives and our own evaluative judgments lend further support to the view that evolutionary forces have played a large role in shaping the content of

[16] See, for instance, de Waal (1996).

our evaluative judgments. We may view many of our evaluative judgments as conscious, reflective endorsements of more basic evaluative tendencies that we share with other animals.

Now note two important complications to this rough evolutionary sketch. First of all, the discussion so far might have suggested that things happened this way: *first* our ancestors began making evaluative judgments, and *then* tendencies to make some of these evaluative judgments rather than others were selected for. In other words, first came the capacity to make evaluative judgments, and then followed the selection of their content. But the actual course of evolution certainly did not take place in these two stages, much less in that order. Consider again the list of widely held evaluative judgments I mentioned earlier. Behavioral and motivational tendencies in the direction of at least some of the pairings of circumstance and response on this list presumably arose and became entrenched in our ancestors long before the rise of any capacity for full-fledged evaluative judgment—where I am understanding the capacity for *full-fledged evaluative judgment* to involve not only an unreflective capacity to experience one thing as "demanding" or "counting in favor of" another (a more primitive capacity that other animals such as chimpanzees might share with us), but also a reflective, linguistically-infused capacity to judge that one thing counts in favor of another, and to step back from such judgments and call them into question. Behavioral and motivational tendencies to do what would help one's offspring, for example—behavioral and motivational tendencies of gradually increasing degrees of consciousness and complexity—presumably vastly predated the sophisticated, linguistically-infused capacity to make the reflective judgment that 'The fact that something would help one's offspring is a reason to do it.' Thus, the capacity for full-fledged evaluative judgment was a relatively late evolutionary add-on, superimposed on top of much more basic behavioral and motivational tendencies.[17]

A second complication is this. In order for evolution by natural selection to take place with respect to a given trait, the trait in question must be genetically heritable. Yet it is implausible to think that the acceptance of a *full-fledged evaluative judgment* with a given content—for example, the

[17] I am indebted to Peter Godfrey-Smith for helpful comments regarding these points.

acceptance of the judgment that 'One ought to help those who help you'—is a genetically heritable trait. That is to say: when individuals in a given population vary with respect to whether or not they make this evaluative judgment (or any other), most if not all of that variation is likely *not* due to genetic differences, but other factors (such as culture or upbringing).[18] In contrast, however, it is plausible to suppose that over the course of much of our evolutionary history, what I have been calling "more basic evaluative tendencies" *were* genetically heritable traits, where a *basic evaluative tendency* may be understood very roughly as an unreflective, non-linguistic, motivational tendency to experience something as "called for" or "demanded" in itself, or to experience one thing as "calling for" or "counting in favor of" something else. We may think of these as "proto" forms of evaluative judgment. A relatively primitive version of such a tendency might be possessed by a bird who experiences some kind of motivational "pull" in the direction of feeding its offspring. A more sophisticated version might be possessed by a chimpanzee who has a motivational and perhaps emotional or proto-emotional experience of certain behaviors as "called for" by certain circumstances (for example, the experience of a threat to its offspring as "demanding" a protective response). It seems plausible to hypothesize that over the course of much of our evolutionary history—perhaps up until relatively recently[19]—when individuals in a given population varied with respect to whether they possessed a given basic evaluative tendency, a significant portion of that variation *was* due to genetic differences. So, for example, when individuals varied with respect to the presence or absence of an unreflective tendency to experience the fact that someone helped them as "counting in favor of" helping the other in return, a significant portion of that variation was attributable to genetic differences.

The upshot of these complications is this. The influence of Darwinian selective pressures on the content of human evaluative judgments is best understood as *indirect*. The most plausible picture is that natural selection has had a tremendous *direct* influence on what I have called our "more basic evaluative

[18] For discussion of the concept of genetic heritability, see Block and Dworkin (1974).

[19] For discussion of how the genetic heritability of a trait can vary over time, see Block and Dworkin (1974), 41.

tendencies," and that these basic evaluative tendencies, in their turn, have had a major influence on the evaluative judgments we affirm. By this latter claim I do not mean that we automatically or inevitably accept the full-fledged evaluative judgments that line up in content with our basic evaluative tendencies. Certainly not: For one thing, other causal influences can shape our evaluative judgments in ways that make them stray, perhaps quite far, from alignment with our more basic evaluative tendencies.[20] For another thing, we are reflective creatures, and as such are capable of noticing any given evaluative tendency in ourselves, stepping back from it, and deciding on reflection to disavow it and fight against it rather than to endorse the content suggested by it. My point here is instead the simple and plausible one that had the general content of our basic evaluative tendencies been very different, then the general content of our full-fledged evaluative judgments would also have been very different, and in loosely corresponding ways.[21] Imagine, for instance, that we had evolved more along the lines of lions, so that males in relatively frequent circumstances had a strong unreflective evaluative tendency to experience the killing of offspring that were not his own as "demanded" by the circumstances, and so that females, in turn, experienced no strong unreflective tendency to "hold it against" a male when he killed her offspring in such circumstances, on the contrary becoming receptive to his advances soon afterwards. Or imagine that we had evolved more along the lines of our close primate relatives the bonobos, so that we experienced sexual relations with all kinds of different partners as "called for" in all kinds of different circumstances. Finally, imagine that we had evolved more on the model of the social insects, perhaps possessing overwhelmingly strong unreflective evaluative tendencies in the direction of devoting

[20] Indeed, it is likely that we were selected above all else to be extremely flexible when it comes to our evaluative judgments—not locked into any particular set of them but rather able to acquire and adjust them in response to the conditions in which we find ourselves. In suggesting that we possess basic evaluative tendencies, then, I am simply suggesting that when it comes to certain core issues such as our individual survival, the treatment of our offspring, and reciprocal relations with others, there are likely to be strong predispositions in the direction of making some evaluative judgments rather than others, for instance (referring back to my earlier lists) judgments (1) through (6) as opposed to judgments (1') through (6').

[21] This counterfactual claim is all I need for the purposes of my argument. While one might inquire into the exact causal process by which basic evaluative tendencies have influenced the content of human evaluative judgments, it is not necessary for me to enter into such questions here.

ourselves to the welfare of the entire community, and only the weakest tendency to look out for our own individual survival, being unreflectively inclined to view that survival as "good" only insofar as it was of some use to the larger community. Presumably in these and other such cases our system of full-fledged, reflective evaluative judgments would have looked very different as well, and in ways that loosely reflected the basic evaluative tendencies in question. My conclusion: The content of human evaluative judgments has been tremendously influenced—*indirectly* influenced, in the way I have indicated, but nevertheless tremendously influenced—by the forces of natural selection, such that our system of evaluative judgments is saturated with evolutionary influence. The truth of some account very roughly along these lines is all that is required for the Darwinian Dilemma to get off the ground.[22]

5. THE FIRST HORN OF THE DILEMMA: DENYING A RELATION

The basic problem for realism is that it needs to take a position on what relation there is, if any, between the selective forces that have influenced the content of our evaluative judgments, on the one hand, and the independent evaluative truths that realism posits, on the other. Realists have two options: they may either assert or deny a relation.

Let us begin with the realist's option of claiming that there is *no* relation. The key point to see about this option is that if one takes it, then the forces of natural selection must be viewed as a purely distorting influence on our evaluative judgments, having pushed us in evaluative directions that have nothing whatsoever to do with the evaluative truth. On this view, allowing our evaluative judgments to be shaped by evolutionary influences is analogous to setting out for Bermuda and letting the course of your boat be determined by the wind and tides: just as the push of the wind and tides on your boat has nothing to do with where you want to go, so the historical push of natural selection on the content of our evaluative judgments has nothing to do with evaluative truth. Of course every now and then, the wind and tides might happen to deposit someone's boat on the shores of Bermuda. Similarly, every now and

[22] In the remainder of the paper, I will often speak loosely about the influence of natural selection on our evaluative judgments, without reiterating the complications I have discussed in this section. These complications should nevertheless be kept in mind throughout.

then, Darwinian pressures might have happened to push us toward accepting an evaluative judgment that accords with one of the realist's independent evaluative truths. But this would be purely a matter of chance, since by hypothesis there is no relation between the forces at work and the "destination" in question, namely evaluative truth.

If we take this point and combine it with the first premise that our evaluative judgments have been tremendously shaped by Darwinian influence, then we are left with the implausible skeptical conclusion that our evaluative judgments are in all likelihood mostly off track, for our system of evaluative judgments is revealed to be utterly saturated and contaminated with illegitimate influence. We should have been evolving towards affirming the independent evaluative truths posited by the realist, but instead it turns out that we have been evolving towards affirming whatever evaluative content tends to promote reproductive success. We have thus been guided by the wrong sort of influence from the very outset of our evaluative history, and so, more likely than not, most of our evaluative judgments have nothing to do with the truth. Of course it's *possible* that as a matter of sheer chance, some large portion of our evaluative judgments ended up true, due to a happy coincidence between the realist's independent evaluative truths and the evaluative directions in which natural selection tended to push us, but this would require a fluke of luck that's not only extremely unlikely, in view of the huge universe of logically possible evaluative judgments and truths, but also astoundingly convenient to the realist. Barring such a coincidence, the only conclusion remaining is that many or most of our evaluative judgments are off track. This is the far-fetched skeptical result that awaits any realist who takes the route of claiming that there is no relation between evolutionary influences on our evaluative judgments and independent evaluative truths.

But the realist may not be ready to abandon this route just yet. Let us grant (sticking with this horn of the dilemma) that the distorting influence of natural selection on the content of our evaluative judgments has been tremendous. One might nevertheless object that to draw a skeptical conclusion from this is unwarranted. For the argument so far ignores the power of a very different kind of influence on our system of evaluative judgments—a kind of influence that one might claim *is* related to the truth and

that has also been tremendous—namely, the influence of rational reflection. After all, we are not unthinking beings who simply endorse whatever evaluative tendencies were implanted in us by evolutionary forces. Over the course of human history, endless amounts of reflection have gone on and greatly altered the shape of our evaluative judgments. According to the objection at hand, just as a compass and a little steering can correct for the influence of the wind and tides on the course of one's boat, so rational reflection can correct for the influence of selective pressures on our values.[23]

I accept one important point that this objection makes. Any full explanation of why human beings accept the evaluative judgments we do would need to make reference to the large influence of rational reflection. The view I am suggesting by no means involves thinking of us as automatons who simply endorse whatever evaluative tendencies are implanted in us by evolutionary and other forces. On the contrary, the view I am suggesting acknowledges the point that we are self-conscious and reflective creatures, and in a sense seeks to honor that point about us *better* than alternative views, by asking what reflective creatures like ourselves should conclude when we become conscious of what Kant would call this "bidding from the outside" affecting our judgments. (Here I have in mind Kant's statement in the third section of the *Foundations of the Metaphysics of Morals* that "we cannot conceive of a reason which consciously responds to a bidding from the outside with respect to its judgments."[24]) The very fact of our reflectiveness implies that something must happen—that something must change—when we become conscious of any foreign influence (such as these Darwinian forces) on our evaluative judgments. What that change should be is exactly what I am exploring in this paper.

Where I think the objection goes wrong, then, is as follows. The objection gains its plausibility by suggesting that rational reflection provides some means of standing apart from our evaluative judgments, sorting through them, and gradually separating out the true ones from the false—as if with the aid of some uncontaminated tool. But this picture cannot be right. For what rational reflection about evaluative matters involves, inescapably, is assessing some evaluative judgments in terms of others.

[23] I owe this last way of putting the point to Paul Boghossian.

[24] Kant (1785 [1959]), 448.

Rational reflection must always proceed from some evaluative standpoint; it must work from some evaluative premises; it must treat some evaluative judgments as fixed, if only for the time being, as the assessment of other evaluative judgments is undertaken. In rational reflection, one does not stand completely apart from one's starting fund of evaluative judgments: rather, one *uses* them, reasons in terms of them, holds some of them up for examination in light of others. The widespread consensus that the method of reflective equilibrium, broadly understood, is our sole means of proceeding in ethics is an acknowledgment of this fact: ultimately, we can test our evaluative judgments only by testing their consistency with our other evaluative judgments, combined of course with judgments about the (non-evaluative) facts. Thus, if the fund of evaluative judgments with which human reflection began was thoroughly contaminated with illegitimate influence—and the objector has offered no reason to doubt *this* part of the argument—then the tools of rational reflection were equally contaminated, for the latter are always just a subset of the former. It follows that all our reflection over the ages has really just been a process of assessing evaluative judgments that are mostly off the mark in terms of others that are mostly off the mark. And reflection of *this* kind isn't going to get one any closer to evaluative truth, any more than sorting through contaminated materials with contaminated tools is going to get one closer to purity. So long as we assume that there is no relation between evolutionary influences and evaluative truth, the appeal to rational reflection offers no escape from the conclusion that, in the absence of an incredible coincidence, most of our evaluative judgments are likely to be false.[25]

6. THE SECOND HORN OF THE DILEMMA: ASSERTING A RELATION

So let us now turn to the realist's other option, which is to claim that there *is* indeed some relation between the workings of natural selection and the independent evaluative truths that he or she posits. I think this is the more plausible route for the realist to take. After all, we think that a lot of our evaluative

[25] If one holds that the assessment of (non-evaluative) factual judgments also proceeds via reflective equilibrium, one might wonder why the points in this paragraph don't apply equally well to rational reflection about scientific matters (for example). The key difference is that in the scientific case, our "starting fund" of (non-evaluative) factual judgments need not be viewed as mostly "off track." For further discussion, see note 35 below.

judgments are true. We also think that the content of many of these same evaluative judgments has been influenced by natural selection. This degree of overlap between the content of evaluative truth and the content of the judgments that natural selection pushed us in the direction of making begs for an explanation. Since it is implausible to think that this overlap is a matter of sheer chance—in other words, that natural selection just happened to push us toward true evaluative judgments rather than false ones—the only conclusion left is that there is indeed some relation between evaluative truths and selective pressures. The critical question is what *kind* of relation. Different metaethical views will give different answers, and we may judge them according to those answers.

The realist has a possible account of the relation that might seem attractive on its face. It is actually quite clear, the realist might say, how we should understand the relation between selective pressures and independent evaluative truths. The answer is this: we may understand these evolutionary causes as having *tracked* the truth; we may understand the relation in question to be a *tracking* relation.[26] The realist might elaborate on this as follows. Surely, he or she might say, it is advantageous to recognize evaluative truths; surely it promotes one's survival (and that of one's offspring) to be able to grasp what one has reason to do, believe, and feel. As Derek Parfit has put the point:[27] it is possible that "just as cheetahs were selected for their speed, and giraffes for their long necks, the particular feature for which we were selected was our ability to respond to reasons and to rational requirements."[28] According to this hypothesis, our ability to recognize evaluative truths, like the cheetah's speed and the giraffe's long neck, conferred upon us certain advantages that helped us to flourish and reproduce. Thus, the forces of natural selection that influenced the shape of so many of our evaluative judgments need not and should not be

[26] I borrow the term 'tracking' from Robert Nozick, who uses it in similar contexts in Nozick (1981).

[27] In correspondence.

[28] Nozick suggests something very similar in Nozick (1981) when he writes that "It seems reasonable to assume there has been some evolutionary advantage in acting for (rational) reasons. The capacity to do so, once it appeared, would have been selected for. Organisms able and prone to act for (rational) reasons gained some increased efficiency in leaving great-grand progeny" (337).

viewed as distorting or illegitimate at all. For the evaluative judgments that it proved most selectively advantageous to make are, in general, precisely those evaluative judgments which are true.

Call this proposal by the realist the *tracking account*. The first thing to notice about this account is that it puts itself forward as a scientific explanation.[29] It offers a specific hypothesis as to how the course of natural selection proceeded and what explains the widespread presence of some evaluative judgments rather than others in the human population. In particular, it says that the presence of these judgments is explained by the fact that these judgments are true, and that the capacity to discern such truths proved advantageous for the purposes of survival and reproduction. So, for instance, if it is asked why we observe widespread tendencies to take our own survival and that of our offspring to be valuable, or why we tend to judge that we have special obligations to our children, the tracking account answers that these judgments are true, and that it promoted reproductive success to be able to grasp such truths.

In putting itself forward as a scientific explanation, the tracking account renders itself subject to all the usual standards of scientific evaluation, putting itself in direct competition with all other scientific hypotheses as to why human beings tend to make some evaluative judgments rather than others. The problem for realism is that the tracking account fares quite poorly in this competition. Even fairly brief consideration suggests that another evolutionary explanation of why we tend to make some evaluative judgments rather than others is available, and that this alternative explanation, or something roughly like it, is distinctly superior to the tracking account.

According to what I will call the *adaptive link account*, tendencies to make certain kinds of evaluative judgments rather than others contributed to our ancestors' reproductive success not because they constituted perceptions of independent evaluative truths, but rather because they forged adaptive

[29] This brings out the interesting way in which non-naturalist versions of value realism, in spite of their insistence that values are *not* the kinds of things that play a role in causal explanations, are ultimately forced (unless they opt for the first horn of the dilemma) to take a stand on certain matters of scientific explanation—in particular, on questions about why human beings tend to make some evaluative judgments rather than others, and on the origins of our capacity to grasp independent evaluative truths. Indeed, as I'll try to show, these realists are forced (again, unless they opt for the first horn) to posit a causal role for evaluative truths in the course of our species' evolution.

links between our ancestors' circumstances and their responses to those circumstances, getting them to act, feel, and believe in ways that turned out to be reproductively advantageous.[30] To elaborate: As a result of natural selection, there are in living organisms all kinds of mechanisms that serve to link an organism's circumstances with its responses in ways that tend to promote survival and reproduction. A straightforward example of such a mechanism is the automatic reflex response that causes one's hand to withdraw from a hot surface, or the mechanism that causes a Venus's-flytrap to snap shut on an insect. Such mechanisms serve to link certain kinds of circumstances—the presence of a hot surface or the visit of an insect—with adaptive responses—the immediate withdrawal of one's hand or the closing of the flytrap. Judgments about reasons—and the more primitive, "proto" forms of valuing that we observe in many other animals—may be viewed, from the external standpoint of evolutionary biology, as another such mechanism. They are analogous to the reflex mechanism or the flytrap's apparatus in the sense that they also serve to link a given circumstance with a given response in a way that may tend to promote survival and reproduction. Consider, for example, the evaluative judgment that the fact that someone has helped one is a reason to help that individual in return. Just as we may see a reflex mechanism as effecting a pairing between the circumstance of a hot surface and the response of withdrawing one's hand, so we may view this evaluative judgment as effecting a pairing between the circumstance of one's being helped and the response of helping in return. Both of these pairings of circumstance and response, at least if the evolutionary theory of reciprocal altruism is correct about the latter case, are ones that tended to promote the reproductive success of ancestors who possessed them.[31]

---

[30] For closely related points, see Blackburn, who writes that an evaluative attitude's "function is to mediate the move from features of a situation to a reaction" (1993), 168; and Gibbard, who writes that the "biological function [of normative judgments] is to govern our actions, beliefs, and emotions" (1990), 110.

[31] In order for a mechanism which *effects a pairing* between the circumstance of a hot surface and the response of withdrawing one's hand to be adaptive, there must of course be a means of *detecting* the presence of a hot surface. Similarly, in order for a "mechanism" which *effects a pairing* between the circumstance of one's being helped and the response of helping in return to be adaptive, there must be a means of *detecting* or *tracking* circumstances in which one is helped. In proposing the adaptive link account, what I mean to be focusing in on are the mechanisms which *effect the pairing* between (perceived) circumstance and response, and *not* the mechanisms which do the (separate) job of *tracking*

Now of course there are radical differences between the mechanism of a reflex response and the "mechanism" of an evaluative judgment. The former is a brute, hard-wired physical mechanism, while the latter is a conscious mental state, subject to reflection and possible revision in light of that reflection. But this does not change the fact that there is a deep analogy between their functional roles. From an evolutionary point of view, each may be seen as having the same practical point: *to get the organism to respond* to its circumstances in a way that is adaptive.[32] Something like a reflex mechanism does this through a particular hard-wiring of the nervous system, while an evaluative judgment—or a more primitive evaluative experience such as some other animals are likely to have—does this by having the organism experience a particular response as *called for*, or as *demanded by*, the circumstance in question. In the latter case, the link between circumstance and response is forged by our taking of the one thing to be a *reason* counting in favor of the other—that is, by the experience of normativity or value.[33]

---

circumstances. While in the case of an automatic reflex mechanism, it may be hard to pull these mechanisms apart, the two jobs are nevertheless theoretically distinct, and the "mechanisms" clearly do come apart in the case of (non-evaluative) factual judgment versus evaluative judgment. Our capacity for (non-evaluative) factual judgment does the job of *tracking circumstances* (tracking, among innumerable other things, which individuals have helped us), whereas our capacity for evaluative judgment does the job of *effecting pairings of (perceived) circumstance and response* (getting us, among many other things, to respond to those who have helped us with help in return).

[32] I do not mean to be offering a full explanation of why we have a capacity to make evaluative judgments. Among other things, I say nothing to address the question: why did we evolve this "normative capacity" as a means of forging links between circumstance and response instead of, for instance, having such links forged solely by brute reflex mechanisms? The answer presumably has to do with the incredible flexibility and plasticity of the former capacity as opposed to reflex mechanisms, but this is not a question that I need to enter into for the purposes of my argument.

[33] Here I have suggested that it's a certain kind of *conscious experience*—for example, the conscious experience of the fact that someone has helped you as "counting in favor of" helping in return—that does the work of forging adaptive links between circumstance and response, and which was selected for. But a qualification is needed here, for it may be that this *conscious experience* was not itself directly selected for, but is rather an incidental byproduct of underlying information-processing and behavior-control systems which *were* selected for. If this is so, it does not pose any problem for my argument, since the only point I need for my argument is that the content of our evaluative judgments has been greatly affected by the influence of natural selection. This point still holds even if what was selected for are certain information-processing and behavior-control systems, which in turn give rise, as an incidental byproduct, to conscious experiences—here, in particular, of some things as "counting in favor of" other things.

For illustration of the differences between the adaptive link account and the tracking account, consider a few examples. Consider, for instance, the judgment that the fact that something would promote one's survival is a reason to do it, the judgment that the fact that someone is kin is a reason to accord him or her special treatment, and the judgment that the fact that someone has harmed one is a reason to shun that person or retaliate. Both the adaptive link account and the tracking account explain the widespread human tendencies to make such judgments by saying that making them somehow contributed to reproductive success in the environment of our ancestors. According to the tracking account, however, making such evaluative judgments contributed to reproductive success because they are *true*, and it proved advantageous to grasp evaluative truths. According to the adaptive link account, on the other hand, making such judgments contributed to reproductive success not because they were true or false, but rather because they got our ancestors to respond to their circumstances with behavior that itself promoted reproductive success in fairly obvious ways: as a general matter, it clearly tends to promote reproductive success to do what would promote one's survival, or to accord one's kin special treatment, or to shun those who would harm one.

We now have rough sketches of two competing evolutionary accounts of why we tend to make some evaluative judgments rather than others. For reasons that may already have begun to suggest themselves, I believe that the adaptive link account wins this competition hands down, as judged by all the usual criteria of scientific adequacy. In particular, there are at least three respects in which the adaptive link account is superior to the tracking account: it is more parsimonious; it is much clearer; and it sheds much more light on the explanandum in question, namely why human beings tend to make some evaluative judgments rather than others.

Let me start with the parsimony point. The tracking account obviously posits something extra that the adaptive link account does not, namely independent evaluative truths (since it is precisely these truths that the tracking account invokes to explain why making certain evaluative judgments rather than others conferred advantages in the struggle to survive and reproduce). The adaptive link account, in contrast, makes no reference whatsoever to evaluative truth; rather, it explains the advantage of making

certain evaluative judgments directly, by pointing out how they got creatures who made them to act in ways that tended to promote reproductive success. Thus, the adaptive link account explains the widespread presence of certain values in the human population more parsimoniously, without any need to posit a role for evaluative truth.[34]

Second, the adaptive link account is much clearer than the tracking account, which turns out to be rather obscure upon closer examination. As we have seen, according to the tracking account, making certain evaluative judgments rather than others promoted reproductive success *because these judgments were true*. But let's now look at this. How exactly is this supposed to work? Exactly why would it promote an organism's reproductive success to grasp the independent evaluative truths posited by the realist? The realist owes us an answer here. It is not enough to say: "Because they are true." We need to know more about *why* it is advantageous to apprehend such truths before we have been given an adequate explanation.

What makes this point somewhat tricky is that on the face of it, it might seem that *of course* it promotes reproductive success to grasp any kind of truth over any kind of falsehood. Surely, one might think, an organism who is aware of the truth in a given area, whether evaluative or otherwise, will do better than one who isn't. But this line of thought falls apart upon closer examination. First consider truths about a creature's manifest surroundings—for example, that there is a fire raging in front of it, or a predator rushing toward it. It is perfectly clear why it tends to promote reproductive success for a creature to grasp such truths: the fire might burn it to a crisp; the predator might eat it up.[35] But there are many other kinds of truths such that it will confer either no advantage or even a disadvantage for a given

[34] For related discussion, see Blackburn (1993), 169, and Gibbard (1990), 107-108.

[35] It is points like this which explain why the Darwinian Dilemma doesn't go through against realism about non-evaluative facts such as facts about fires, predators, cliffs, and so on. In short, the difference is that in the case of such non-evaluative facts, unlike in the case of evaluative facts, the tracking account prevails as the best explanation of our capacity to make the relevant sort of judgment. In order to explain why it proved advantageous to form judgments about the presence of fires, predators, and cliffs, one will need to posit in one's best explanation that there *were indeed* fires, predators, and cliffs, which it proved quite useful to be aware of, given that one could be burned by them, eaten by them, or could plummet over them. For related discussion, see Gibbard (1990), chapter 6, and Gibbard (2003), 253-258.

kind of creature to be able to grasp them. Take, for instance, truths about the presence or absence of electromagnetic wavelengths of the lowest frequencies. For most organisms, such truths are irrelevant to the undertakings of survival and reproduction; hence having an ability to grasp them would confer no benefit. And then one must also take into account the significant costs associated with developing and maintaining such a sophisticated ability. Since for most organisms, this would be energy and resources spent for no gain in terms of reproductive success, the possession of such an ability would actually be positively *disadvantageous*.

With this in mind, let us look again at the evaluative truths posited by realists. Take first the irreducibly normative truths posited by non-naturalist realists such as Nagel, Dworkin, Scanlon, or Shafer-Landau. A creature obviously can't run into such truths or fall over them or be eaten by them. In what way then would it have promoted the reproductive success of our ancestors to grasp them? The realist owes us an answer here, otherwise his or her alleged explanation of why it promotes reproductive success to make certain judgments in terms of the *truth* of those judgments is no explanation at all. To say that these truths could kill you or maim you, like a predator or fire, would be one kind of answer, since it makes it clear how recognizing them could be advantageous. But such an answer is clearly not available in the case of the independent irreducibly normative truths posited by the non-naturalist realists. In the absence of further clarification, then, the non-naturalist's version of the tracking account is not only less parsimonious but also quite obscure.

Value naturalists would appear to have better prospects on this point than non-naturalist realists. Since value naturalists construe evaluative facts as natural facts with causal powers, it is much more comprehensible how grasping such facts could have had an impact on reproductive success. I return to this issue in the following section. For the time being, note the following. The naturalist's proposed version of the tracking account, so far, is this: Making some evaluative judgments rather than others tended to promote reproductive success because those judgments constituted perceptions of evaluative facts, which just are a certain kind of natural fact. At least so far, this isn't much of an explanation either. What kinds of natural facts are we talking about, and exactly why did it promote reproductive success to

grasp them? The naturalist can certainly try to develop answers to these questions, but at least on the face of things, the prospects appear dim. Take the widespread judgment that one should care for one's offspring, for example. Exactly what natural fact or facts does the evaluative fact that one should care for one's offspring reduce to, or irreducibly supervene upon, and why would perceiving the natural fact or facts in question have promoted our ancestors' reproductive success? It seems unattractive to get into such complexities when one can just say, as the adaptive link account does, that ancestors who judged that they should care for their offspring met with greater reproductive success simply because *they tended to care for their offspring*—and so left more of them.

I've argued that the adaptive link account is both more parsimonious and clearer than the tracking account. My third and final point is that the adaptive link account does a much better job at actually illuminating the phenomenon that is to be explained, namely why there are widespread tendencies among human beings to make some evaluative judgments rather than others. To return to our original questions, why do we tend to judge that our survival is valuable, rather than worthless? Why do we tend to judge that we have special obligations to care for our children, rather than strangers or distant relatives? Why do we tend to view the killing of other human beings as a much more serious matter than the killing of plants or other animals? The adaptive link account has very good answers to such questions, of the general form that ancestors who made evaluative judgments of these kinds, and who as a result tended to respond to their circumstances in the ways demanded by these judgments, did better in terms of reproductive success than their counterparts. It is quite clear why creatures who judged their survival to be valuable would do much better than those who did not, and so on. Now compare the tracking account's explanation. It tries to answer these same questions by saying that these judgments are *true*: that survival *is* valuable, that we *do* have special obligations to care for our children, that the killing of human beings *is* more serious than the killing of plants or other animals. Such answers do not shed much light. In particular, the tracking account fails to answer three questions.

First, how does the tracking account explain the remarkable coincidence that so many of the truths it posits turn out to be exactly the same judgments that forge adaptive links between circumstance

and response—the very same judgments we would expect to see if our judgments had been selected on those grounds alone, regardless of their truth? The tracking account has no answer to this question that does not run right back into the parsimony and clarity problems just discussed.

Second, what does the tracking account have to say about our observed predispositions to make other evaluative judgments which (we may decide on reflection) are *not* true? For instance, we observe in human beings a deep tendency to think that the fact that someone is in an "out-group" of some kind is a reason to accord him or her lesser treatment than those in the "in-group." The adaptive link account offers a promising explanation of this, namely that having this evaluative tendency tended to promote reproductive success because those who possessed it tended to shower their assistance on those with a higher degree of genetic relatedness, or on those most able or likely to reciprocate. The tracking account's preferred explanation, however, falls flat, since in this case it is not plausible to answer that this evaluative predisposition developed because it is *true* that the fact that someone is in an "out-group" is a reason to accord him or her lesser treatment than those in the "in-group." More and more, many of us are coming to think that this is *not* true. The tracking account is thus left with nothing in the way of an explanation as to why we observe such deep tendencies to make the contrary judgment.[36]

Finally, consider the question of all those normative judgments that human beings *could* make but don't. As I have noted, the universe of logically possible evaluative judgments is huge, and we must think of all the possible evaluative judgments that we *don't* see—from the judgment that infanticide is laudable, to the judgment that plants are more valuable than human beings, to the judgment that the fact that something is purple is a reason to scream at it. Here again the adaptive link account has something potentially informative to point out, namely, that such judgments—or evaluative tendencies in these general sorts of directions—forge links between circumstance and response that would have been useless

[36] Someone drawn to the tracking account might argue that when it comes to our *corrected* evaluative judgments—for example, our judgment that membership in an "out-group" is no reason to accord a person lesser treatment—the tracking account provides a better explanation. But this is not so. It is perfectly compatible with the adaptive link account that we come to reject some of our basic evaluative tendencies on the basis of other evaluative judgments we hold.

or quite maladaptive as judged in terms of reproductive success. The tracking account has nothing comparably informative to say. It can just stand by and insist that such judgments are false—reaffirming our convictions but adding nothing to our understanding of why we have them.

To sum up, the set of evaluative judgments that human beings tend to affirm appears to be a disparate mishmash, ranging across all kinds of unrelated spheres and reflecting all kinds of unrelated values—some self-interested, others family-related, still others concerning how we should treat non-relatives and other forms of life, and so on. The power of the adaptive link account is that it exposes much of this seeming unrelatedness as an illusion; it illuminates a striking, previously hidden unity behind many of our most basic evaluative judgments, namely that they forge links between circumstance and response that would have been likely to promote reproductive success in the environments of our ancestors. The tracking account has no comparable explanatory power. Its appeal to the truth and falsity of the judgments in question sheds no light on why we observe the specific *content* that we do in human evaluative judgments; in the end, it merely reiterates the point that we *do* believe or disbelieve these things. When we couple this final point with the points about the parsimony and clarity of the adaptive link account as compared to the tracking account, it is clear which explanation we should prefer. The tracking account is untenable.

One last point remains in order to close off the Darwinian Dilemma. The tracking account was the most obvious and natural account for the realist to give of the relation between selective pressures on our evaluative judgments and the independent evaluative truths that he or she posits. In the wake of the tracking account's failure, one might think that the realist still has the option of developing some alternative account of this relation. But this is not so. Rather, insofar as realism asserts any relation at all between selective pressures on our evaluative judgments and evaluative truths, the position is forced to give a tracking account of this relation. The reason for this stems from the very nature of realism itself. The essence of the realist position is its claim that there are evaluative truths that hold independently of all of our evaluative attitudes. But because it views these evaluative truths as ultimately independent of our evaluative attitudes, the only way for realism *both* to accept that those attitudes have been deeply

influenced by evolutionary causes *and* to avoid seeing these causes as distorting is for it to claim that these causes actually in some way *tracked* the alleged independent truths. There is no other way to go. To abandon the tracking account—in other words, to abandon the view that selective pressures pushed us *toward* the acceptance of the independent evaluative truths—is just to adopt the view that selective pressures either pushed us *away from* or pushed us in ways that *bear no relation to* these evaluative truths. And to take *this* view is just to land oneself back in the first horn of the dilemma, in which one claims that there is no relation between selective pressures on our evaluative judgments and the posited independent truths. Realism about value, then, has no escape: it is forced to accept either the tracking account of the relation or else the view that there is no relation at all, and both of these options are unacceptable.[37]

7. FIRST OBJECTION: AN OBJECTION BY THE VALUE NATURALIST

At this point, an important objection remains open to the value naturalist, whose position I touched on only quickly in the argument of the previous section.[38] As we have seen, according to the value naturalist, evaluative facts are identical with (certain) natural facts. As also mentioned earlier, some value naturalists take the position that we may never be able to provide a reduction specifying exactly *which* natural facts evaluative facts are identical with, but let us set this point aside for the moment and assume for the sake of argument that it is agreed upon by all that evaluative facts are identical with such-and-such ordinary natural facts. Since these ordinary natural facts are in the same general category as facts about fires, predators, cliffs, and so on, presumably there is going to be a plausible evolutionary account available as to why we were selected to be able to track them, just as I myself have supposed there is a plausible evolutionary account available as to why we were selected to be able to track facts

[37] There is one other option available here: to posit that evaluative truths are in some way a *function* of our evaluative attitudes. This is exactly the way to go, in my view. But to make this move—to accept that evaluative truths are ultimately a function of our evaluative attitudes—is just to abandon value realism and embrace antirealism.

[38] I am indebted to Nishi Shah for very helpful comments and discussion regarding this objection and the material throughout this section.

about fires, predators, cliffs, and so on.[39] There may even be a good evolutionary account of why these natural facts are ones that we take a particularly strong interest in. It thus might seem that we have the outlines of a perfectly good answer to my question regarding the relation between evolutionary pressures on our evaluative judgments and independent evaluative truths. In particular, the relation is this: in ways roughly analogous to the ways in which we were selected to be able to track, with our non-evaluative judgments, facts about such things as fires, predators, and cliffs, so we were also selected to be able to track, with our evaluative judgments, *evaluative* facts, which are just identical with such-and-such natural facts.

This response, I will argue, ultimately just puts off a level the difficulties raised for realism by the Darwinian Dilemma. But first I need to distinguish between versions of value naturalism which count as genuinely realist in my taxonomy and those which don't. My taxonomy, while of course not the only legitimate understanding of realism, is far from ad hoc.[40] Rather, it zeroes in on the important question: Does the view in question understand evaluative truths as holding, in a fully robust way, independently of all our evaluative attitudes?

Suppose the value naturalist takes the following view. Given that we have the evaluative attitudes we do, evaluative facts are identical with natural facts *N*. But if we had possessed a completely different set of evaluative attitudes, the evaluative facts would have been identical with the very different natural facts *M*. Such a view does not count as genuinely realist in my taxonomy, for such a view makes it dependent on our evaluative attitudes *which* natural facts evaluative facts are identical with. On such a view, there is an important sense in which we need only alter our evaluative attitudes in order to change the evaluative facts, for by altering our evaluative attitudes we change which natural facts the evaluative facts are identical with. Views of this kind count as antirealist in my taxonomy, and as such are not a target of my argument; instead they escape the Darwinian Dilemma in the way I discuss in section 10.

[39] See note 35 above.

[40] Thanks to Dale Jamieson for pressing me to be explicit about this.

Peter Railton's account of individual non-moral good is an example of such a view.[41] According to Railton's proposal, roughly understood, an individual's non-moral good is identical to what that person would desire to desire under conditions of full information. Suppose, then, that what Ann would desire to desire under conditions of full information is (in part) her own longevity. In that case, her individual non-moral good is identical (in part) to her own longevity. But now suppose that Ann undergoes a significant change in her evaluative attitudes, such that it is no longer true of her that under conditions of full information she would desire to desire her own longevity. In that case, her individual non-moral good is no longer identical to her longevity, but is instead identical to something else (whatever it is that she'd now desire to desire under conditions of full information). There is an important sense in which Ann need only alter her evaluative attitudes in order to change the evaluative facts, for by altering her evaluative attitudes she changes which natural facts the evaluative facts are identical with. Railton's proposal therefore counts as antirealist in my taxonomy.[42]

In order to count as genuinely realist, then, a version of value naturalism must take the view that *which* natural facts evaluative facts are identical with is independent of our evaluative attitudes. For ease of expression, let us put the point this way: In order to count as realist, a version of value naturalism must take the view that facts about *natural-normative identities* (in other words: facts about exactly *which* natural facts evaluative facts are identical with) are independent of our evaluative attitudes. On the kind of view I have in mind, evaluative facts are identical with natural facts *N*, and even if our evaluative attitudes had been entirely different, perhaps not tracking those evaluative/natural facts *N* at all, but instead tracking some very different natural facts *M*, the evaluative facts *still* would have been identical

[41] See Railton (1986).

[42] One might object that while it depends on Ann's evaluative attitudes that her good is identical (in part) to her *longevity*, it presumably does not depend on her evaluative attitudes that her good is identical to *what she would desire to desire under conditions of full information*. This latter identity holds independently of her evaluative attitudes. Does that mean that Railton's view is realist after all on my taxonomy? The answer is no, since a view which *identifies* evaluative facts with facts about our evaluative attitudes (identifying them in particular with what those attitudes pick out as valuable under certain conditions) cannot properly be said to hold that evaluative facts are *independent* of our evaluative attitudes—any more than a view which identifies water with H_2O can properly be said to hold that facts about water are independent of facts about H_2O.

with natural facts *N*, and *not* natural facts *M*. On this sort of view, for example, Ann's individual non-moral good might be identical (in part) with her longevity, and even if Ann's evaluative attitudes were entirely different—such that she possessed no concern whatsoever for her longevity, and would fail to be concerned with it even if fully informed, and so on—her individual good would nevertheless *still* be identical (in part) with her longevity.

There is one artful way of achieving this sort of view that I also wish to rule out as genuinely realist, and that is by means of the move of "rigidifying."[43] Consider, for instance, a view which says that *which* natural facts evaluative facts are identical with is fixed in some way by our *actual* evaluative attitudes (in other words, by *our* attitudes, here and now). And suppose that our actual attitudes determine it that the evaluative facts are identical with natural facts *N*. On such a view, even if we had had entirely different evaluative attitudes, it *still* would have been the case that the evaluative facts are identical with natural facts *N*, since those are the ones picked out by our *actual* evaluative attitudes. Such a view is not genuinely realist in my taxonomy, however, for on such a view, there is no robust sense in which other creatures (including other possible versions of ourselves) would be making a *mistake* or *missing anything* if their evaluative attitudes tracked natural facts *M*, say, instead of natural facts *N*. For those other creatures could also pull the rigidifying move. And the upshot is that when we say "The good is identical to *N*" and they say "The good is identical to *M*," we will not be *disagreeing* with each other, with one of us correct and the other incorrect about which natural facts the good is identical to, but rather simply talking past each other, with the reference of our word 'good' fixed by *our* actual evaluative attitudes, and the reference of their word 'good' fixed by *their* actual evaluative attitudes.[44] We can of course go on using the word 'good' in our sense, according to which we're right to think that the good is identical to N, and they can go on using the word 'good' in their sense, according to which they're right to think that the good is identical to *M*, but there is, on such a view, no standard independent of all of our and their

[43] For discussion of the "rigidifying" move, see Darwall et al. (1992), 162-163.

[44] For similar points, see Hare (1952), 148-149; Horgan and Timmons (1991) and (1992); and Smith (1994), 32-35.

evaluative attitudes determining whose sense of the word 'good' is right or better; neither of us can properly accuse the other of having made a mistake.[45] For this reason, views that achieve "independence from our attitudes" by way of the rigidifying move do not count as genuinely realist in my taxonomy.

In sum, what I will call *genuinely realist versions of value naturalism* hold that *which* natural facts evaluative facts are identical with is independent of all our evaluative attitudes, and they do not achieve this result by means of the rigidifying move. When it comes to the case just sketched, a genuinely realist version of value naturalism will hold that even if the two communities' uses of the word 'good' track different natural properties, the communities are nevertheless (at least potentially) using the word 'good' in the same sense—genuinely disagreeing with one another about the correct natural-normative identity—and that there is a fact of the matter about which (if either) of us is right that obtains independently of all of our and their evaluative attitudes.[46]

Genuinely realist versions of value naturalism are vulnerable to the Darwinian Dilemma. To see this, the first thing to note is the following. How, according to these views, do we figure out the correct natural-normative identities? We may assume that the answer is the one given by value naturalists such as Nicholas Sturgeon and David Brink. Sturgeon writes that if a full account of which natural facts evaluative facts are identical with is to be had, then this account "will have to be derived from our best moral theory, together with our best theory of the rest of the world."[47] And Brink agrees that "Determination of just which natural facts and properties constitute which moral facts and properties is a matter of substantive moral theory."[48] These theorists do not propose some completely new approach to substantive moral theory; on the contrary, they think we should proceed in roughly the way we currently do proceed—starting with our existing fund of evaluative judgments, giving more weight to those evaluative judgments which strike us as correct if anything is (for instance, the judgment that Hitler was

[45] This assumes that each community has correctly identified the natural properties tracked by its own evaluative attitudes.

[46] Brink seems to take such a view in section VII of Brink (2001).

[47] Sturgeon (1985), 59.

[48] Brink (1989), 177-178.

morally depraved[49]), and then working to bring our evaluative judgments into the greatest possible coherence with each other and with our best scientific picture of the rest of the world.

It's at this point that the Darwinian Dilemma kicks in. The genuinely realist value naturalist posits that there are *independent facts about natural-normative identities*. But the value naturalist also holds that in trying to figure out what those identities are, we will have to rely very heavily on our existing evaluative judgments. Yet, as we have seen, those evaluative judgments have been tremendously influenced by Darwinian selective pressures. And so the question arises: What is the relation between evolutionary influences on our evaluative judgments, on the one hand, and the independent truths about natural-normative identities posited by the realist, on the other? In trying to figure out which natural facts evaluative facts are identical with, we have no option but to rely on our existing fund of evaluative judgments: I judge that Hitler was morally depraved, for instance, and in doing so steer toward the view that the evaluative fact of someone's being morally depraved is roughly identical to her having a psychology of such-and-such a character (naturalistically described)—a psychology that is like Hitler's in certain relevant ways (exactly *which* ways is to be determined by relying on further evaluative judgments of mine). But in relying on these and other evaluative judgments, I rely on judgments that are saturated with evolutionary influence. What then is the relation between that influence and the independent truths I'm seeking to uncover—these independent truths about natural-normative identities?

As before, the realist has two options: he or she may either assert or deny a relation. Suppose that the realist denies that there is any relation. As before, a highly skeptical result follows. If there is no relation whatsoever between evolutionary influences on our evaluative judgments and independent truths about natural-normative identities, then all our hypothesizing about natural-normative identities is hopelessly contaminated with illegitimate influence. Due to the distorting pressure of Darwinian forces, we are, for all we know, tracking natural facts *M* with our evaluative judgments, whereas we ought to be

[49] See Sturgeon (1985).

tracking (say) the entirely different set of natural facts *N*, the ones which are *really* identical with evaluative facts.

Suppose, on the other hand, the realist value naturalist claims that there *is* a relation between evolutionary pressures on our evaluative judgments and independent truths about natural-normative identities. Here, as before, the realist's only option for spelling out this relation is some version of a tracking account, according to which we were somehow selected to be able to track with our evaluative judgments independent facts about natural-normative identities. But if the tracking account failed as a scientific explanation when it came to arguing that we were selected to track independent evaluative truths, then it will fail even more seriously when it comes to arguing that we were selected to track independent facts about natural-normative identities. For it is even more obscure how tracking something as esoteric as independent facts about natural-normative identities could ever have promoted reproductive success in the environment of our ancestors. The adaptive link account again wins out: the best explanation of why human beings tend to make some evaluative judgments rather than others is not that these judgments constituted an awareness (however imperfect) of independent facts about natural-normative identities, but rather that the relevant evaluative tendencies forged links between our ancestors' circumstances and their responses which tended to promote reproductive success.

I conclude that any genuinely realist version of value naturalism runs headlong into the same basic dilemma I've been sketching. To the extent that a view insists on there being evaluative facts which hold independently of all our evaluative attitudes, it is impossible to reconcile that view with a recognition of the role that Darwinian forces have played in shaping the content of our values. Once we become fully conscious of this powerful "bidding from the outside" with respect to our evaluative judgments, I suggest, our response must be to adjust our metaethical view so as to become antirealists.

8. SECOND OBJECTION: THE BYPRODUCT HYPOTHESIS

While the first objection belongs to value naturalists, a second objection might be voiced by realists generally. Confronted with the Darwinian Dilemma, the realist may suggest the following

alternative evolutionary hypothesis. Perhaps the human ability to grasp independent evaluative truths was not itself selected for, but is instead the byproduct or outgrowth of some other capacity which *does* have an explanation in terms of natural selection (or else some other, non-selective evolutionary explanation). Many human capacities, after all, are like this: our ability to do astrophysics, for instance, was surely not itself directly selected for, but is instead the byproduct or outgrowth of other capacities which likely do have an explanation in terms of natural selection. Perhaps in some similar fashion our ability to grasp independent evaluative truths has emerged as a byproduct or outgrowth of some other capacity—call it *capacity C*.

This objection has not been properly developed until the realist has explained exactly what capacity *C* is, how it evolved, and what relation it bears to the capacity to grasp independent evaluative truths. However these details might be filled out, though, the Darwinian Dilemma arises again for such a proposal—this time with regard to capacity *C*. In particular, the question for the realist becomes this: What relation, if any, does the realist claim obtained between the evolution of capacity *C* and the independent evaluative truths that he or she posits?

Suppose the realist answers "no relation." Suppose, in other words, that the realist claims that the capacity to grasp independent evaluative truths arose as a complete fluke, as the wholly incidental byproduct of some other capacity *C* that was selected for on the basis of factors that had nothing whatsoever to do with the task of grasping evaluative truths. If the realist takes this route, then the coincidence point is triggered again: how incredible (not to mention how extraordinarily convenient for the realist) that, as a matter of sheer coincidence, a capacity happened to arise (as the entirely incidental byproduct of some totally unrelated capacity *C*) which operates to grasp precisely the sort of independent truths postulated by the realist.

To this charge of an implausible coincidence, the realist might protest that this sort of thing happens all the time in evolution—in other words, that one trait arises as the completely incidental byproduct of selection for some other trait. While it is quite true that this happens, the suggestion that this is what happened in the case of our ability to grasp independent evaluative truths is very implausible

given the nature of the trait in question. The task of grasping independent evaluative truths presumably requires a highly specialized, sophisticated capacity, one specifically attuned to the evaluative truths in question. The capacity at issue is not a simple, brute sort of feature—not, presumably, if we have any reasonable chance of grasping the truths posited by the realist. But the more complicated and uniquely specialized a faculty is, the less plausible it is to hypothesize that it could have arisen as a sheer fluke, as the purely incidental byproduct of some unrelated capacity that was selected for on other grounds entirely. It is completely implausible, for instance, to suggest that the human eye in its present developed form emerged as the purely incidental byproduct of selection for some other, unrelated capacity.[50] I suggest that it is no more plausible to claim that the sophisticated ability to grasp independent evaluative truths emerged as such a byproduct.

The realist's other option is to maintain that there *is* some relation between the evolution of capacity *C* and the independent evaluative truths that he or she posits. According to this proposal, it is no fluke that the ability to grasp evaluative truths emerged as a byproduct of capacity *C*, because there is some relation between capacity *C* and the capacity to grasp evaluative truths. But now the challenge for the realist is to explain what this relation is. And it's hard to see how the realist can say anything except that capacity *C*, whatever it may be, involves at least some *basic* sort of ability to grasp independent evaluative truths, of which our present-day ability to grasp evaluative truths is a refined extension, in much the same way that our present-day ability to do astrophysics is presumably a refined extension of more basic abilities to discover and model the physical features of the world around us.[51] But at this point the realist has to give some account of how this more basic sort of ability to grasp independent evaluative

[50] Of course it may be that the very first *rudiments* of the human eye emerged as the purely incidental byproduct of selection for some other capacity, and then these rudiments conferred advantages of their own and began to be selected for more directly. And the realist might claim that something similar could have happened in the case of our ability to grasp independent evaluative truths: first came an extraordinarily rudimentary form of this ability, and then the ability began to be selected for more directly. The realist may certainly pursue this proposal. But if he or she does so, then he or she has opted for the second horn of the dilemma, claiming that there *was* indeed a relation between the operation of selective pressures on the relevant capacity and the independent evaluative truths he or she posits.

[51] For related discussion, see Gibbard (2003), 265-267.

truths arose. And given what has to be the complexity and specialization of even this more basic ability (a point of comparison is the complexity and specialization of the more basic abilities on which the ability to do astrophysics is based), it is implausible to suggest that the emergence of this more basic ability was a mere fluke. The only alternative to saying that the emergence of this ability was a fluke is to claim that we were in some way selected to *track* the independent evaluative truths posited by the realist, yet this proposal, for the reasons I've already given, is scientifically unacceptable. The byproduct hypothesis, while it pushes matters off a step by hypothesizing an intervening capacity or set of capacities, does not permit escape from the Darwinian Dilemma for the realist about value.

9. THIRD OBJECTION: THE BADNESS OF PAIN AS AN ALLEGED INDEPENDENT TRUTH ABOUT VALUE

The case of physical pain—for instance, in the various forms associated with burns, cuts, bruises, broken bones, nausea, and headaches—serves as one of the strongest temptations toward realism about value. Realists frequently appeal to the case of pain when defending their views,[52] and when presented with the Darwinian Dilemma, another such appeal may seem attractive. One possibility is for the realist to argue along the following lines. There are obvious evolutionary explanations of why we tend to feel physical pain when we do: roughly, we tend to feel it in conjunction with bodily conditions or events that diminish reproductive success, such as a cut to the skin or a blow to the head. Pain itself, moreover, due to its very nature, is bad independently of whatever evaluative attitudes we might hold. Together these points provide at least a rough answer to the question of what the relation is between evolutionary pressures and independent evaluative truths: in short, evolutionary pressures led us to feel pain under such-and-such kinds of circumstances, and that experience is, of its very nature, bad independently of all our evaluative attitudes, its badness therefore demanding a realist construal. Taking a slightly different tack, the realist might also argue: it is presumably no mystery from an evolutionary point of view why

[52] One prominent such appeal is Nagel's discussion of pain in Nagel (1986), 156-162.

we're able to "track" pain, and pain itself is an evaluative fact, bad independently of all our evaluative attitudes. So it's fairly clear how we were selected to track (at least these) independent evaluative facts.

While such ideas have some intuitive appeal, evolutionary considerations—including one more application of the Darwinian Dilemma—help us to see how the badness of pain *does* in fact depend on our evaluative attitudes, revealing a realist understanding of its badness to be mistaken. For the purposes of the ensuing discussion, let us focus on the following evaluative claim: Someone's pain counts as a reason for *that person* to do what would avoid, lessen, or stop it.[53] It is no doubt plausible to think that this evaluative truth holds independently of all our evaluative attitudes, just as a realist about value claims. Of course many realists would maintain in addition that someone's pain also counts as a reason for *other people* to do what would avoid, lessen, or stop it, and that this evaluative truth too holds independently of all our evaluative attitudes. I take it, however, that realism about the badness of pain *for the person whose pain it is* is a more formidable target than realism about the badness of pain *for people whose pain it isn't*, and so I'll focus on the former sort of realism. If I can raise questions about realism's viability in this toughest, most basic case, then questions about the viability of the latter sort of realism will follow *a fortiori*. In the remainder of this discussion, then, when I talk about the "badness of pain" or say that "pain is bad," I'm using such expressions as a shorthand way of talking about the badness of pain *for the person whose pain it is*.

To start out, we need to clarify what is meant by 'pain.' Consider, for the sake of argument, the following proposal:

> Pain is a sensation such that the creature having the sensation unreflectively takes that sensation to count in favor of doing whatever would avoid, lessen, or stop it.

Let me now call attention to several points about this definition.

---

[53] More precisely: Someone's pain counts as a *pro tanto* reason for that person to do what would avoid, lessen, or stop it, where a *pro tanto* reason is a reason that is good as far as it goes, but which may be outweighed by other considerations. (So, for example, the fact that it will be painful is a good reason *not* to go to the dentist, so far as that reason goes, but it may well be the case that this reason is counterbalanced and ultimately outweighed by good reasons in favor of going.)

First, the definition draws on the notion of unreflective valuing that I introduced in section 4. In its most rudimentary form of all, such valuing might involve some primitive conscious experience of a motivational "push" or "pull" in the direction of a certain behavior; in its more sophisticated forms, such valuing might involve some emotional or proto-emotional—yet still non-reflective and non-linguistic—experience of a behavior as "demanded" or "counted in favor of" by the circumstances—an experience such as a chimpanzee or grizzly bear might be capable of. Thus, according to the definition of pain under consideration, if a creature does not at an *unreflective* level take a given sensation to "count in favor of" doing what would alleviate it—in other words, if a creature has a sensation that it in no way feels motivationally "pushed" or "pulled" to avoid, lessen, or stop—or if, more complexly, the creature feels no distress at the sensation's presence, no relief when it subsides, and so on—then the sensation in question does not count as a pain. Since unreflective valuing is something that many or most animals are capable of, this definition is consistent with the idea that many or most animals can experience pain.

Second, according to the definition of pain at hand, the word 'pain' technically refers to the sensation that is the object of the specified negative unreflective evaluative reaction, as opposed to the composite of the sensation plus the unreflective evaluative reaction. This is analogous to the way in which the expression 'Juliet's beloved' refers to Romeo, as opposed to the composite of Romeo plus Juliet's love for him. Yet this does not imply that the unreflective evaluative reaction is not a necessary element of the pain experience. On the contrary, according to the definition at hand, just as Juliet's love for Romeo is what makes Romeo her beloved, so a creature's negative unreflective evaluative reaction to a sensation is (at least part of) what makes that sensation a pain.

Third, accepting this definition of pain involves accepting that there are two elements involved in the experience of pain: a sensation plus an unreflective evaluative reaction to that sensation.[54] But it

[54] My treatment of pain has been influenced by Christine M. Korsgaard's treatment in Lecture 4 of Korsgaard (1996), particularly when it comes to the idea that there are two elements involved in the experience of pain. One of the most important differences between Korsgaard's treatment of pain and mine, however, is this. Korsgaard takes the position that pain itself never provides a reason for its sufferer to do what would alleviate it; it is rather merely the perception of some *other* reason that the

should be noted that accepting the definition does *not* involve commitment to the idea that it will always or even often be possible to *separate* these two elements in a creature's experience of pain. On the contrary, one might think that in many or perhaps all cases of pain, the sensation and the unreflective evaluative reaction to it are merged inextricably into a single, unified experience—such that it is impossible to have one element of the pain experience without the other, or even to be able to tell the two elements apart when one examines one's pain introspectively. But none of this means that no theoretical analysis of pain into these two elements is possible. Compare the moment when Juliet sees that Romeo is dead. It would be impossible for most of us ever to separate the sight of our beloved's dead body from our evaluative reaction to it, but this does not mean that there is no distinction to be drawn between the two elements of the experience. The experience of pain may well be similar: it may often be impossible, as a practical or introspective matter—but not as a theoretical matter—to separate the sensation that is involved from one's unreflective taking of that sensation to be bad.

Fourth, as it so happens, it appears that it *is* sometimes possible to separate out the two elements involved in the experience of pain. For example, patients who have been suffering from terrible pain sometimes report that after receiving certain drugs or undergoing certain surgeries they feel the same sensation as before and yet it no longer bothers them.[55] Such cases lend support to the idea that there are indeed two distinct elements involved in the experience of pain. Note, however, that according to the definition we're considering, the moment such a separation occurs in practice—in other words, the moment a pain sensation ceases to be the object of a negative unreflective evaluative reaction—it thereby ceases to be a pain, and becomes just another sensation.

Understood as a statement of a *necessary* condition of a sensation's being a pain (and that is how I will be understanding it), I think the proposed definition of pain is a plausible one.[56] Nevertheless, my

---

sufferer has. In contrast, on my view, pain itself (at least virtually always) *does* provide a reason for its sufferer to do what would alleviate it. I take this to be the more intuitively plausible position.

[55] For references, see Richard J. Hall's discussion, to which I am indebted, in Hall (1989).

[56] One might well doubt whether the definition states a sufficient condition of a sensation's being a pain. One might, for instance, be concerned about the way in which the definition would apparently count

argument does not depend on one's accepting it. Rather, I'll offer another argument in the form of a dilemma (calling this dilemma the *Pain Dilemma* to distinguish it from the Darwinian Dilemma). I'll argue that the realist about the badness of pain runs into trouble no matter whether he or she accepts or rejects this statement of a necessary condition of a sensation's being a pain.

Suppose, to begin with, that the definition of pain I have just sketched is rejected by the realist. In that case, pain is understood as a sensation such that the creature having it does *not* necessarily unreflectively experience that sensation as counting in favor of whatever would avoid, lessen, or stop it. Instead, pain is given some other definition, according to which there is no inconsistency in supposing that an individual or even an entire species could unreflectively take pain sensations to count in favor of doing whatever would *bring them about and intensify them* rather than whatever would stop or lessen them. On this view, it is perfectly possible, as a conceptual matter, that instead of disliking pain the way we all happen to do, we could naturally enjoy it and be inclined to seek it out, unreflectively experiencing it as counting in favor of what would cause it—just the way we feel about the sensations associated with a massage, for example.

I take it that if pain is understood in such a way, then realism about its status as a reason to do what would avoid, lessen, or stop it is no longer very tempting. Take, for example, the case of a patient who is having a pain sensation (defined however the person opting for the first horn of the Pain Dilemma

---

sensations like the taste of something rancid, the smell of rotten eggs, or the sound of fingernails on a chalkboard as pains, since we have unreflective negative evaluative reactions to these sensations too. One might also think that a complete definition would say more about how pain involves not only an unreflective negative evaluative reaction to a *sensation*, but also to the *bodily condition* of which that sensation is (at least in many cases) a perception. With regard to this latter point, however, the definition at hand already addresses it, at least up to a point. For note that the definition may be understood as positing *two* unreflective evaluative reactions: first, an unreflective taking of the *sensation* in question to be bad, and second, an unreflective taking of *whatever would avoid, lessen, or stop the sensation* to be good. Assuming that an underlying bodily condition is what is causing the sensation in question, and that the elimination of that bodily condition would stop it, the second unreflective evaluative element involves taking that bodily condition to be bad and the elimination of it to be good. So, for example, if one is experiencing pain due to a broken leg, part of what this involves, according to the definition, is the unreflective taking of whatever would stop this sensation to be good, which in turn involves the unreflective taking of the healing of one's leg to be good (and the unreflective taking of its current broken condition to be bad). Thanks to Hilla Jacobson for discussion of related issues.

would like). And suppose that thanks to medication, the person is no longer bothered by this sensation in the slightest, feeling no motivation whatsoever to avoid, lessen, or stop it, no experience of the sensation as something to be gotten rid of. (Someone opting for the first horn of the Pain Dilemma, by definition, must admit that this is possible.) Suppose further that with the help of some new miracle drug, the patient comes positively to *enjoy* the sensation in question—which is to say that he unreflectively feels inclined to treat the sensation as counting in favor of whatever would bring it about or intensify it. (Again, someone opting for the first horn of the Pain Dilemma must agree that this is possible.)

A realist about the badness of pain so understood would have to say that even in such a case, the *sensation itself* still provides the patient with reason to do whatever would avoid, lessen, or stop it, and that this person is making a *mistake* if he goes ahead and endorses his unreflective tendency to think that the sensation counts in favor of doing what would bring it about or intensify it. Such a position is not very plausible. Of course it might be the case that this patient has *other* reasons to take actions that would, as a matter of course, *happen* to stop the sensation in question; indeed, this is likely to be the case if the sensation in question is being caused by an underlying bodily condition that is bad for the person in other respects—for example, because the bodily condition is a hindrance to the pursuit of the person's other ends. But it does not seem plausible to insist that in such a case the *sensation itself* constitutes a reason for the person to do whatever would stop it. On the contrary, if the bodily condition responsible for the sensation is (for example) a broken leg, then it seems that our patient would be perfectly correct to reason as follows: "The fact that allowing my leg to heal would end this sensation that I've come to enjoy (thanks to the miracle drug) is a small thing that counts *against* letting it heal, but all things considered I should go ahead and let it heal anyway, since after all I need to be able to walk to pursue most things that are important to me." Thus, if pain is understood as a sensation such that it is perfectly conceivable that we could unreflectively be inclined to view it as counting in favor of what would bring it about or intensify it, then realism about its status as a reason to do what would avoid, lessen, or stop it is unattractive.

Suppose, however, there is remaining doubt and some are still tempted by a realist position on the badness of pain so understood. It's at this point that the Darwinian Dilemma arises again for the realist. To see how, suppose again that pain is given some definition according to which it is *not* a necessary feature of pain that we unreflectively experience it as counting in favor of what would avoid, lessen, or stop it. In that case, the following becomes a legitimate scientific question: given that it is perfectly conceivable that we all could have ended up taking pain sensations to count in favor of what would cause them and intensify them rather than in favor of what would lessen them and stop them, what explains the fact that such a huge percentage of us so consistently do the latter? Here, as in earlier cases, there is a powerful evolutionary answer. I've left it open how the person opting for the first horn of the Pain Dilemma is defining pain (so long as that definition makes no reference to the idea that pain is a sensation that we unreflectively take to count in favor of what would stop it). But if the proposed definition is to be plausible at all, then it will pick out (predominantly, one assumes) sensations associated with the sorts of bodily conditions that we normally consider painful, such as cuts, burns, bruises, broken bones, and so on. And it is of course no mystery whatsoever, from an evolutionary point of view, why we and the other animals came to take the sensations associated with bodily conditions such as these to count in favor of what would avoid, lessen, or stop them rather than in favor of what would bring about and intensify them. One need only imagine the reproductive prospects of a creature who relished and sought after the sensations of its bones breaking and its tissues tearing; just think how many descendants such a creature would leave in comparison to those who happened to abhor and avoid such sensations.

As in earlier cases, the realist faces a problem when confronted with such an explanation. For once again we see that there is a striking coincidence between the content of the independent evaluative truth posited by the realist, on the one hand, and the content that evolutionary theory would lead us to expect, on the other. The realist tells us that it is an independent evaluative truth that pain sensations (however he or she defines them) are bad, and yet this is precisely what evolutionary theory would have predicted that we come to think. And once again the realist is unable to give any good account of this coincidence. To insist that the coincidence is *mere* coincidence is implausible. The realist's alternative,

here as in earlier cases, is to defend some sort of tracking account, according to which we were selected to be able to discern independent evaluative truths, among them the truth that these pain sensations (however the realist is defining them) are bad. Yet here as in earlier cases, the tracking account is scientifically unacceptable. In order to explain why we came to think that these sensations are bad, we need make no reference whatsoever to the *fact* that they are bad; we need only point out how it tended to promote reproductive success to *take* them to be bad (due to their connection with bodily conditions that tended to diminish reproductive success).

The realist, then, is forced to the other horn of the Pain Dilemma. To salvage realism about the badness of pain, he or she is forced to understand pain as a sensation such that the creature who has it unreflectively takes that sensation to count in favor of whatever would avoid, lessen, or stop it. But now notice what this means. In order to salvage his or her view of pain as bad independently of our evaluative attitudes, the realist must admit that pain's badness depends on its being a sensation such that the creature who has it is unreflectively inclined to *take* it to be bad. But this, in turn, is just to admit that its badness depends in an important sense on our evaluative attitudes—in particular, on our being unreflectively inclined to take it to be bad. Pain may well be bad, in other words, but if it is so, its badness hinges crucially on our unreflective evaluative attitudes toward the sensation which pain is. The realist is thus forced to recognize the role of our evaluative attitudes in determining the disvalue of pain. Though initially plausible, it is a mistake to say that pain is bad independently of our evaluative attitudes. Pain, if it is plausibly to be construed as bad independently of our other evaluative attitudes, must be understood as a sensation such that we have a certain evaluative attitude toward it—and it's that evaluative attitude which (at least in part) *makes* the sensation bad.

I conclude that appeals to pain are not a promising avenue for the realist who wishes to escape the Darwinian Dilemma. Appeals to pleasure are no more promising, for there exists an analogous argument against realism about pleasure's status as a reason to do what would bring it about. This argument would center around, and propose an analogous dilemma with regard to, the following definition of pleasure:

pleasure is a sensation such that the creature having it unreflectively takes that sensation to count in favor of doing whatever would bring it about, intensify it, or make it continue.

10. HOW ANTIREALISM SIDESTEPS THE DARWINIAN DILEMMA

Let me now sketch how antirealist views on the nature of value sidestep the dilemma for realism that I have described in this paper. Antirealist views understand evaluative facts or truths to be a function of our evaluative attitudes, with different versions of antirealism understanding the exact nature of this function in different ways. For instance, according to the constructivist view mentioned in section 2, the truth of the evaluative judgment that '*X* is a reason for agent *A* to *Y*' is a function of *A*'s evaluative attitudes—in particular, of whether that judgment would be among *A*'s evaluative judgments in reflective equilibrium. Such a view, as I pointed out earlier, leaves room for the possibility of evaluative error. If, for example, *A* thinks that the fact that someone is a member of some "out-group" is a reason for him to accord that person lesser treatment, then *A*'s judgment is mistaken if that judgment would not be among his evaluative judgments in reflective equilibrium. It is not my purpose to develop or defend such a view here. The point is to give one example of an antirealist view, and to emphasize that antirealist views can leave room for the possibility of evaluative error, even though the standards determining what counts as an error are understood ultimately to be "set" by our own evaluative attitudes.

What then does an antirealist say about the relation between evaluative truths and the evolutionary influences that have shaped our evaluative judgments? First of all, the antirealist opts for what I have said is the more plausible horn of the Darwinian Dilemma, arguing that *of course* there is some relation at work here—of course it is no coincidence that there is such a striking overlap between the content of evaluative truths and the content that natural selection would have tended to push us toward. Of course it's no coincidence that, say, breaking one's bones *is* bad and that's also exactly what evolutionary theory would have predicted we think. But whereas the realist is forced to offer the scientifically unacceptable tracking explanation of this overlap, the antirealist is able to give a very different account.

According to the antirealist, the relation between evolutionary influences and evaluative truth works like this. Each of us begins with a vast and complicated set of evaluative attitudes. We take the breaking of our bones to be bad, we take our children's lives to be valuable, we take ourselves to have reason to help those who help us, and so on. Our holding of each of these evaluative attitudes is assumed by the antirealist to have some sort of causal explanation, just like anything else in the world. And the antirealist grants without hesitation that one major factor in explaining why human beings tend to hold some evaluative attitudes rather than others is the influence of Darwinian selective pressures. In particular, the antirealist has no problem whatsoever with the adaptive link account, if something along those lines turns out to be the best explanation. These and other questions about the best causal explanations of our evaluative attitudes are left in the hands of scientists. Whatever explanation the natural and social scientists ultimately arrive at is granted, and then evaluative truth is understood as a function of the evaluative attitudes we have, however we originally came to have them. Take the constructivist view I've been mentioning as an example. What exactly is the relation between selective pressures and evaluative truth on this view? It may be put this way: evaluative truth is a function of how all the evaluative judgments that selective pressures (along with all kinds of other causes) have imparted to us stand up to scrutiny in terms of each other; it is a function of what would emerge from those evaluative judgments in reflective equilibrium.

Where the realist's tracking account and the antirealist's account divide, then, is over the *direction of dependence* that they take to be involved in the relation between evaluative truths and the evolutionary causes which influenced the content of our evaluative judgments. The realist understands the *evaluative truths* to be prior, in the sense that evolutionary causes are understood to have selected us to track those independent truths. The antirealist, on the other hand, understands the *evolutionary causes* to be prior, in the sense that these causes (along with many others) gave us our starting fund of evaluative attitudes, and evaluative truth is understood to be a function of those attitudes. Both accounts offer an explanation of why it is no coincidence that there is significant overlap between evaluative truths and the kinds of evaluative judgments that natural selection would have pushed us in the direction of. The

difference is that the antirealist account of the overlap is consistent with science. Antirealism explains the overlap not with any scientific hypothesis such as the tracking account, but rather with the metaethical hypothesis that value is something that arises as a function of the evaluative attitudes of valuing creatures—attitudes the content of which happened to be shaped by natural selection. The breaking of our bones *is* bad, in other words, and we're well aware of this. But the explanation is not that it is true independently of our attitudes that the breaking of our bones is bad and we were selected to be able to notice this; the explanation is rather that we were selected to *take* the breaking of our bones to be bad, and this evaluative judgment withstands scrutiny from the standpoint of our other evaluative judgments (to speak, for example, in the voice of the constructivist antirealist).[57]

11. CONCLUSION

By understanding evaluative truth as ultimately prior to our evaluative judgments, realism about value puts itself in the awkward position of having to view every causal influence on our evaluative judgments as either a tracking cause or a distorting cause. In the end, this is a difficult position to be in no

[57] In objection to this section's argument, someone might try to replicate against antirealism the move I made in section 7 against realist versions of value naturalism. In particular, an objector might charge that the antirealist, in arriving at his or her view on the way in which evaluative truth is a function of our evaluative attitudes, must rely heavily on our substantive evaluative judgments (regarding practical reasons). Since those judgments are contaminated with evolutionary influence (and since the antirealist presumably wishes to say that his or her metaethical view is true independently of our evaluative attitudes), the objector might argue that the Darwinian Dilemma threatens antirealism as much as it does realism. There is not space to address this objection in depth here, but in brief my reply is that in arriving at his or her metaethical view, the antirealist does *not* need to rely on our substantive evaluative judgments (regarding practical reasons). This may be seen by imagining an alien investigator who (1) quite recognizably possesses evaluative concepts; (2) accepts evaluative judgments (regarding practical reasons) with *entirely different* substantive content than our own; and who nevertheless (3) arrives at the same metaethical view as the human antirealist; and (4) does so based on the exact same considerations. Examples of such considerations might include the Darwinian Dilemma itself (which the alien investigator could accept), or the types of considerations that David Lewis offers in favor of his analysis of value in Lewis (1989). (Lewis's proposal counts as antirealist in my taxonomy; see his remark that if we had been disposed to value seasickness and petty sleaze, then in one good sense (though not the only sense) of 'value,' "it would have been true for us to say 'seasickness and petty sleaze are values'" (133).) I assume here for the sake of argument that the alien investigator would share our judgments regarding *epistemic reasons*; this assumption is complicated by the fact that (as mentioned in note 2) I believe the Darwinian Dilemma can be extended to apply against realism about epistemic reasons. A full discussion of this objection would address such complications.

matter what kind of causal influence is at issue. I have focused on the case of Darwinian influences on our evaluative judgments because I think it raises the problem for realism in a particularly acute form. In principle, however, an analogous dilemma could be constructed using *any* kind of causal influence on the content of our evaluative judgments. For the argument to work, two conditions must hold. First, the causal influence in question must be extensive enough to yield a skeptical conclusion if the realist goes the route of viewing those causes as distorting. Second, it must be possible to defeat whatever version of the tracking account is put forward with a scientifically better explanation.

At the end of the day, then, the dilemma at hand is not distinctly Darwinian, but much larger. Ultimately, the fact that there are *any* good scientific explanations of our evaluative judgments is a problem for the realist about value. It is a problem because realism must either view the causes described by these explanations as distorting, choosing the path that leads to normative skepticism or the claim of an incredible coincidence, or else it must enter into the game of scientific explanation, claiming that the truths it posits actually play a role in the explanation in question. The problem with this latter option, in turn, is that they don't. The best causal accounts of our evaluative judgments, whether Darwinian or otherwise, make no reference to the realist's independent evaluative truths.

Consider again the old dilemma whether things are valuable because we value them or whether we value them because they are valuable. The right answer, according to the view I've been suggesting, is somewhere in between. Before life began, nothing was valuable. But then life arose and began to value—not because it was recognizing anything, but because creatures who valued (certain things in particular) tended to survive. In this broadest sense, valuing was (and still is) prior to value. That is why antirealism about value is right. But I've emphasized that antirealist views can make room for the possibility of evaluative error, such that we can be wrong about any given evaluative judgment even as we recognize that the standards for such errors are ultimately "set" by our own evaluative attitudes. Because of this, talk of normative perception still makes sense. Now that there are creatures like us with marvelously complicated systems of valuings up and running, it is quite possible to come to value something because one recognizes that it has a value independent of oneself—not in the realist's sense,

but in an antirealist's more modest sense. Thus, although valuing ultimately came first, value grew to be able to stand partly on its own. It grew to achieve its own, limited sort of priority over valuing—a priority that we can understand while at the same time being fully conscious of great biddings from the outside.

REFERENCES

Avers, C. J. (1989): *Process and Pattern in Evolution*, New York: Oxford University Press.

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

Barkow, J. H., Cosmides, L. and Tooby, J. (eds.) (1992): *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, New York: Oxford University Press.

Blackburn, S. (1984): *Spreading the Word: Groundings in the Philosophy of Language*, Oxford: Clarendon Press.

Blackburn, S. (1993): *Essays in Quasi-Realism*, New York: Oxford University Press.

Blackburn, S. (1998): *Ruling Passions*, Oxford: Clarendon Press.

Block, N. J. and Dworkin, G. (1974): 'IQ, Heritability, and Inequality, Part 2,' *Philosophy and Public Affairs* 4, 40-99.

Boyd, R. (1988): 'How to Be a Moral Realist,' in Geoffrey Sayre-McCord (ed.), *Essays on Moral Realism*, Ithaca: Cornell University Press.

Brink, D. O. (1989): *Moral Realism and the Foundations of Ethics*, Cambridge: Cambridge University Press.

Brink, D. O. (2001): 'Realism, Naturalism, and Moral Semantics,' *Social Philosophy & Policy* 18, 154-176.

Buss, D. M. (1999): *Evolutionary Psychology: The New Science of Mind*, Boston: Allyn and Bacon.

Darwall, S., Gibbard, A. and Railton, P. (1992): 'Toward *Fin de siècle* Ethics: Some Trends,' *Philosophical Review* 101, 115-189.

Dworkin, R. (1996): 'Objectivity and Truth: You'd Better Believe It,' *Philosophy and Public Affairs* 25, 87-139.

Gibbard, A. (1990): *Wise Choices, Apt Feelings*, Cambridge, MA: Harvard University Press.

Gibbard, A. (2003): *Thinking How to Live*, Cambridge, MA: Harvard University Press.

Gould, S. J. and Lewontin, R. C. (1979): 'The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme,' *Proceedings of the Royal Society of London, Series B, Biological Sciences* 205, 581-598.

Hall, R. J. (1989): 'Are Pains Necessarily Unpleasant?,' *Philosophy and Phenomenological Research* 49, 643-659.

Hamilton, W. D. (1963): 'The Evolution of Altruistic Behavior,' *American Naturalist* 97, 354-356.

Hamilton, W. D. (1964): 'The Genetical Evolution of Social Behavior, I and II,' *Journal of Theoretical Biology* 7, 1-52.

Hare, R. M. (1952): *The Language of Morals*, Oxford: Clarendon Press.

Horgan, T. and Timmons, M. (1991): 'New Wave Moral Realism Meets Moral Twin Earth,' *Journal of Philosophical Research* 16, 447-465.

Horgan, T. and Timmons, M. (1992): 'Troubles for New Wave Moral Semantics: The Open Question Argument Revived,' *Philosophical Papers* 21, 153-175.

Kant, I. (1785 [1959]): *Foundations of the Metaphysics of Morals*, L. W. Beck (trans.), New York: Macmillan Library of Liberal Arts.

Korsgaard, C. M. (1996): *The Sources of Normativity*, Cambridge: Cambridge University Press.

Lewis, D. (1989): 'Dispositional Theories of Value,' *Proceedings of the Aristotelian Society*, suppl. 63, 113-137.

Nagel, T. (1986): *The View From Nowhere*, Oxford: Oxford University Press.

Nozick, R. (1981): *Philosophical Explanations*, Cambridge, MA: The Belknap Press of Harvard University Press.

Pigliucci, M. and Kaplan, J. (2000): 'The fall and rise of Dr. Pangloss: adaptationism and the *Spandrels* paper 20 years later,' *Trends in Ecology and Evolution* 15, 66-70.

Railton, P. (1986): 'Moral Realism,' *Philosophical Review* 95, 163-207.

Scanlon, T. M. (1998): *What We Owe to Each Other*, Cambridge, MA: Harvard University Press.

Shafer-Landau, R. (2003): *Moral Realism: A Defence*, Oxford: Clarendon Press.

Smith, M. (1994): *The Moral Problem*, Oxford: Blackwell Publishers.

Sober, E. (1984): *The Nature of Selection: Evolutionary Theory in Philosophical Focus*, Cambridge, MA: MIT Press.

Sober, E. and Wilson, D. S. (1998): *Unto Others: The Evolution and Psychology of Unselfish Behavior*, Cambridge, MA: Harvard University Press.

Sturgeon, N. (1985): 'Moral Explanations,' in D. Copp and D. Zimmerman (eds.), *Morality, Reason and Truth*, Totowa, NJ: Rowman and Allanheld.

Trivers, R. (1971): 'The Evolution of Reciprocal Altruism,' *Quarterly Review of Biology* 46, 35-57.

de Waal, F. (1996): *Good Natured: The Origins of Right and Wrong in Humans and Other Animals*, Cambridge, MA: Harvard University Press.

Department of Philosophy
New York University
503 Silver Center
100 Washington Square East
New York, NY 10003
E-mail: sharon.street@nyu.edu