# Ethics, Naturally

## On Science and Human Values


# Tore Skålevik


## Master's Thesis in Philosophy


Advisor: Bjørn Torgrim Ramberg Spring 2013

Department of Philosophy, Classics, History of Art and Ideas

# Abstract

In his 2010 book "The Moral Landscape" neuroscientist Sam Harris claims that science can determine human values. This thesis investigates and evaluates Harris' claim in the context of contemporary meta-ethics. I want to show that this kind of scientific moral realism has the resources to face up to the explanatory challenges found in meta-ethical literature, and to explain how something can be said to be morally true without relying on empirically unsupported evaluative premises. Harris suggests that morality is much like health. We can never know how healthy it is possible to become, or if everyone can be as healthy as possible at the same time, but we can distinguish a healthy person from a sick one and use science to identify the causes of health (that which is *good* for you). We humans *will* care about what happens in our world, and while we can never know how satisfied it is possible to become with the state of the world, we can distinguish a satisfied person from an unsatisfied one and—as Harris suggests—knowledge about the causes of well-being (values) "may one day fall within the reach of the maturing sciences on mind". As long our nervous systems share roughly the same anatomy—like our bodies do—we should expect our well-being to systematically depend on the same external factors. It is uncertain whether Harris' theory qualifies as genuine moral realism, and if it is actually able to distinguish right from wrong practices, even given complete descriptive knowledge. The meta-ethical distinction between facts and values, and the tasks moral judgments are supposed to accomplish in order to be called "true" provide some serious explanatory challenges to Harris' theory. I will look at the meta-ethical framework before Harris' theory, so as to be able to present Harris' theory as an answer to the meta-ethical challenges. I first present Harris' theory as a cognitivist theory, and then consider how the cognitive moral judgments it arrives at can be said to be true. As it turns out Harris provides a challenge of his own to the current meta-ethical framework. It seems to reveal a deep disagreement about what qualifies as moral truth. I ultimately think that Harris makes a good case for the plausibility of the maturing sciences of mind developing moral prescriptions and valuable practical moral guidance. I also think that many such prescriptions can be considered true on par with other scientific propositions. If this doesn't qualify as moral truth, then perhaps that demands of meta-ethics are too strict.

# Preface

This thesis is the culmination of five years of study at the University of Oslo. While the thesis itself is focused on Sam Harris' theory of how science can determine human values, the general topic—science and morality—is representative of these five years. My interest in understanding human behavior, social organization and ideology lead me towards studying both the empirical approach of the social sciences and psychology, and the philosophical approach of moral and political philosophy. When engaging in academic life one soon recognizes the enormity of the cumulated knowledge possessed by humans, and sadly realize how little of it one is able to learn in five years. Even so, there is much material that I wish could have made its way into this thesis, but which there was no space to discuss properly. I feel particularly indebted to Daniel Dennett's naturalist and multi-disciplinary approach to philosophical questions, which have had no small role in shaping my own naturalist views, but which is too far removed from the meta-ethical perspective of this thesis to be included. It was also Dennett who lead me to Sam Harris, through their collaboration as critics of religion and other empirically unjustified belief. I have had a great time researching and writing this thesis, and I hope it will prove enjoyable for the reader as well.

**Tore Skålevik** Oslo, May 2013

# Contents

# Introduction: Questions of Value

In the grand scheme of philosophical debate the question "can science can determine human values?" is relatively small. Most people seem to believe that science just isn't the kind of method that can address questions of value, and as such waste little time really considering the issue. This is a mistake according to neuroscientist Sam Harris.

To Harris, questions about how to act, how to live our lives and how to organize our societies are "the most important questions in human life". For naturalists, like Harris, attempting to understand how such questions can be answered in the context of science is a pivotal task. Given that our universe—including brains and minds—is lawful, and that empirical observation provides the basis for knowledge, leaving reason with an instrumental, relational and "model-building" role, we see that if there are any moral facts they must in one way or another rely upon scientific knowledge. Can there be such facts?

In his 2010 book "The Moral Landscape"[1] Harris argues that science can in fact determine human values. He argues that it must be so, given what we currently know about human nature and the nature of the universe in general. Harris' theory will not allow us to immediately answer all normative moral questions. Instead it argues that the questions we can ask today do have objective answers which we can seek, and which we may be able to find at some point in the future, given substantially more scientific knowledge. Harris' theory is almost exclusively meta-ethical. It wants to explain what moral questions and moral judgments are about, and what is required for a moral judgment to be true. It wants to argue that we have knowledge of at least some moral truths today, and that there are more moral truths to be known given more scientific knowledge. In principle—Harris argues—we can know all moral truths given full scientific knowledge. This is how Harris describes his thesis:

> "I want to be very clear about my general thesis: I am not suggesting that science can give us an evolutionary or neurobiological account of what people do in the name of "morality." Nor am I merely saying that science can help us get what we want out of life. These would be quite banal claims to make—unless one happens to doubt the truth of evolution, the mind's dependency on the brain, or the general utility of science. Rather I am

[1] The full title is "The Moral Landscape: How Science Can Determine Human Values". I will use the abbreviation "ML" when referring to this book.

> arguing that science can, in principle, help us understand what we should do and should want—and, therefore, what other people should do and should want in order to live the best lives possible. My claim is that there are right and wrong answers to moral questions, just as there are right and wrong answers to questions of physics, and such answers may one day fall within reach of the maturing sciences of mind." (ML p. 28)

A more modest goal of "The Moral Landscape" is to "begin a conversation about how moral truth can be understood in the context of science" (ML p. 2). It is possible—and I think appropriate—to view Harris' theory as the first contribution, setting the stage for this discussion. Viewed this way the theory appears more tentative and less explicit then its subtitle "How Science Can Determine Human Values" seems to imply. The ambition of "The Moral Landscape" as a philosophical project is not to put forward an exact formula for the scientific determination of human values, but to show that we don't need to invoke *a priori* reasoning or make empirically unsupportable assumptions in order to talk about moral truth.

*

Like Harris, I consider it to be a very important task to seriously consider whether science can help us answer the most important questions in human life, and if so how. My goal in this thesis is to consider whether Harris' theory has the resources to face up to the explanatory challenges found in contemporary meta-ethical literature. Meta-ethics is an extremely complex field of study with a large number of competing theories all seeking to explain morality,[2] and there is no way I will be able to do justice to all of Harris competitors in this thesis. My focus will be on Harris' theory itself, and whether it constitutes a viable naturalist explanation of morality, which can rise up to challenge the various existing explanations. In line with Harris' agenda I want to show that it is *possible* to understand moral truth in the context of science; that it is *possible* that our moral judgments have scientifically determinable truth conditions; and that it therefore is a possibility that science can help us answer the most important questions in human life. By discussing Harris' particular view on *how* science can determine human values, I hope to show that it is plausible enough so that it—and similar theories—can't be dismissed as non-starters, on the ground that they can't *possibly* meet the explanatory challenges in meta-ethics. This is what I mean when I say that I

[2] This is one of the conclusions of Darwall, Gibbard and Railton's 1992 review of meta-ethics.

will consider if Harris' theory constitutes a *viable* naturalist explanation of morality. I don't intend to show that Harris' particular explanation of morality is the *best* currently available explanation. Other kinds of explanations—notably rationalist and non-cognitivist—have been in development for a very long time, and one wouldn't expect a relatively undeveloped explanation, such as the one discussed here, to compare favorably to them right at the outset, at least not before it has been acknowledged as a viable kind of explanation of morality.

*

In order to achieve my goal there are several things I need to do. If I want to show that Harris theory is a viable meta-ethical alternative, I need to show that it has the resources to meet the explanatory challenges found in contemporary meta-ethical literature. To do so, we need to know what these challenges are. Harris is facing roughly the following two tasks: (i) Showing that moral judgments are truth-apt mental states, like beliefs. (ii) Showing that there exist empirically available truth-conditions capable of justifying at least some moral judgments. Harris must show both these things without compromising a fundamental feature of moral judgments, namely that they provide us with a reason for acting in accordance with them. Moral truth would be trivial if at the end of the day believing or disbelieving a moral proposition had no normative effect on our behavior. The challenges from contemporary meta-ethics are particularly well formulated in Michael Smith's "The Moral Problem" (1994), which will be discussed in chapter 2.

I will of course also have to present Harris' actual theory. I start by developing what I take to be Harris' core theory of value in chapters 3 and 4. These chapters attempt to explain value as a natural phenomenon, but don't address, explain or define *moral* value. The account I develop in chapters 3 and 4 is very similar to the account of (natural and non-moral) value defended by naturalist philosopher Peter Railton (Railton 1986a & 1986b), and his thoughts will be very helpful in specifying the position. The theory of natural value forms the foundation for a naturalist solution to the moral problem. It rejects what is known as "value absolutism" as well as "value subjectivism", and argues instead that values are relational. What is "good for" or "valuable to" any given organism or group of similarly constituted organisms will depend on what they are like and how they are affected by events in the world at large. The purpose of the account is to show that value-claims are truth-apt, by showing that they are really claims about physical reality.

Chapter 5 and 6 discuss how our concept of "moral value"—as characterized in "The Moral Problem"—fits into the theory of natural value. Can we accept that claims about *moral* value are relational claims on par with claims about what is *personally* valuable, or do moral values need to be absolutistic in order to achieve the appropriate binding force? On this question Harris parts with Railton and traditional naturalist though. Like most philosophers—including several of Harris' critics—Railton doesn't seem to think that relational value-claims accurately capture what we mean by "moral value" and that to answer moral questions we need to (re)define "moral value" as something like "the maximization of natural value". This amounts to making the unscientific assumption that "the maximization of natural value" is (absolutely) valuable to everyone. Because of the unscientific nature of such an evaluative premise, I think this a move is unavailable to Harris, who wants to show that *science* determines human values. I take Harris to argue that moral value-claims really are a form of relational value-claims.

Chapter 7 mounts some serious criticism against Harris characterization of moral value. The critics I discuss each in some way attribute an evaluative premise—like the assumption that "the maximization of natural value" is (absolutely) valuable to everyone—to Harris, some claiming that he explicitly supports such a premise and some saying that it is a hidden assumption. In all cases they mean to show that without making this, or a similar, assumption Harris' model of the moral landscape fails to reveal any moral facts.

The rest of the thesis (chapters 8 through 11) formulates and discusses the success of Harris response to this critique. Notably Harris thinks that his relational definition of moral value is analogous to our definition of what is healthy. As long as we consider medicine to generate medical truths, we must also allow a science of natural value to generate moral truths. This analogy is very interesting, because if Harris is correct we can use the arguments designed to refute Harris' particular kind of moral realism to refute medical realism. What Harris seems to be arguing is that even if we can't assume an evaluative premise to ground moral truth, we don't have to abandon moral truth, like his critics suggest we must. Harris main strategy for making this point seem plausible is to elaborate the analogy between morality and health.[3] He wants to convince us that a science of morality enjoys the same kind of normative relation to the life of human beings as medicine does. The main question that

[3] The analogy is a prominent feature in "The Moral Landscape", but the extent of Harris' reliance on the analogy is most evident in his extensive "Response to Critics" (2011), where he among other uses, clearly reveals its foundation role, and why it is reasonable to call it his main strategy: "Unless you understand that human health is a domain of genuine truth claims—however difficult "health" may be to define—it is impossible to think clearly about disease. I believe the same can be said about morality. And that is why I wrote a book about it..."

arises out of Harris model is whether it can deal with what we may call "substantial moral questions". While health is a reasonably robust concept, at least intuitively, which provides a great deal of practical guidance, there is even in medicine real disagreement not only about what, in particular, is conducive to health—what is "healthy"—but also on how to conceive of health. Maybe we can agree that blowing up the earth is morally bad, like we can agree that arsenic is unhealthy, because both seems to be obviously true for all human beings. But are these the kind of moral truths we are after? Can a science of morality modeled on medicine even approach the hard questions of ethics? Or does the analogy run out of fuel, and power of conviction, just at the point where ethics gets interesting—and difficult?

"The Moral Landscape" presents us with an unusual approach to ethics, and to properly understand the theory I think that it is important to understand why Harris wrote it and what it is meant to achieve. Placing the theory into its appropriate context before attempting to analyses it in the context of meta-ethics will shed some light on Harris' apparent failure to address some of the issues I will subject it to in this thesis. This is the task I turn to first, in chapter 1.

# Chapter 1: When Scientists Talk About Morality

Everyone comes to meta-ethics with different backgrounds, and with different answers to foundational questions about, for example, the roles and limits of reason and science. This first chapter is about Harris' background, and some of the assumptions and definitions underlying the ideas he brings to meta-ethics.

Sam Harris' PhD is in neuroscience, and his primary field of research is belief. Harris also seems to have an inclination towards philosophy, and before becoming a neuroscientist he acquired a BA in philosophy. However, Harris is probably best known as an advocate of science and empiricism and a critic of faith and religion. In 2004 Harris convinced the major publisher W. W. Norton to pick up and support his first book "The End of Faith", the first book to feature the core view on the immorality of religion and other faith-based belief which has now become known as "New Atheism". The book won the PEN/Martha Albrand award for first nonfiction in 2005,[4] and spent a total of 33 weeks on the New York Times best seller list for paperback nonfiction.[5] This commercial success made easier the publication of similar books including Richard Dawkins' "The God Delusion" (2006), Daniel Dennett's "Breaking the Spell" (2007), Victor Stenger's "The New Atheism" (2009) and the late Christopher Hitchens' "God is not Great" (2009). Sam Harris can be said to have had a foundational role in launching the new atheist movement.

One unsurprising aspect of the New Atheist movement is that they reject God. However some of the new atheists—in particular Sam Harris—have drawn wider ethical and normative conclusions from their general arguments employed against the existence of gods. The new atheist's critique of faith does not limit itself to religious belief but targets all faith-based beliefs in all areas. Further—and more important for this thesis—they claim that acting on the basis of faith-based beliefs is *immoral*.

One central new atheist figure, Victor Stenger, explains that the new atheists argue for "a far less accommodating attitude" towards any kind of irrational or faith-based belief, which is defined as "belief in the absence of empirical evidence, and often in the face of contrary evidence" and that "to act on the basis of faith can often be to act in conflict with reason. We

[4] http://www.pen.org/literature/2005-literary-awards-winners (viewed May 4th 2013)
[5] http://www.nytimes.com/2007/02/18/books/bestseller/0218bestpapernonfiction.html (viewed May 4th 2013)

New Atheists claim that to do so is immoral, and dangerous to society" (Stenger 2010). While not necessarily accurate for all new atheists is sums up Harris' position very well.

In "The End of Faith" Harris argues for the common atheist position that religious beliefs are absurd, but the main force of the book is its moral argument: that religious beliefs, even moderate ones, are immoral (Harris 2004). This line of moral argumentation is present, though perhaps less prominent, in the other books mentioned. Dawkins brings up the point that teaching children that hell is a real place—in complete absence of empirical evidence for this claim—is a form of child abuse equivalent to, and perhaps even worse than, physical abuse (Dawkins 2006 p. 356). As far as one can justify the moral wrongness of child abuse then, the wrongness of this particular religious belief follows. On Harris account, "beliefs are principles of actions: whatever they may be at the level of the brain, they are processes by which our understanding (and *mis*understanding) is represented and made available to guide our behavior (Harris 2004 p. 52).[6]

A brief analysis of the moral argument from "The End of Faith" reveals that it is a consequentialist argument, condemning actions following from unjustified religious action-guiding beliefs for the harmful consequences they cause. The fact that the New Atheist movement started shortly after September 11th 2001 is no accident. While suffering in and of itself is only a direct consequence of some extreme religious beliefs, the more general conclusion of "The End of Faith" is that all beliefs which are out of accord with empirical reality in the very least threaten to produce harmful consequences by causing believing agents to act against empirical evidence, and by preventing them from seeking the truth. This is particularly referring to actions like suicide bombings which are caused, on Harris account, by beliefs in a variety of positive personal and social consequences following from such actions. In Dawkins' case the belief in hell causes—through the act of teaching—great psychological trauma to the children exposed to this belief.

Before Harris engaged the issue in "The Moral Landscape" none of the New Atheists had attempted to produce a philosophical foundation to back up their consequentialist moral views, or refute criticism aimed at this particular position.[7] Their alleged inability to produce

[6] This characterization of belief is similar both in content and terminology to ideas from the pragmatist tradition dating back to C. S. Peirce (Peirce described beliefs as "rules for action") (see for example Hookway 2008). Harris abandons this terminology in ML, perhaps to distance himself from the pragmatist position which he rejects as a whole (despite agreement on several points) because of its relativistic implications for morality (Harris 2004 p.179).

[7] Daniel Dennett has written much on a naturalistic understanding of values. However Dennett doesn't seem to share Harris moral realism, and seems to reject the idea that we in practice can arrive at moral truth by considering the consequences of actions. Dennett (1995 p.494-499) discusses the "Three Mile Island Effect". The meltdown at the nuclear plant at Three Mile Island had tragic consequences, but it also had positive

a sophisticated secular and scientific alternative to the religious moral foundations they set out to condemn and destroy was—and remains—a general source of criticism. According to Harris, this criticism comes from believers and non-believers alike:

> "People who draw their worldview from religion generally believe that moral truth exists, but only because God has woven it into the very fabric of reality; while those who lack such faith tend to think that notions of "good" and "evil" must be the products of evolutionary pressure and cultural invention. [...] My purpose is to persuade you that both sides in this debate are wrong." (ML p. 2)

The "sides" that Harris refers to are sides in the ongoing "culture wars" being waged "both in the United States, between secular liberals and Christian conservatives, and in Europe, between largely irreligious societies and their growing Muslim populations." (ML p. 4).

Harris seems to be claiming that there currently is no *generally acknowledged* source of moral value and moral truth for secularists to ground their moral beliefs in. I think the observed failure to agree on moral foundations, both in philosophy and elsewhere, make this a reasonable assessment. His further claim—that "those who lack [...] faith tend to think notions of "good" and "evil" must be the products of evolutionary pressure and cultural invention"—amounts to the claim that secularists commonly believe that morality is not *real*, making them moral relativists or moral nihilists. This claim is admittedly based on feedback that Harris has received on his previous books and on talks he has given on morality. Exactly how dominant moral relativism and moral nihilism is in the scientific and secular communities is not relevant to my thesis, and I will not attempt to evaluate it. However, only from a realist position can Harris defend the kind of moral condemnation he wields against religion, and we note that one of Harris' primary goals in "The Moral Landscape" is the rejection of moral anti-realism.

This goal may explain much of the philosophical interest in Harris' project. Within academic philosophy the debate between moral realism and moral anti-realism is prominent. According to a recent survey, 56.4% of philosophers accept or lean towards moral realism, while 27.7% accept or lean towards anti-realism, and 15.9% answered 'other'.[8] This debate is

---

consequences on nuclear safety. Was it a good thing? The same ambiguity could be attributed to the consequences of the September 11th attacks. Ultimately Harris doesn't end up endorsing this kind of utilitarian calculus.

[8] philpapers.org/surveys . In the same survey 72.8% were atheists, 14.6% theists and 12.6% answered other.

complicated, with several issues being debated as well as a myriad of rival positions on each side of the realist/anti-realist divide. Three of the most prominent contemporary moral philosophers describe the scene as "remarkably rich and diverse" in their influential article on the historical development and current state of meta-ethics (Darwall, Gibbard & Railton 1992), and in chapter 2 we will see why this is the case.

*

I now want to outline the approach to knowledge underlying Harris thesis. As an empiricist Harris thinks that all knowledge traces back to observation. On this view, claims about reality can only be justified, or count as true, when tested against observations of the natural world. This is basically to say that science is the only method for justifying claims about reality, including moral propositions. As a philosophical position, empiricism is not without competitors. There are some good arguments suggesting that some propositions can be justified *a priori*, meaning that they can be justified and count as true based *only* on reason and logic, independent of any empirical evidence. That debate goes beyond the scope of this thesis. My goal, as stated in the introduction, is to consider if science can reveal moral truth. As such it doesn't matter all too much if there are also other ways to justify moral truth, even if Harris denies that there can be. Our empirical approach means, however, that we can't view conceptual problems as preceding our current scientific understanding of reality. This does not mean that conceptual problems aren't real; it means that they don't arise, and certainly can't be solved, prior to experience. To illustrate this point consider our concepts of "something" and "nothing". The conceptual distinction between these two concepts is about as clear as anything can be. It seems as though it must be true independent of any observation of an external world that "something" is different than "nothing". Based on this it becomes a conceptual problem to explain how the universe (something) can come from nothing. No matter how much "something" we observe by scientific means, it seems impossible to explain how this something could have come from nothing. Basically this *a priori* "knowledge" that we have of "something" and "nothing" is preventing us from taking anything science tells us as answers to questions about the origin of the universe.

In their book "The grand Design" (2010) physicists Stephen Hawking and Leonard Mlodinow makes the claim that science is now capable of answering all questions which traditionally has belonged to philosophy—including "why is there something rather than nothing?" as well as "why do we exist?" and "why this particular set of laws and not some

other?"—and that in this sense "philosophy is dead" (Ibid p. 5). What they seem to mean by "philosophy" here is precisely such *a priori* reasoning. Their claim is not that such questions are answerable by people in white lab-coats doing experiments, but that questions such as these can't be answered outside the context of modern science. The model that modern science, particularly physics, has created of reality is so strange and so counter-intuitive that without it we are bound to get our concepts of "something", "nothing", "space", "time", "past", "future", "causation", "motion", "infinity", "complexity" and so on, wrong by simply considering them independent of empirical reality. The conceptual problems that philosophers deal with are therefore often confused and only follow from "naive" and incomplete models of reality (Ibid p. 7). One of the things Hawking and Mlodinow argue is that science has shown us that "something" and "nothing" isn't really that different at all. There is in a sense no such thing as "nothing" as we typically understand it. An even better, and more recent, formulation of these arguments in non-mathematical terms is found in Lawrence M. Krauss' "A Universe from Nothing" (2012).

Harris too distinguishes between a broad and a narrow definition of science. In the narrow sense science is limited to careful observations, measurements and experiments. Harris thinks that science in this narrow sense "should be considered a specialized branch of a larger effort to form true beliefs about events in our world." (ML p. 195 endnote 2). This "larger effort" is the broad definition of science, and it depends on the use of reason to produce theoretical models of the world—including concepts such as "right" and "wrong"—which agree with and explain relevant (narrow) observations, and as such it is a joint effort by scientists and philosophers alike. Harris seems to think about the human nervous system, like Hawking and Mlodinow think about physical reality. Harris seems to think that an updated scientific model of the human nervous system and how it is affected by the world shows us that many of our common sense concepts about mental states and value—in particular our conception of facts and values as distinct existences—are wrong. The main reason that Harris gives for not engaging most of the meta-ethical literature is that he didn't develop his theory based on this literature, but "came to [his position on the relationship between human values and the rest of human knowledge] by considering the logical implications of our making continued progress in the sciences of mind." (ML p. 197 en. 1).

Harris' scientific approach to moral questions provides several challenges when giving a philosophical evaluation of his theory. Harris is not a professional philosopher and his writings are not aimed at academic philosophy. While his arguments are fundamentally philosophical they have been rhetorically adjusted to convince other scientists and other

interested parties in the general population. Harris has decided that his target audience would be bored by too much philosophical complexity and jargon. While he might be correct, his choice has the unfortunate effect of leaving his position on some important philosophical issues unclear and ambiguous. It will therefore be necessary to present Harris' ideas in a more complete philosophical guise then he does himself, which will involve some interpretation on my part. This gives me the opportunity to develop, or at least clarify, Harris' ideas, so hopefully some good will come of it. Before jumping to his actual theory, I will in the next chapter look at some of these conceptual problems which Harris largely avoids addressing directly.

# Chapter 2: The Moral Problem

In this chapter I will attempt to make sense of the main challenges facing those who wish to engage in contemporary meta-ethics, whether they come from philosophy or neuroscience. These are ultimately the challenges Harris will have to overcome if his moral theory is to be successful. I briefly mentioned these challenges in the introduction. What I didn't mention in the introduction is that it is considered close to impossible to provide a coherent explanation of all the aspects of morality. For example, if Harris can explain how moral judgments are truth-apt—like we will see him do in chapters 3 and 4—then he apparently can't also hold that moral judgments provides the necessary motivating reason for acting in accordance with them. The argument I will present in this chapters quite accurately predicts the kind of criticism we will see applied to Harris' theory in chapter 7.

Why do we think that it is necessary for a proper *moral* judgment be both truth-apt (like a belief) and motivating (like a desire), and why is it so difficult to explain how a judgment can have both these attributes? My answer to these questions is based on what seems to me the best formulation of the problem to date, Michael Smith's "The Moral Problem" (1994). This book was published two years after the already mentioned essay by Darwall, Gibbard & Railton, which describes metaethics as "remarkably rich and diverse". Smith attempts to explain why this is the case:

> In my view, the reason can be traced to two of the more distinctive features of morality, features that are manifest in ordinary moral practice as it is engaged in by ordinary folk. The philosopher's task is to make sense of a practice having these features. Surprisingly, however, these features pull against each other, so threatening to make the very idea of morality altogether incoherent" (Smith 1994 p. 4-5)

So, what exactly are these two features, and how do they "pull against each other"? The first feature is objectivity. Morality, by appearance, is an objective enterprise. When we make moral judgments—for example about actions—we are making objective claims. Or rather, we can say that we are assigning a truth-value to objective moral propositions. If we judge the proposition that "X is morally wrong" to be true, we seem to not simply make a claim about what we like, but a claim about an independent world of moral facts. It is true that we

sometimes also use objective-sounding language to describe what we like. We can say things like "chocolate is good". How is this claim different from the (moral) claim that "helping those in need is good"? The distinction between likes and moral judgments is roughly that it is impossible to convince others of the objective truth of a like. Whether or not one likes chocolate is a matter of taste. There is no rational argument to be made that would change a persons like of chocolate. It is still an unresolved question whether any claims employing the terms "good" or "bad" is really open to change by rational argument, or if they are all matters of taste. However, in ordinary moral practice we observe disagreement and the use of rational argument to convince each other of the truth-value of various moral propositions. In doing so we don't—indeed can't—ground the objectivity of our claim in our own subjective feelings, but rather we refer to what we might call *moral facts* taken to be independent of any particular person. It is precisely the existence of moral arguments and the insistence on the objective truth of one's own believes that allows us to distinguish moral questions from questions of taste. Smith summarizes this objective feature of morality as follows:

> "We seem to think that moral questions have correct answers; that the correct answers are made correct by objective moral facts; that moral facts are wholly determined by circumstances; and that, by engaging in moral conversation and argument, we can discover what these moral facts determined by circumstances are." (Smith 1994 p. 6)

*

The second feature of morality is about moral motivation. Again this is a feature we observe: Having made a moral judgment, fully believing this to be objectively true, the judge finds himself motivated to act on the judgment. To be clear, such motivation doesn't have to be overriding, because the judge can easily have stronger motivation to do something else. We can say that having made a moral judgment, the judge finds himself with a *motivating reason* to act in accordance with the judgment, and that he will act on the judgment in the absence of conflicting motivating reasons. Judging terrorism to be wrong, for example, somehow entails being motivated to refrain from terrorism. This is not some arbitrary definition of moral judgments, which could be otherwise; this is what distinguishes a *moral* judgment from a factual judgment of empirical reality. Smith very appropriately says that discussing whether or not X is morally right seems equivalent to discussing whether or not one has a reason to X.

So, by coming to believe that X is morally right, one simultaneously accepts that one has a reason to X, and thereby being motivated to X, at least in the absence of some other overriding motivation (Smith 1994 p. 6-7). Without this feature morality would be quite uninteresting, as it would fail to have practical implications. Consider by comparison a regular belief about the world, for example "diamonds are hard". In stark contrast to moral judgments, this belief by itself entails no reasons for action. It isn't normative. If the judgment that "terrorism is wrong" did not entail a reason to refrain from terrorism, it would be unproblematic to say things like "sure, I believe terrorism to be wrong, but I would really like to scare some people by blowing up a building, and I don't see any reason why I shouldn't", and the entire purpose of engaging in moral discourse would be lost. Smith summarizes the second feature as follows:

> "Moral judgments seem to be, or imply, opinions about the reasons we have for behaving in certain way, and, other things being equal, having such opinions is a matter of finding ourselves with a corresponding reason to act." (Smith 1994 p. 7)

*

Accepting the first feature—the objectivity of moral judgments—has two important implications. The first implication is the psychological implication that moral judgments constitute a kind of mental state which can be objectively true or false. This means that they have propositional content about reality that can be either true or false. The philosophical position encapsulating this core claim is called *cognitivism*. The second implication is the metaphysical implication that there exists a realm of moral facts, capable of justifying at least some of our moral propositions (more or less the same way that the proposition "snow is white" is justified by the fact that snow *is* white).

Explaining these two implications corresponds to the two tasks I listed in the introduction, and as I said, Harris must do this without compromising the motivational feature of moral judgments, in order to establish moral realism.

Accepting the second feature of morality—the motivational feature—has implications of its own. Believing something to be morally right necessarily provides the agent with a reason to act accordingly. This implication is—as we have seen—absent in the case of ordinary factual believes. This implication must be a part of both the explanation of

cognitivism and of moral facts. In explaining cognitivism the challenge is to show how reasoning can change the agent's desires in the cases of moral judgment, when it seemingly can't in cases of similarly expressed judgments of taste. What I take to be the more difficult challenge is to explain how there can be moral facts capable of achieving these changes in motivation. The success of the two tasks ahead depends to a large extent on how plausible their incorporation of the motivational implication is.

*

Why do we think it is problematic to give a coherent account of moral judgment explaining both these features? A moral judgment—as just described—in part aims to represent a true aspects of (moral) reality and in part expresses the agent's desire for the object of the judgment. This makes moral judgments look like a composite mental state, part belief and part desire. An initial reason to suspect that a moral judgment can't both represent the world as it is and motivate us to promote the object of the judgment is that beliefs and desires seem to have different *directions of fit* (Smith 1994 p. 116). This is a metaphor, but a useful one. Beliefs on this view are mental states characterized by their aim to represent, or *fit*, the world. Desires on this view are mental states characterized by their aim to conform, or *fit*, the world to their content. The content of a desire describe the future state that we wish to achieve, regardless of what we perceive the current state of the world to be. The content of a belief describes the current state of the world as we perceive it, regardless of whether or not we wish for this state. Consider that a moral judgment needs to describe the current state of the (moral) world as we perceive it. There seems to be nothing preventing different people who make such a judgment from having different and contradictory desires. Like holding the proposition "diamonds are hard" to be true is compatible with desires both to change this state and for it to remain, holding the proposition "caring is morally good" to be true is compatible with desiring to care and desiring not to care. The point is that if moral judgments are like beliefs in this regard their truth value is determined by matters of fact over which the individual has no control. As such it seems one doesn't have to be motivationally affected by the moral judgment, because one just *observe* that it is true. If believing that X is good doesn't entail a desire to X, the only way to secure the required motivating reason seems to be to treat moral judgments as expressions of desire, and desires aren't truth-apt, because they make no claims about the world.

The background theory of mental states that informs the problem we have just seen, is commonly called "the humean theory of psychology", because its roots go back to David Hume (Hume 1739-1740). The central claim is that beliefs and desires are distinct existences and that it would be impossible for a single mental state to perform the functions of both a belief and a desire. Hume though that both a desire and a belief are required for action: a desire to secure the motivation, and a belief about how to change the world to fit the content of the desire. Successful actions change the world from what it is to what we desire. On the humean account, desires control actions while reason and beliefs only have an instrumental role. Even if beliefs and desires always occur together on Hume's account, the desire is always necessarily prior to the belief, which leads Hume to say that moral judgments aren't truth-apt.

Hume's argument and Smith's metaphor are very abstract, so perhaps it is better to explain how this distinction affects moral theories in practice. An appropriate example is the normative moral theory known as utilitarianism which originated with Jeremy Bentham (1780) and John Stewart Mill (1863). Like Harris, utilitarians are consequentialist who hold the moral status of actions to depend (in some way) on the consequences on the action. Utilitarians define moral value as the maximization of total well-being, and claim that an action is the morally right action if it is the action that causes the most total well-being as a consequence. This simple premise allows utilitarians to identify the morally right action in any situation is so far as they are able to accurately calculate the consequences of the available actions. Because such calculations are factual and scientific, there will in each case be an objective answer to the question "what is the right thing to do?".[9] These results could be shown to anyone who would have as little reason to doubt them as they would to doubt that life on earth is carbon based. Based on what has been said, we can see that the belief that "X causes the most total well-being" is not sufficient to motivate those who come to (correctly) believe it to act so as to produce X. We can for example easily imagine that the action which maximizes total well-being will harm at least some people. A simple calculation reveals that an action which subtracts fifty percent from the well-being of half the population and doubles the well-being of the other half will increase the total well-being by 25 percent. Half the population is in this case harmed and it strikes us as strange if they became motivated to act in this way simply by recognizing the truth that the action maximizes total well-being.

[9] There are several substantial problems with performing such calculations in practice, as the aforementioned "Three Mile Island effect" is supposed to show.

Peter Singer, who is perhaps the most influential contemporary proponent of consequentialism, lends his full support to the humean distinction between beliefs and desires (Singer 1981 p. 72-86), which he describes as the "best known tenet of modern moral philosophy: the doctrine that there is an unbridgeable gulf between facts and values, between descriptions of what *is* and prescriptions of what *ought* to be." (ibid p. 73)

Utilitarianism is a cognitivist theory, since it claims that moral judgments are beliefs about matters of fact. Another cognitivist theory which suffers the same problem is John Rawls' theory of "justice as fairness" (Rawls 1971/1999). Rawls argues that the right action is the one that secures a just society, according to his principles of justice. We see that it is entirely possible to recognize that X has this effect while at the same time desiring to live in a society with inequality and competition. Because of the possibility of accepting the moral judgment but not desiring to act on it, cognitivists typically reject that moral judgments themselves are motivating. This rejection follows from accepting both cognitivism and the humean theory of motivation. If moral judgments are beliefs, which are distinct existences from desires, and incapable of being motivating, then the required motivational aspect of the moral judgment must be secured by something other than the moral judgment itself. These different cognitivist theories must of course somehow explain why it is that everyone has a motivating reason to act in accordance with their particular answer to moral questions.

One alternative is to reject cognitivism and accept that moral judgments are desires. By doing so non-cognitivists have no problem explaining why people are motivated to act on the moral judgments they make, but they have to concede that moral judgments can't be true in the traditional sense. Yet another alternative is to reject any deep distinction between beliefs and desires, and claim that moral judgments constitute a special "composite" class of mental states, which are both representational and motivational. By doing so one would have to convincingly explain the link between accepting a factual claim and then necessarily be motivated to act in a certain way.

The point of this chapter has not been to argue that it is impossible to produce a coherent explanation of morality, but that it is very difficult, given the conceptual problems described. The fundamental distinction between beliefs and desires seem very plausible, as do both the objective and the motivational feature of morality, as they each seem to be the best available explanation of real and observable phenomenon. Smith's analysis of the disagreement in metaethics concludes that since each of the major positions in meta-ethics reject at least one of the two features of morality, or the belief/desire distinction, they are all

"bound to end up denying something that seems more certain then the theories they themselves go on to offer. Moral nihilism quite rightly looms." (Smith 1994 p. 13).

With this in mind, I now turn to Harris' theory. I will return to the moral problem later, to see how that problem appears in the context of Harris' approach. If it turns out that Harris can provide an attractive response to dilemma Smith describes, than I think we ought to count this as a significant point in his favor.

# Chapter 3: Natural Value

This chapter and the next consider a naturalist version of cognitivism, and aims to establish that judgments of value (evaluations) are representational and truth-apt mental states, upon which reasons for actions are contingent. It does so by way of a naturalist theory of value which is the focus of this chapter. The purpose of this theory is to define the concept of value; what it is, where it comes from and how we come to know it. In addition to providing a naturalistic definition of value, it is also a semantic account, since it involves claims about what fixes the reference of our evaluative terms like "good" and "bad". The account attempts to show that by performing evaluations—including moral judgments—we are in effect making claims about reality, thus indicating that evaluations have propositional content, and further that abandoning one's belief in the propositional content will cause one to abandon the evaluation as a whole. This account will also attempt to explain how some evaluations—judgments of personal value—can be considered objectively true. The question of whether or not *moral judgments* can be true is saved for later, because the answer to this question will depend on what we mean by "moral".

The account of natural value is developed form a perspective of naturalism. Naturalism is not a clearly defined philosophical position. It involves a rejection of the supernatural—vaguely definable as causes we in principle can't have empirical knowledge about—but this rejection is compatible variety of ontological and epistemological positions (see for example Papineau 2007). Harris' version of the theory of natural value builds on a view of the universe as deterministic and a causally closed system, which is especially evident in Harris' treatment of the question of free will (ML p. 102-6, Harris 2012). The only required ontological premise is the possibility of the reduction of mental states, like desires, beliefs and evaluations to states of the brain, and further to causal physical processes.[10] The account of natural value could be rejected if our desires bore no relation to our physical brains and bodies and were constituted outside the causal structure of the universe and as such were principally outside our empirical reach. I don't consider this a very plausible model of reality, and as such not a serious objection.

[10] I am not sure if the possibility of full reduction is actually required for the theory to work. It might be sufficient to accept that the mental supervenes on the physical, in such a way as to allow us to establish reliable correlations between the physical structure of nervous systems and mental content. Harris clearly supports full reduction, but doesn't discuss the technicalities involved.

*

The account itself starts from a premise which Harris takes to be self-evident, that "consciousness is the only intelligible domain of value" (ML p. 32). This formulation is somewhat vague. What we get out of this formulation is that the existence of value depends on the existence of consciousness, but it doesn't specify how. The point that Harris wishes to make is that only to conscious organisms can something matter. The important distinction is between material structures to which nothing can be said to matter (such as rocks) and material structures capable of caring (such as humans). This distinction is at the level of capability; some structures have the capacity to care, and some don't. Harris undoubtedly thinks of caring as a conscious capacity. Organisms which simply react to their environment but lack conscious experience don't *really* care on Harris account. The capacity to have positive and negative experiences caused by internal or external factors seems to be what Harris is after. This capacity is the capacity to value. From now on I will call a conscious organism with the capacity to value "agent".

Inherent in the concept of valuing is a dichotomy consisting of positive and negative value, "good" and "bad". As soon as an agent values X, necessarily the negation of X (~X) becomes something to be avoided. Without speculating in what an agent values in particular, Harris asserts that everything which is really valuable to the agent is constitutive of its "well-being". This suggests that an agent's ultimate goal is to secure its well-being. In fact it could not be different. It would be impossible to an agent to value its own suffering; suffering being the negation of well-being. What is "good for" an agent is thus to be in a state of well-being, and what is "bad for" an agent is to be in a state of suffering, independent of what causes these states.[11]

Harris' definition of well-being and suffering is distinct from traditional concepts of well-being and suffering holding them to be natural properties like pleasure and plain, like is the case with the aforementioned utilitarianism of Bentham (1789) and Mill (1863) as well as most later versions of the theory. It is a common critique of such theories that it is possible for

[11] We should note that this position is similar but not identical to the position known as psychological egoism. Psychological egoism is the position that humans are always motivated by self-interest. Harris' position addresses the physical—or structural—level of the organism, while psychological egoism addresses the psychological level. It is in principle possible on Harris account that a human organism is so constituted that helping another organism in need is genuinely conducive of the well-being of the helping organism. At the level of psychology this organism would experience a genuinely altruistic motivation. Harris' position is more similar to the idea of "the selfish gene" (Dawkins 1976/2006). Genes must be "selfish", or they wouldn't get selected. The various organisms that genes are responsible for, however, can be programmed to be truly altruistic.

an agent to value pain and to recent pleasure. Such a critique does not work against Harris. In case an agent values pain, the experience of pain is included in Harris' concept of well-being for that agent.

The "open question argument" (Moore 1903) is one formal version of this kind of critique. It claims that no natural property (such as pleasure) can be equated with the property of "goodness", because it will always be meaningful to ask whether any particular instance of pleasure is good, and so "pleasure" is not analytically equivalent to "goodness". Harris can similarly say that it is meaningful to ask whether any particular instantiation of a natural property or phenomenon is conducive of well-being, without questioning the goodness (or "well-beingness") of well-being itself.

Well-being is nevertheless a natural property. Given the naturalistic premise that mental states are identifiable by their physical underpinnings, an organism in a state of well-being will look different to science than an organism in a state of suffering, at the physical level. It can be hard to pinpoint exactly what Harris means by well-being, except that the term refers to the physical states of agents that the agents experience as positive and "values". Philosopher, and one of Harris' critics, Russell Blackford contributes the following observation:

> "When the drift of the argument presses [Harris] towards defining well-being, he says that he is not talking about feelings of pleasure; instead, he tends to invoke ideas of deep satisfaction or fulfillment." (Blackford 2010)

Harris' consequentialism gives us a good idea of what he means by "well-being". Positive and negative experiences—roughly characterizable as "well-being" and "suffering"—are lawfully caused by events in the world on Harris naturalistic account. This makes them consequences. "Feelings of pleasure" are consequences of events in the world, and so are feelings of "satisfaction or fulfillment". I think that Harris wants to argue that agents are finite beings, composed of finite material parts, which can be arranged in a finite number of configurations. Some of these configurations involve the experience of well-being and some involve the experience of suffering. When external events affect the organism, it doesn't result in a single change of state, but can be seen as starting chain reactions lasting until the organism dies. Taking a drug can cause intense pleasure at first, but the effects of this pleasure on the physical structure might cause a lasting state of suffering. As such there is *in principle* a metric for measuring well-being in individual organisms. Given sufficient empirical

knowledge about the structure of the organism and how it is affected by external causes, we could predict how much total well-being will result from a given cause.

I think Harris' characterization of "well-being" is very similar to the concept of homeostasis. The structure of a human organism only allows for a limited number of possible configurations of this structure. Rearrange the material parts outside these configurations and the structures breaks down, and with it the possibility of value. Human bodies typically needs to hold a certain temperature, have a beating heart, and so on. As fellow neuroscientist Antonio Damasio (2010) points out, an experience of well-being seems to correlate with the structure of the organism being in a homeostatic state. According to Damasio, the nervous system of our ancestors at some point evolved a monitoring function of the structural state of the body, allowing the body to take action when the necessary internal conditions for continued existence were threatened. As these organisms evolved consciousness, the internal states of the body became represented by experiences of well-being and suffering, depending on how close the body was to optimal homeostatic conditions. This explains Harris talk of well-being as "deep satisfaction or fulfillment" as opposed to mere pleasure, which is sometimes conducive to homeostasis and well-being, and sometimes not. The fact *that* homeostasis is experienced as well-being is contingent, as we can imagine organisms with entirely automated regulatory functions, whose experience of well-being is entirely determined by other factors. Even so, the possibility of experiencing well-being would depend on maintaining homeostasis. The point is that the experiencing of well-being—which is valuable on Harris' account—depends on the agent being in one of a limited number of possible physical configurations, and it is as such these physicals states which are valuable.

We can say all this simply by observing agents, without making any claim as to what in particular causes well-being or suffering in the agents. Peter Railton shares this view, and explains it like this:

> "It seems to me that notions like good and bad have a place in the scheme of things only in virtue of facts about what matters, or could matter, to beings for whom it is possible that something matter. Good and bad would have no place within a universe consisting only of stone, for nothing could matter to stones. Introduce some people, and you will have introduced the possibility of value as well. It *will* matter to people how things go in their rock-strewn world." (Railton 1986b p. 47).

The above definition of value, in conjunction with the premise that mental states are reducible to (in principle) observable physical processes allows us to show three things, which I will now outline. We can reject value absolutism; we can reject value subjectivism and relativism; and we can establish a version of value relationalism. This value relationalism will be foundational for the later account of moral truth.

*

"Value absolutism" is the view that value is determined independent of particular agents, so as to be objectivity valuable, or valuable to all conscious organisms capable of valuing. The above theory of value does not by itself rule out value absolutism, but it demands that *if* there is an absolute value it must be the case that all agents will value it. This line of argumentation is familiar from Kantian ethics. Kant himself argued that all rational agents are ends in themselves, meaning that the value of their existence is absolute (Kant 1785). Kant's methodology does involve truth claims held to be true *a priori*—independent of any empirical knowledge—and does of course stand in stark opposition to empiricism, which underlies the account of natural value. Classical utilitarianism also argues for a form of absolute value, which is compatible with empiricism. Much like the theory of natural value argues, pleasure and the avoidance of pain, are taken to be the ultimate values *for* agents. However, utilitarians—as we have seen—typically go on to claim that maximizing the total amount of pleasure is an absolute value, and the *moral* (as opposed to personal) worth of any action is to be determined by this principle, known as the *principle of utility*. The theory of natural value can't justify this move, unless we *knew* as a matter of fact that every single agent cared for this maximization, even at the expense of their own pleasure and fair treatment. We would have to be able to show that the well-being of every agent includes the maximization of total well-being.

Both Railton and Harris recognize the possibility that agents can be differently constituted. On earth alone there are many different species of agents. The only criteria given so far for something to qualify as an agent is that it has the conscious capacity to value. The relevance of cognitivism and of moral truth does however depend on the agent having a cognitive capacity; a capacity to form beliefs. To simplify the discussion, "agents" will now refer only to organisms with such a capacity and be short for "moral agents" or "rational

agents". I will assume that the human species is the only species of such agents.[12] Even so, the possibility that agents can be differently constituted remains. Human beings are similar, not identical, and we know that human beings are motivated to pursue different ends. As Railton puts it: "What in particular will matter, or could matter, to […] people will depend upon what they are like." (Railton 1986b/2003 p. 47). Neither can we rule out the possibility of radically differently constituted species of agents living on other planets. As long as we can't prove that all agents are sufficiently identical to value exactly the same thing, we must reject value absolutism and stop looking for absolute values.

*

Value absolutism is often seen as a central part of realist theories. If we don't accept that there is something which really is valuable to everyone it is sometimes difficult to see how moral realism could be true, and how we can avoid lapsing into moral relativism. This two-sided debate is for example represented in "Moral Relativism and Moral Objectivity" (Harman & Thompson 1996). Moral relativists, like Harman, share the rejection of value absolutism with Harris and Railton. They claim instead that the values we observe people to be motivated by are *subjective*. We *choose* them for ourselves, much in the same way that normative moral theorists—such as utilitarians, virtue-theorists or Kantians—can be said to *choose* a foundational value which they think (usually for good reasons) apply to everyone. Since everyone choose their own values, moral relativism can explain the motivational feature of moral judgments. Those who accept utilitarianism value the maximization of total well-being, while other can choose to value other thing, like the virtues kindness and benevolence. We can observe that people have many different basic values, including but not limited to the values suggested by the dominant philosophical theories. Relativists claim that none of these basic values are objectively right or objectively better than any other. However, the moral judgments that people make are said to be either true or false *relative to* these different basic values, or "moral frameworks". For example, the judgment that "X is morally good" is true *relative to* a utilitarian moral framework, if X is conducive to the most total well-being, but false *relative to* a Kantian framework, if X is in conflict with a moral duty. Because our own moral framework is salient to us as judges, and our judgments can be said to be true relative to this salient framework, relativists can explains why we (mistakenly) claim that our judgments

[12] Since the rational/moral capacity is an evolved feature, we expect to find, and do find, at least precursors to human rationality, as well as moral behavior, in other animals.

are objectively true; we are making an honest and natural mistake. Moral relativism is thus a form of value subjectivism.

This is significantly different from what the theory of natural value claims. Upon rejecting value absolutism, the theory of natural value does not resort to a subjectivist explanation of values. Determining if X is valuable to a given agent does not require that the agent first subjectively *choose* his basic value or values. X will be valuable to the agent if the agent is so constituted that X is conducive of his well-being. He is not free to choose whether or not to value X. His "choice" is constrained by the particular way in which X affects him. If X causes him suffering, then he can't value X. Railton explains that while the theory of natural value "denies the existence of absolute good, it may yield an objectively determinate two-place predicate 'X is part of Y's good'" (Railton 1986b/2003 p. 49).

Let me attempt to clarify. A person consists of physical matter in a particular configuration, just like everything else in the universe, and there are facts to be known about how this matter is affected by the various particular objects and events (X) external to it. Well-being might be a phenomenological state, but it is still instantiated in the physical agent, and more importantly it is caused by these external events according to lawful physical processes. An agent can *believe* that well-being is a consequence of the object or action he has a desire for, but he could be *wrong*, even about his most basic values. No one is born with knowledge about what is constitutive of their well-being. We are borne with certain dispositions and instincts which will automatically arouse in us desires to act and respond in certain ways. But we are not borne knowing whether or not acting on these instincts is good for us, in our current or future circumstances. We have to learn such things. Like many other species, humans can learn of such consequences by conditioning, and unlike other species we can also learn through the cognitive acceptance of facts. So, the claim is not that agents making value-claims are consciously aware what propositions they are making. The claim is that: Once we form the belief, or otherwise learn, that acting on a desire causes suffering and not well-being, we lose our motivation to act on it, and *will* cease acting on it. All the facts are out there to be experienced, observed and mapped, although conscious organisms have varying and often very limited abilities to do so.

We can reject relativism and subjectivism, because value-claims are not about what is actually being valued subjectively (what the agent can be said to have choosen), but about whether or not what is being desired is really conducive of well-being. The core claim of the theory of natural value is that the value of a particular object or event (X) is determined by the *relation* between X and a particular conscious organism (A). X is valuable to A if X stands in

a certain relation (so as to cause well-being) to A. The theory is therefore *relational*. What is and isn't valuable to A is therefore a matter of objective facts, that serve as truth conditions for his value-claims. As we saw, this does not mean that the theory is absolutistic, because we can only determine the value of X to a specifically constituted agent or identical agents. Two agents who are differently constituted would not stand in the same relation to X. As long as we can't show it to be necessary that all rational agents stand in the same relation to X, we can't claim that X has absolute value.

This leads us to the question of what exactly claims of *moral* value are on this account. Does a claim of moral value have to be absolutistic? Do we have to show that X is valuable to all possible moral agents for it to qualify as moral value? Ultimately Harris' moral theory is about the reality of moral value, and I will further discuss what he takes this to mean in chapter 5. We will not be surprised to find that it is a claim about the relation between a generalized human kind and the world. Before doing so I will in the next chapter consider if the theory of natural value can indeed establish cognitivism, without compromising the motivational aspect of value judgments, and how it deals with the distinction between beliefs and desires.

# Chapter 4: Facts and Values

The theory of natural value claims that judgments of value are both truth-apt *and* intrinsically motivating. From what we saw of the moral problem, this means that the distinction between (truth-apt) beliefs and (motivating) desirers is being challenged. This is correct, and in this chapter I will address this issue. I will not be talking directly about beliefs and desires, but about facts and values, which for the purposes of this discussion is the same distinction, using different terms. That value relationalism by itself provides a great challenge to the fact/value distinction is a claim shared by both Railton and Harris, but it is Railton who provides the clearer philosophical explanation, and I will specify this claim here. Harris provides two other claims against a fundamental distinction between facts and values, which I will also consider. One is related to the account itself and to the further theory, and one is a new and separate argument derived from Harris own neuroscientific research. But, first let's make sure that the theory of natural value can claim both cognitivism and internalism.

I start by considering if the theory of natural value is indeed cognitivist. We must acknowledge that "the moral problem" addresses *cognitivism about moral judgments*, but the theory of natural value seeks to establish cognitivism about judgments of value in general, in particular judgments about the value of X in relation to a single agent (A). This seems like a problem, because we could in principle accept that judgments of personal value were cognitive, while retaining that judgments about *moral* value simply were expressions or projection of our personal value onto the world. We could accept that when I judge X to be valuable to *me*, this involves me *believing* that X really is constitutive of my well-being. But that claims about moral value involves no further believes. I don't think this problem actually arises from the theory of natural value itself. The theory does not demand that agents consciously recognize the propositions—about the effect of acting on the desire on their well-being—underlying their judgments of value. The point is rather that such propositions are implied, and when the agent comes to believe the proposition is *false*, the motivation changes. We can say, along these lines, that when an agent having judged X to be morally valuable, he comes to believe that X in fact causes the suffering (the negation of well-being) of everyone but himself, his moral conviction will change, and he will stop trying to convince others that X is valuable to them. Logically he can't believe that X is both conducive to their well-being and suffering, as we saw. He could however believe that X causes everyone physical pain, but

that pain is conducive to their well-being, in which case his moral conviction would hold.[13] If moral judgments are about what is valuable to everyone—what really matters to them, what their good or their well-being consists in, what they would be motivated to do—then a belief that something is not conducive to well-being will prevent one from judging it morally good.

The theory of natural value states quite directly that judgments of value are beliefs about whether or not a desire is conducive of well-being. We desire well-being, and every particular thing in reality is either conducive to it or not, and thus valuable or not valuable to us. It is straightforward enough on this account that desiring X does not by itself constitute a reason for pursuing X. It must be combined with the acceptance of the proposition that X is conduce if one's well-being.

We realize that this does not in any significant way violate the standard humean theory of motivation. A desire (for well-being) constitutes our fundamental motivation, and the theory of natural value argues that it could not be otherwise. This desire for well-being is not a contingent desire; it isn't a desire for anything in particular. "Desire" might not even be the appropriate word. To say that caring agents desire well-being is to acknowledge the fact that events in the world are capable of affecting the agent in ways experienced by the agent as well-being or suffering, good or bad. Discovering which events are conducive of well-being, and how to initiate such events, are instrumental tasks to which we apply reason based on empirical observation of how the agents are affected by the events. It is by reason we determine the value of any particular thing (X) for agents, after carefully observing the effects of X on the structure of the agent. We were led to believe by "the moral problem", that since instrumental reason only discovers facts and result in representational beliefs, it couldn't possibly reveal value.

The reason why the theory of natural value works is the rejection of value absolutism. It is Peter Railton who points this out (Railton 1986b/2003 p. 47). There seems to be *no fundamental distinction between facts and values when values are seen as relational*. In we mean for the fact/value distinction to separate facts from values at the ontological level, thus creating a form of fact/value dualism—which is precisely what the humean theory of

[13] If an agent has judged X to be morally valuable, and then comes to believe that X causes *physical pain* to everyone, but also happens to believe that physical pain is conducive of well-being, his moral conviction would not change. This could occur because of a belief in things like the cleansing effects of pain, or a belief that pain in this life causes well-being in the next.

motivation attempts to establish, as we saw Smith argue—we must accept value absolutism *as a premise.* [14]

According to Railton, we arrive at fact/value dualism if we accept reason instrumentalism, internalism (the claim that moral judgments are intrinsically motivating), as well as demand that values be absolute. We can't use reason instrumentally to establish that X is absolutely and intrinsically valuable to all agents, but we can use reason instrumentally to establish that X is intrinsically valuable to a particular agent, because we know that the agent *must* care for something, and we can use reason to determine what that is.

The rejection of fact/value dualism is central to Harris' entire project, and to naturalist ethics in general, because it allows us to treat values as facts (and motivating desires as beliefs) which in principle places them under the domain of scientific observation and instrumental reason, what Harris takes to be the "larger effort to form true beliefs about events in the world" (ML p. 195, en. 1). Rejecting value absolutism might seem like a heavy price to pay for moral realists, however, because we can no longer argue that X is absolutely and objectively valuable and as such exerts a normative force on all agents. Perhaps there is good reason for having value absolutism as a (hidden) premise in the moral problem. But recognizing value absolutism as a dead end allows us to explore other and possibly scientific solutions to our moral problems.

*

Harris specifies three reasons rejecting any deep distinction between facts and values (ML p. 11). The first reason reveals that he shares Railton's rejection of fact/value dualism on the back of treating value as relational: "(1) whatever can be known about maximizing the well-being of conscious creatures—which is, I will argue, the only thing we can reasonably value—must at some point translate into facts about brains and their interaction with the world at large;". This is the essence of the theory of natural value, and I have said enough about that for now. We can note that the reason is ontological. It explains value as a real phenomenon in the physical universe, and rejects the possibility that something can be valuable unless it is an ontological fact that it matters to agents. Consider with Harris that something "cannot affect the experience of any creature (in this life or in any other). Put this

[14] It can still be useful to distinguish facts from values for many purposes—like we distinguish the mind from the brain—even if these turn out to be the same at the most fundamental level.

thing in a box, and what you have in that box is—it would seem, by definition—the least interesting thing in the universe." (ML p. 32).

The second reason is an epistemological reason: "(2) the very idea of "objective" knowledge (i.e., knowledge acquired through honest observation and reasoning) has values built into it, as every effort we make to discuss facts depends upon principles that we must first value (e.g., logical consistency, reliance on evidence, parsimony, etc.);". This is meant to blur the fact/value distinction even further. Much of the discussion of the fact/value distinction, is about the problem of going from the empirical "is" to the moral "ought", that is to logically move from facts to values. What Harris wants to point out is that we can't get to "is" without relying on "ought". This reason is very interesting, because it reminds us of our subjective viewpoint and or limitations as seekers of true knowledge. At first it might even seem like this reason works against its purpose, because it seems to block the necessary acquisition of a motivating reason for action following the acquisition of empirical "facts" about X's conduciveness to an agent's (A) well-being. I have said that such "facts" determine the value of X for A, and that A upon believing such facts has a reason for action. However, in case A does not value logical consistency, reliance on evidence, parsimony, etc., the scientific facts alone might entirely fail to convince him of the scientifically determined value of X. This does not affect the theory of natural value itself, since not valuing science would simply cause the agent to not challenge the propositional content underlying his value in light of scientific evidence: He would still *believe* that X really isn't valuable. In Railton's version of the theory of natural value, he doesn't assign science the determinative role that Harris does and this, I think, causes a difference in their subsequent theories of what would constitute moral value. Railton imagines an agent (A) judges X to be valuable to himself, and asks what would cause him to change his belief that X is in his interest (Railton 1986a, 1986b). As we saw, in case A does not value science, empirical evidence will fail to change his belief. Railton introduces the hypothetical A+, a version of the agent (A) who has full knowledge of the consequences of all A's possible action on his well-being. A+, being a future version of A, will be concerned that A does the right thing, and so A will trust in (or value) the advice of A+. Railton does however demand that A+ be fully rational, and it is hard to imagine how this is not simply to attribute the values of science and reason to A+. Harris attempts to remove the subjective perspective from the picture entirely. He argues that:

> "To really believe [a] proposition is also to believe that you have accepted it for legitimate reasons. It is, therefore, to believe that you are in compliance

> with certain norms—that you are sane, rational, not lying to yourself, not confused, not overly biased, etc." (ML p. 14)

Again, this is an implied belief. Everyone initially seems sane to themselves from their subjective point of view. It is when one comes to believe that one has been insane, biased or confused when forming a belief, that one starts to doubt the belief. So if A believes that X is conducive of his well-being, but comes to believe that the formation of this belief was due to say the confirmation bias, his value-judgment fails to provide him with a motivating reason for action. This is as such a second order belief, underlying both moral and factual beliefs. Knowing all there is to know about air and water for example, one could not believe that air was the denser of the two, unless one had a false belief about one's own rationality. If this is the case it seems we can blame the failure of a scientific fact to motivate on the agent having a false second order belief, insofar as we can establish a common standard and common norms of rationality. The point is that facts in all areas of knowledge, not only in the moral domain, depend on values, and that this according to Harris helps to undermine any deep distinction between facts and values.

I will address one further issue in this chapter, the third reason Harris gives against the fact/value distinction: "(3) beliefs about facts and beliefs about values seem to arise from similar processes at the level of the brain: it appears that we have a common system for judging truth and falsity in both domains.". This is a conclusion from Harris' own research on beliefs (ML ch. 4, Harris, Sheth & Cohen 2008, Harris et.al. 2009). The experiment was done by placing subjects in an fMRI scanner and asking them to evaluate the truth of various propositions from several domains, including both ethics and mathematics. The questions had three types of answers: true, false or indeterminate. It turns out that the same areas of the brain are activated when we judge a mathematical proposition as true as when we judge a moral proposition as true. In fact this goes for propositions from all areas tested, and for false and indeterminate answers as well. It is unclear what exactly these result shows, and I shall therefore not focus too much on them. Harris himself takes them to be a clear argument for accepting cognitivism over non-cognitivism (ML p. 225 en. 35). Non-cognitivism does assert that judgments of value—judgments that motivate—must be (primarily) desires, whereas judgments about matters of fact must be beliefs. This distinction is—as we saw—strictly conceptual, and it is hard to tell what, if any, differences between these judgments would be evident at the level of the brain. However, the striking similarity in processes involved in the generation of both types of judgments could easily be taken as support that judgments of

value include propositional content which is judged to be true. If we accept at face value that moral judgments are motivational as well, as they appear to be, this might at least help shift the burden of proof from those who reject fact/value dualism (since the distinction is not evident in our observations) to those who claim it is real.

All in all I think that the rejection of fact/value dualism is well supported by the material I have presented so far. I also think the account of natural value appears very plausible as an explanation of value judgments, including both their truth-aptness and their ability to motivate. There are of course several competing positions which can also—by the looks of it—explain value judgments. All these accounts do in themselves provide a challenge to the theory of natural value by comparison. I think the theory of natural value could claim an advantage by being an entirely naturalist approach to the problem, and unlike non-cognitivist explanations, it can explain how value-judgments are truth-apt as well as motivational. We have, however, rejected the possibility of absolute values, which means that there is still a long way to go if we want to show that moral judgments can be true while also being normative for all humans, who do—as we know—subscribe to different values, even with access to the same empirical data. I now turn to discuss how our concept of "moral value" fits into the theory of natural value.

# Chapter 5: Moral Value

What kind of theory is the theory of natural value? Is it a normative theory, a descriptive one, or both? It certainly is descriptive. It asserts—based on observation—that human beings are material organisms who are lawfully affected by external causes. It asserts that some external objects and events lawfully cause suffering in human organisms and some external objects and events cause well-being in human organisms. Which objects and events causes suffering and which causes well-being (and which has no effect) is entirely dependent on how the organism is physically constituted because this again determines how it responds to external factors. Suffering refers to the various possible internal states that the human organism will act so as to avoid, and well-being refers to the various possible internal stats that the organism will act so as to produce. Being in a state of suffering produces the unpleasant experience of suffering or misery in the organism and being in a state of well-being produces the pleasant experience of well-being or satisfaction in the organism. There are several ways for organisms to initiate behavior that will avoid suffering and produce well-being. The behavior can be "hard-wired", as is the case when humans entirely non-cognitively removes their hand from hot stoves. We can have dispositions towards specific or less specific behaviors, meaning we experience motivation to say punch people we don't like without considering the consequences on our well-being. We can learn to avoid things like poisonous mushroom, even when they look almost identical to nourishing ones. The point is that once the organism forms the belief that any action is conducive to its well-being or to its suffering, these beliefs are action-guiding in the sense that the organism *will* act on them. The claim that Harris wants to make is that it is impossible for human beings to knowingly act so as to produce their own suffering. These are all descriptive claims, but these descriptive claims help fix the reference of value-claims. We know (at least accept) that A *will* perform the actions that he *believes*[15] are conducive of his well-being. It will therefore make no sense for A to claim that what is valuable to him is anything but his well-being and the objects and events conducive to it.[16] We also know that it is a matter of fact—on the physical level—what objects and actions are

[15] Explicitly or implicitly.

[16] Once again, this does not constitute psychological egoism, because A could believe that purely altruistic behavior is conducive to his well-being, and he could in principle be correct.

conducive of A's well-being. A's claim that X is valuable to him, really is the claim that "X is conducive to my well-being" which is a claim about physical matters of fact.

When asking what is the *morally* right thing to do, it seems we are typically not interested in knowing what is in *our own* best interest to do or what is of personal value to us. What is the difference between the claims "X is valuable to me" and "X is morally valuable"? The theory of natural value doesn't straightforwardly tell us what we mean by moral value-claims. "The Moral Problem" however helps us define moral value. A moral judgment—as we remember—is both about matters of fact and it must provide the judge with a motivating reason. How does this definition fit into the theory of natural value?

Consider that A makes the judgment "X is morally valuable". In order for this judgment to provide A with a motivating reason for action, the theory of natural value tells us that A must believe that X is conducive to his own well-being. We can therefore say that the moral judgment "X is morally valuable" entails the value judgment "X is valuable to me". As we have seen, judging something to be *morally* valuable—as opposed to just valuable *for* some given organism—is to judge that everyone who makes the judgment (which you believe is objectively true) also has a motivating reason to act in accordance with it. By the terms of the theory of natural value this seems to mean that a moral judgment can only be the judgment that "X is conducive to the well-being of everyone". Harris says that "when we believe that something is factually true or morally good, we also believe that another person, similarly placed, should share our belief." (ML p. 14). I take it that this is what Harris thinks we are doing when making moral judgments: generalizing from our own belief that X is conducive of our well-being to the moral belief that X is also conducive to the well-being of similarly situated people. My interpretation of Harris is that he supports the definition of moral judgments as the judgment that "X is conducive to the well-being of all human beings, including myself". "X" in this case can stand for an in principle any object or event in a specified context. For example "drinking water is conducive to the well-being of all humans", "drinking water when one is thirsty is conducive to the well-being of all humans" as well as "drinking water on Fridays is conducive to the well-being of all humans" are all possible moral judgments.

It isn't obvious that Harris wholly supports this relational definition, and this is one of the ambiguities I mentioned in the introduction. Harris seems to argue for two related—but different—conceptions of moral value in "The Moral Landscape". The one that has gotten the most attention is not the relational definition, but a utilitarian definition. Thomas Nagel (2010) in particular attributes such a definition to Harris right from the start, and there seems to be

good reasons for doing so. The theory of natural value seems like a perfect backdrop for utilitarianism, because it argues that science can determine what is valuable to each individual, and as such it would be the task of science to determine what causes the most total well-being.

> "The structure of utilitarianism as a moral view suits it to someone who wishes to emphasize the role of scientific knowledge in settling moral questions, for it depends only on one simple evaluative premise, namely that we should do what will promote the welfare of conscious creatures and should not do what will diminish it." (Nagel 2010)

Utilitarianism requires us to define moral value as the maximization of total well-being. This definition is absolutistic, at it holds one value to be absolutely valuable, namely the value of the most total well-being. Even if we could somehow determine that all existing humans valued the most total well-being—unlikely as this is—the most total well-being would not qualify as binding on unborn humans—according to the theory of natural value—because we can't rule out that they might be differently enough constituted so as to not value the most total well-being. In chapters 6 and 7 I will explore the reasons why critics such as Nagel criticize Harris for relying on a non-scientific premise.

There is however at least one very important reason why we shouldn't automatically attribute this "simple evaluative premise" to Harris the way Nagel does. Accepting an absolutistic principle breaks squarely with Harris scientific agenda. Even if we do in chapter 6 and 7 find that the model of the moral landscape could be in trouble if human beings didn't in fact have a basic concern for things like other people's well-being and for social stability, we should not automatically dismiss the possibility that these concerns are contingent truths following from a model of scientifically determined relational human value. I think the more fundamental definition of moral value found in "The Moral Landscape" is the relational one, and that is what I will be arguing for.

The main problem with the relational definition is its apparent inability to render any moral judgments true, even if they are truth-apt. As I have defined it in this chapter, a claim about moral value is the claim that "X is conducive to the well-being of all humans", possibly also including unborn ones. Since human beings aren't identical, such a judgment seems to involve a unique factual claim about the relation of every single human being to the world, all of which has to be true of the judgment as a whole to be true. I don't think this is the final

definition that Harris would want to support, and I will attempt to develop and nuance this definition throughout the rest of this thesis, in light of the criticisms and responses that can be given.

# Chapter 6: The Moral Landscape

In this chapter I present Harris's model of the moral landscape. The model relies on, or follows from, what I will call 'the argument from the most possible suffering', which is itself a continuation of the logic which established the definition of personal value in a material universe. The argument is designed to shed light on what we can and can't mean by "moral value" in such a universe, and seeks to establish a realist concepts of moral value not unlike the concept of personal value, in line with the definition I outlined in chapter 5. Harris doesn't treat personal value separately from moral value, but discusses both concepts simultaneously, which can sometimes be confusing. My main reason for treating them separately is that the argument seems to work much better for personal value, which has allowed me to establish some of Harris' points about the nature of value and science role in identifying it without addressing the additional difficulties involved in applying the argument to moral value. This chapter introduces the missing pieces of Harris argument about the moral landscape, which includes a realist concept of moral, as well as personal, value.

What is perhaps the most foundational premise in Harris' theory is the premise that all claims about value are really claims about the well-being of conscious creatures. I have explained what this claim amounts to at the level of personal value, and when we step up to questions about moral value *no new sources of value are introduced* and nothing foundational changes.

> "The concept of "well-being" captures all that we can intelligibly value. And "morality"—whatever people's associations with this term happen to be—really relates to the intentions and behaviors that affect the well-being of conscious creatures." (ML p. 32-33)

This is a semantic point; about where our moral terms can and do get their meaning from. However, many moral theories—both philosophical and religious—identify values which appear to have nothing to do with well-being. Perhaps most notable is John Rawls' already mentioned theory of "justice as fairness" (Rawls 1971/1999), claiming that the moral value of fairness runs deeper than a concern for well-being, making it our moral duty to act so as to promote fairness rather than well-being. This is a concern because it suggests that well-being

might not in fact be the only source of value. Harris' use of the term "well-being" suggests that he thinks that a concern for fairness and other actually held moral values are really, or result from, a deeper concern for well-being:

> "At bottom this is purely a semantic point: I am claiming that whatever answer a person gives to the question "Why is religion important?" can be framed in terms of a concern about someone's well-being (whether misplaced or not)." (ML p. 198 en. 9)

And:

> "All other philosophical efforts to describe morality in terms of duty, fairness, justice, or some other principle that is not explicitly tied to the well-being of conscious creatures, draw upon some conception of well-being in the end." (ML p. 33)

What agents care about are the only things in existence which we can reasonably call valuable. Logically this means that all claims about value must somehow relate to what conscious creatures care about. Claiming that X is valuable—personally or morally—if X is the kind of thing that no one cares about, in the sense that it is conducive to no one's well-being, makes no sense at all on this model. We must remember that when this theory claims that agents "care about" something, or that something is conducive to an agent's well-being, it doesn't mean subjective caring—or caring at the psychological level—but what is good for the material organism as a whole. Setting yourself on fire is bad for you, if you are so constituted as to have your well-being reduced by being burned. This is true even if the organism is psychologically motivated to perform such an act, which it could be for various reasons. It could for example (wrongly) believe that being burned brings about a cleansed state which is what its well-being really consist in.

It is important to see that Harris is not arguing for or against the value of fairness, pleasure, justice or even purity or a duty to always tell the truth. These would be normative arguments of the kind we saw him make in "The End of Faith". The model of the moral landscape aims to show how we must go about it if we wish to make such claims. So, in case these particular possible values are in fact valuable they are so *because* conscious creatures are concerned about them as a function of their biological constitution and relation to the world at large. It might very well be the case that the just and fair world is preferable to a

world in which well-being in any narrow sense (like pleasure) is maximized, but this would be *because* the inhabitants of that world were so constituted as to prefer justice to pleasure, at the causal and physical level. The core point is that actual norms, duties or moral imperatives—like Kant's categorical imperative—"only qualify as a rational standard of morality given the assumption that it will be generally beneficial" (ML p. 198 en. 10). Let me explain this more carefully. I think this point is pivotal, and so it is worth the risk of repetition, to ensure that Harris' point is as clear as I can make it. Any imperative, belief or norm that you accept as true guides you actions. If you *really* believe that it is right to "act only according to that maxim whereby you can, at the same time, will that it should become a universal law",[17] then that is how you will act. From what we know of how human beings form beliefs, we know that it is possible to hold action-guiding beliefs of this sort for many reasons; I have mentioned that they could be caused by dispositions, by learning and by confirming to social norms. The fact that you are psychologically motivated to act on a held belief, doesn't tell you anything about how acting on it will affect you as an organism in your current circumstances. If you came to believe—for example by scientific knowledge of the actual consequences—that acting on the belief was conducive to your suffering, the action-guiding belief would no longer qualify as a "rational standard of morality". This is to say that any norm we can believe in—and we can believe in many—will only qualify as a rational standard for morality if it is in fact the case that that norm is conducive to the deeper well-being of those it affects. If a norm was demonstrably conducive of the suffering of human beings this norm could not be adopted as a moral standard *unless* we were afflicted with false believes about the consequences of adhering to this norm.

However, understanding and acknowledging (the fact) *that* moral value must in some way relate to the experienced personal well-being of agents doesn't take us very far by itself. The difficult task is to provide a satisfactory explanation as to *how* exactly moral value relates to the experienced personal well-being of agents. Harris employs the model of the moral landscape to explain moral value.

> "Throughout this book I make reference to a hypothetical space that I call "the moral landscape"—a space of real and potential outcomes whose peaks correspond to the heights of potential well-being and whose valleys represent the deepest possible suffering. Different ways of thinking and

[17] This is a formulation of Kant's "categorical imperative" from "Groundwork for the Metaphysical of Morals" (1785)

> behaving—different cultural practices, ethical codes, modes of government, etc.—will translate into movements across this landscape and, therefore, into different degrees of human flourishing." (ML p. 7)

This hypothetical space is described by Harris as a continuum of possible experiences from the worst possible misery for everyone and various degrees of well-being up to and including the most possible well-being for everyone. The moral landscape results from an argument starting with the premise that the most possible misery for everyone is bad; morally bad. While being personally bad for each agent—who is in a state of the worst possible misery—it is also morally bad in the sense that it will be impossible for anyone to argue that such a state of the universe is desirable and thus valuable.

> "We simply must stand somewhere. I am arguing that, in the moral sphere, it is safe to begin with the premise that it is good to avoid behaving in such a way as to produce the worst possible misery for everyone." (ML p. 40)

The reason why it is safe to begin with this premise is that there is no possibility that any of the involved agents will desire to produce this state. The state is of course hypothetical. Nothing needs to be said about what this state consists in or what actions lead to it. In practice it could be something like simultaneously subjecting all agents to a harmful neurotoxin (ML p. 39). Accepting that the worst possible misery for everyone is morally bad might seem obvious, but it has an important logical implication: All other possible states of the universe are now morally *better*. If one finds oneself in this state of the most possible suffering for everyone would be morally obligated to act so as to abolish it. The most possible misery for everyone corresponds to the lowest depth on the moral landscape. There are many other valleys on the moral landscape as well. These correspond to states of the universe where all agents suffer, though not as much as they could. All these states are better than the lowest depth, but there are other types of states—the peaks—which correspond to all agents experiencing well-being. Any agent finding himself in a valley will be motivated to climb a peak.

Moral goodness and badness is defined within the context of the moral landscape independent of any particular experience of suffering or well-being. This does by no means mean that the model and its definitions are *a priori*. I think the best way to view the model is as a logical implication of the scientific knowledge we currently have about physical reality including consciousness and motivation. Even if we don't know what states of the universe

constitutes the worst possible misery and the most possible well-being we can say that there is a clear observable difference between these two types of state. This difference is more clear-cut in the case of individual conscious agents. In the language of the moral landscape we can say that it is right for the agent to move towards peaks on the landscape, since the peaks involve his own well-being. The landscape itself is set independently of the agent's subjectivity, and he can be right or wrong about how to move across the landscape and whether or not he has reached the highest peak. But we can't get away from the fact about motivation: The agent will be motivated to move away from valleys and he will be motivated to move towards peaks once he recognizes where he is located.

What constitutes "right" on this model is not the achievement of any particular state, but upwards movement on the suffering/well-being continuum. It will always be right for an agent in a particular state to move towards a state of higher well-being, and it will always be wrong to move towards a state of more suffering. I already covered this for individual agents, and I am repeating in here because the same is the case for moral value.[18] What constitutes a state of well-being or suffering for a group of agents is however much less clear-cut. There might be states where some of the agents experience well-being and some suffering, but the definition of right and wrong as movement between states of suffering and well-being is still supposed to hold. As long as the extremes of the continuum remains possible, talk about moving away from the worst possible misery for everyone towards the most possible well-being *for everyone* still makes sense.

> "Once we admit that the extremes of absolute misery and absolute flourishing—whatever these states amount to for each particular being in the end—are different and dependent on facts about the universe, then we have admitted that there are right and wrong answers to questions of morality." (ML p. 40)

We can imagine states where all agents suffer and we can imagine states where everyone experiences well-being. It will thus always be morally right to move towards states where

[18] We might think that once we figured out what was valuable for every single agent, there would be no need to introduce the concept of "moral value". If we accept that personal value is normative for every agent, and it is "right" for them to act on it, we seem to have arrived at a version of moral egoism. Figuring out additionally that a group of agents share some value doesn't seem to change the motivation of any of the members, or how it is "right" for anyone to act. However, moral discourse just is to identify moral values, even if a moral value just is a value shared by all humans. For now we are trying to figure out if there are such moral values.

everyone experiences more well-being then they do in the current state, and morally wrong to move towards states where everyone suffers more.

> "Grounding our values in a continuum of conscious states—one that has the worst possible misery for everyone at its depths and differing degrees of well-being at all other points—seems like the only legitimate context in which to conceive of values and moral norms." (ML p. 41)

The model of the moral landscape doesn't promise to solve all moral disagreement.[19] It argues that the various ways we can conceive of moral value are constrained by the naturalism and cognitivism that are imposed. We can't give a full account of what in particular is valuable before we have full scientific information about the causal structure of the universe and how the various actions available to us affect well-being, but even today there are a few actions we can clearly conceive of as conducive to the well-being or suffering of everyone and therefore morally good or bad.

> "The fact that it might be difficult to decide exactly how to balance individual rights against collective interests, or that there might be a thousand equivalent ways of doing this, does not mean that there aren't objectively terrible ways of doing this." (ML p. 42)

These "terrible ways" are all morally worse than the contenders for what is "best". I have now presented the basic model of the moral landscape, and added some nuance to the relational definition of moral value. Moral judgments are true—on this account—if the action they judge good produces a state of the universe which is further away from a state of the most possible suffering for everyone and closer to a state of the most possible well-being for everyone. There are however some serious problems with this model which must be addressed.

[19] Moral disagreement will be discussed in detail in chapters 9 and 10.

# Chapter 7: Considering a Principle of Maximization

This chapter will begin to answer questions about the model of the moral landscape. Is it actually possible to talk of moral value in the context of the moral landscape—as I have described it—while relying only on scientific knowledge? Answering this question will give us a better insight into what kind of theory the moral landscape amounts to.

The first thing I will do is to reveal what can seem to be a critical flaw in the logic of the moral landscape. The moral landscape is described by Harris as the continuum between the worst possible misery for everyone and the most possible well-being for everyone. Everyone presumably has a motivating reason to move away from the worst possible misery and towards the most possible well-being. Such movement seems inherent in the concept of motivation. The ability to care means that there are some things you want to avoid and some things you want to bring about. Exactly what one is motivated to pursue changes with one's beliefs about how actions affect one's body. Ultimately a caring agent must care to experience well-being—as we have defined it—and adjust his contingent desires accordingly. We can know that any given agent necessarily would be motivated to avoid the most possible suffering for everyone, and he would be motivated to act as to bring about the most possible well-being for everyone. We can know this because both these hypothetical scenarios include the agent himself. In the last chapter I quoted Harris saying that:

> "Once we admit that the extremes of absolute misery and absolute flourishing—whatever these states amount to for each particular being in the end—are different and dependent on facts about the universe, then we have admitted that there are right and wrong answers to questions of morality." (ML p. 40)

This passage actually admits that the existence of moral truth depends upon there being possible states of absolute misery and absolute well-being (or flourishing, by which Harris seems to mean the same) and that there is a distinction between these states. If we substitute the last word "morality" for "personal value", this passage neatly summarizes the theory of natural value. I have argued, with Harris and Railton, that it is in principle possible to determine which states of the universe constitutes the maximum well-being, and the

maximum suffering for a particular agent (A). I have explained that the reason why this works is that A can't contradict himself. He can't for example value something that is conducive to his suffering. Given a that A is the only existing agent there could only be one—either positive or negative—relation between A and all possible objects and events (X) that affect A. As such X would be either valuable or not, in fact X would even qualify as morally valuable by being valuable to all existing agents.

For particular agents the distinction between well-being and suffering is pretty clear-cut. The idea that each agent's well-being and suffering can be maximized seems very plausible. Given that the universe is material and lawful, as it appears to be, it also seems plausible that science in principle can determine what these states consists in for each agent.

Now, introduce another agent (B), and this scheme changes. Questions of *morality* are about what all human agents have motivating reason to pursue. We must therefore ask if, and how, the hypothetical states of absolute well-being and absolute suffering mentioned in the Harris quote can apply to all human agents. Clearly we can't rule out the possibility that there are possible states where both A and B experience suffering (morally bad) and states where both experience well-being (morally good). However, we also can't rule out the possibility that there are no such states. Because A and B are not identical, it is a possibility that X is conducive to A's well-being while at the same time being conducive to B's suffering. It seems like the logical possibility that there is no state which can be called good and no state which can be called bad will remain. To keep things simple, consider that X is involved in A's maximized well-being and in B's maximized suffering. Given this logically possible scenario, there could not be states of "the most possible suffering for everyone" and "the most possible well-being for everyone". There might not even be any states where both A and B experience well-being, if A and B are sufficiently different. It might even be the case that A's suffering is conducive to B's well-being, and vice versa.

The point I wish to make here is that any vague definitions of absolute, maximized or most possible well-being and suffering will not suffice. The model simply fails to establish the existence of moral truth without making more explicit claims as to what is involved in the states we are all supposedly motivated to move away from and towards. One way to do this—as we have seen—is to define the states in terms of the most aggregate well-being and suffering. This would mean setting as a criterion for the morally right action that it is the one which produces the most total well-being all agents considered.

It remains unclear to this day whether or not Harris makes this move. There certainly are passages in "The Moral Landscape" where Harris appears to show full commitment to a

principle of maximization, but then again the general argument in the book—especially Harris' persistent insistence on the foundational role of science—strongly indicates that he is not willing to accept such a principle. The philosopher Troy Jollimore clearly identifies Harris' confusion on this point is his review of the book:

> "At times [Harris] seems to use "consequentialism" simply to imply that the consequences of an action, in terms of conscious creatures' well-being, are what determine that action's moral rightness or wrongness. This is a quite modest view that is compatible with all sorts of accounts of *how* such well-being matters. (For instance, the claim that I should always maximize my own self-interest, and not be concerned with anyone else's well-being, is in this sense a consequentialist view.) But at other times he goes much further, seeming to suggest that he has somehow established that the consequences must matter in a certain way: well-being in the universe at large (and thus not simply my own well-being, or that of myself and those I care about) must be *maximized*—even where doing so involves violating the basic rights of some particular person, or sacrificing the few for the sake of the many." (Jollimore 2010)

Jollimore points to Harris' discussion of Robert Nozick's thought experiment regarding so-called "utility monsters" as the strongest evidence for Harris acceptance of a principle of maximization. Utility monsters are a hypothetical species of moral agents who would receive great happiness from devouring the human species; comparatively much greater than the suffering that humans would receive in being devoured. Obviously, humans would suffer as a result of such an interaction. Given what has been said about consciousness and motivation, it seems impossible that humans be motivated to be devoured given only the true belief that aggregate universal well-being would be increased as a consequence. Harris nevertheless outright supports such an act of self-sacrifice as morally good:

> "Nozick [. . .] asks if it would be ethical for our species to be sacrificed for the unimaginably vast happiness of some superbeings. Provided that we take the time to really imagine the details (which is not easy), I think the answer is clearly "yes." There seems no reason to suppose that we must occupy the highest peak on the moral landscape." (ML p.211 en.50, quoted by Jollimore 2010)

The critic who sees Harris most clearly as a utilitarian is perhaps Thomas Nagel, who spends little time dwelling with the mentioned ambiguity and states quite directly that "Harris is a utilitarian, in the style of Mill", although he admits that it can occasionally "sound as though Harris is denying the distinction between facts and values altogether" (Nagel 2010). This gets right at the point. Being a utilitarian, and accepting the principle of utility, just is to accept an evaluative premise—a value—conceptually prior to any scientific knowledge. It is to define "the most total well-being" as an absolute value, independent of what any particular conscious agents actually value. It is Nagel's claim that Harris is committed to the truth of an irreducible value judgment that specifies which facts determine the difference between good and bad, right and wrong.

The most extensive review of "The Moral Landscape" comes from Russell Blackford, who also—perhaps legitimately—attributes an evaluative premise to Harris:

> "Harris overreaches when he claims that science can *determine* human values. Indeed, it's not clear how much the book really argues such a thing, despite its provocative subtitle. Harris presupposes that we should be motivated by one very important value, namely the well-being of conscious creatures, … ." (Blackford 2010)

The reason why Blackford thinks Harris presumes this value is because he sees it as the only way Harris account can be logically coherent.

> "There might be a determinate, objectively correct answer to what maximizes global well-being, but no such answer to the ancient questions, "How am I to act?" and "How am I to live?" It's *these* questions that really matter, if we're looking for guidance for our actions." (Blackford 2010)

We are starting to see even more clearly why it is insufficient to simply say that morality is about the well-being of conscious creatures. There are several different ways to care about the well-being of conscious creatures. Blackford points out that we can ask "Why, for example, should I not prefer my own well-being, or the well-being of the people I love, to overall, or global, well-being?" (2010). This is basically what Harris must explain. When facing a moral dilemma, where one course of action increases only your personal well-being and another course of action increases global well-being: How is one to act? If we acknowledge that different people have different interests; that is, if we can somehow know that different

actions really are conducive to the well-being of different people, how are we to identify the morally right action in a given situation?

So, what if Harris is a utilitarian, and accepts a principle of maximization as part of his theory? It certainly would satisfy the need for a clearer definition of what is involved in the morally valuable state which we are obligated to direct our actions towards. However, there are many problems with attributing such a principle to Harris. Accepting an evaluative premise of this sort is not compatible with the rejection of fact/value dualism. As Blackmore (2010) goes on to show there is nothing irrational or self-defeating about an agent who doesn't act so as to increase global well-being given that he simply doesn't value global well-being. In other words, he is not doing anything wrong. Accepting a principle of maximization on rational grounds also seems to defeat Harris scientific agenda. Science is not showing us that it is right to prioritize global well-being over other kinds of well-being. Finally, Harris doesn't to any relevant extent address or deal with the well-known problems associated with the acceptance of such a principle. This leads Nagel—who is under the impression that Harris is a utilitarian—to conclude that "Since Harris skips over the hard substantive questions of right and wrong that occupy moral philosophers, the book is too crude to be of interest as a contribution to moral theory." (Nagel 2010).

Accepting the principle of utility and then go on to not discuss the implications seems strange, and it would indeed make Harris theory philosophically uninteresting. For these reasons I don't think that Harris really intends to employ a principle of maximization as a premise in his theory, and that such an interpretation overshadows the real issues. Addressing the issue, as I have done in this chapter, has nevertheless identified and specified the explanatory challenges Harris is facing. The challenge then is to find a different explanation as to *how* the well-being of conscious creatures matter. If it is not as simple as the morally good world being the one with the most aggregate well-being, then what is this state we are motivated to move towards?

I will in a later chapter return to the issue of global well-being, as I do think that this proposed value occupies a central part of Harris moral theory, much like the critics points out. I do however not think that it occupies the role of an absolute and logically prior value, but can rather be considered a 'human value' pertaining to the human kind *qua* our common human nature and our circumstances. As such this value doesn't go into the grounding framework of the moral landscape, but follows from it and contributes to making it a useful version of moral realism.

I now turn to the more direct arguments for how to understand the concept of moral value, which hopefully will bring us closer to what I take to be Harris' final definition of this concept.

# Chapter 8: The Analogy with Health

So far I have attempted to come to terms with Harris' moral theory. I have laid out his core arguments with as much philosophical accuracy as I think can be done with a basis in "The Moral Landscape". My aim—as I stated in the introduction—has been to see if this theory fits into the framework of contemporary meta-ethics and can qualify as a legitimate version of moral realism, like Harris claims it is. But we have run into trouble. I have explained what Harris thinks of value, and arrived at the position that all claims about value are really claims about the well-being of conscious creatures. This is clearly a consequentialist position, but as Harris' critics correctly point out, this by itself leaves us wondering how X must affect the well-being of conscious creatures for X to count as morally good. The answer we have seen Harris give to this question seems insufficient. The answer is roughly that X needs to move all human being towards a state of the most possible well-being for everyone or away from the state of the most possible suffering for everyone. This corresponds to the characterization of a moral judgment as the value judgment "X is conducive of the well-being of everyone". We have established that such a judgment is truth-apt, because the well-being of each material organism is lawfully affected by X, but can such a judgment possibly be true? What does it take for us to be able to say that X is conducive to the well-being of everyone, or moves everyone towards a state of maximized well-being?

When asked how to justify the truth of the kind of moral value-claims he has been arguing for, I think that Harris parts with what he seem to regard as the "philosophical confusion" of traditional meta-ethics. From what we have seen of the critique against Harris, it seems that before we can say that X is conducive to the well-being of everyone we need to specify in concrete terms what we mean by "the well-being of everyone", and this could be for example "fairness", "the most total well-being" or "virtuous character". Since we allegedly can't show by scientific means that any of these particular possible values really are conducive to the well-being of everyone, we must rely on an evaluative premise; we must accept them based on non-scientific justification. The reason why we supposedly can't observe X to be conducive to the well-being of everyone is—as we have seen—that we can never rule out that at least one human or future human will not be so constituted as to value X. Harris has no objections to this logical possibility, and as we remember, this was precisely the

point given as an argument against value absolutism and for the rejection of fact/value dualism in the theory of natural value.

What Harris objects to is the claim that because we can't observe X to be valuable to all humans, we can't hold X to be morally valuable and we can't hold that X provides everyone with a reason for action (making it *wrong* for them to act against it), unless we accept non-scientific assumptions. This claim would be plausible in case values were subjectively determined. If A's values are determined by his subjectivity, independent of the physical constitution of the organism as a whole, we couldn't predict whether or not X will be valuable to A based on empirical observation.[20] But values on Harris account are relational, determined by the objectively available relation between a valuing organism and objects and events in the world. Knowledge of the structure of a given organism will therefore allow us to predict what is valuable to it, and with increased knowledge comes greater accuracy in our predictions.

Consider now that we observe that X causes in the nervous systems of everyone it affects a state which we know—also through observation—correlates perfectly with well-being. That is we know that whenever a person's nervous systems displays this particular configuration the organism is motivated to maintain it, and we know that X systematically causes this state in people. We can predict that X will continue to do so in all future cases, but we can't of course know for certain that this prediction will hold. Consider that X is the act of taking a drug which enhances intelligence. The prediction we are making—based on a large amount of data—is that taking this drug is conducive to the well-being of every human. The objection—that we can't be certain that the prediction will hold—amounts to saying that it is always an open question whether or not X will be good for, or conducive to a well-being of, a given person. We remember Moore's open question argument from chapter 3. The argument is supposed to show that we can never equate the property of "goodness" (or in Harris' case the property of "conduciveness to human well-being") with any natural property, such as enhancing intelligence. While this might in the strictest sense be true, it doesn't matter for moral truth on Harris' account. In our example, it seems that A has a motivating reason to X, just in case he values empirical evidence. What Harris wants to emphasize is that if science tells us that X systematically causes well-being in humans, than they have as good a reason for acting on X that they have to believe (and act in accordance with) the proposition that water is denser than air and other similarly established scientific facts.

[20] I have already conceded that the theory of natural value could be rejected if values are constituted outside the causal structure of the universe and as such were principally outside our empirical reach. The account depends on the supervenience of the mental upon the material, and the possibility of reducing values to their physical underpinnings.

There is one branch of science which Harris considers analogous to what a science of morality would look like, and that is medicine. Moral truth, on Harris' account can be understood by analogy to medical truth, where moral goodness is analogous to good health, and different actions and practices are considered morally good, just as some actions and practices are considered healthy.

Relying on an analogy to prove one's points can sometimes be risky. In Harris' case I think the use of the analogy with health is a good way to demonstrate some points which it might otherwise be very difficult to explain. As Harris admits, the analogy isn't perfect, but he asserts in his response to critics (Harris 2011) that it is good enough to obviate the criticisms commonly wielded against his theory, like the ones I have just discussed. I don't think the analogy is needed to establish any part of Harris' theory as I have laid it out over the last five chapters. The purpose of the analogy is rather to show why the kind of theory we have now arrived at deserves to be categorized as a genuine form of moral realism, despite its problem in specifying a moral "first cause". It is thus meant to refute the claim that moral judgments about what is conducive of the well-being of humans *in general*, based on the relational definition of moral value, constitute a category of propositions which can't *possibly* be true. Medicine is precisely about making true claims about what is healthy for human beings *in general*. It is important to carefully consider if the analogy holds for the kind of moral questions that we are interested in finding the answer to, the sort of questions being debated in normative ethics, like human rights, the death penalty, obligations to future generations and so on. If it turns out that the hard questions of ethics fall outside the reach of a science of morality modeled on medicine, then we will have to consider if the analogy is really useful.

*

What exactly is health? What does it mean to be in a healthy or unhealthy state? By what criteria can we call substances or practices healthy or unhealthy? We know, or at least seem to know, that being healthy has something to do with being physically fit and being free of disease. We also seem to have few reservations when we call practices and substance which are known to induce such states in humans 'healthy' and 'unhealthy'. Are we willing to recognize that there is a genuine distinction between healthy and unhealthy states? Are we further willing to recognize that science is the way to determine the various causes of these

states? Harris answers both these questions with a clear "yes". More importantly he is claiming that doing so is entirely uncontroversial; that these are obvious facts.

But what does it mean to be "physically fit" and "free of disease"? If we spend only a short time dwelling on the meaning of these terms, we find that they are impossible to define with any precision. To be physically fit means something like being able to do what it is normal for people of the same age to do. What counts as physical fitness can't be defined in absolutistic terms. That is, we can't say that being able to walk a kilometer at age forty is healthy for all humans, always has been always will be. This feat can easily enough be achieved by unfit—or less than average fit—people in most parts of the world today. Not long ago being alive at age forty was uncommon, and as Harris points out "there may come a time when not being able to run a marathon at age five hundred will be considered a profound disability." (ML p. 12). A clear difference between the ability to walk a mile and the lack of such ability nevertheless remains strong. At the extremes, "the difference between a healthy person and a dead one is about as clear and consequential a distinction as we ever make in science." (ML p. 12). Despite the problems with giving precise definitions of physically fit, a dead person can never be considered healthy.

The absence of disease is also a criterion for being healthy, but what constitutes a disease? There are dictionary definitions defining disease as "a disordered or incorrectly functioning organ, part, structure, or system of the body…".[21] Of course, the notion that something can function "incorrectly" presumes that there is a correct way for that thing to function in the first place. Any strict definition of correctness will have built into it a kind of normativity that can't be observed in nature. All we can observe are bodies functioning. Often we will observe two bodies functioning differently, but there is nothing inherent in the functioning that tells us which way is "correct". However, as long as being healthy has something to do with being fit and living long and free of pain we can distinguish between bodily functioning which supports and induces these states and bodily states which do the opposite. As long as people are not identical there will probably always be disputes about what is healthy and what constitutes a disease. The important point is that such disputes will be about whether or not X is conducive to health (being fit and living long and free of pain), and not whether being fit and living long and free of pain is healthy. It seems hard to even imagine someone arguing that dying in agony is healthier then living long and enjoying life. Or less extreme, that living with Alzheimer is healthier than not.

[21] http://dictionary.reference.com/browse/disease (Viewed May 4th 2013)

In what way is health like morality? First we notice that the terminological framework is very similar. There is good health and bad health and the external factors responsible are healthy or unhealthy, just like there are desirable and undesirable states and morally good or bad actions leading to them.

The definition of "good health" is what Harris calls "open-ended". We don't know—and probably never will—how "healthy" it is possible for human beings to be. We don't know how long it is possible for humans to live, what physical feats we can be capable of in the future or how much painful and disabling conditions can be reduced. Harris is suggesting that the same is the case with the definition of the "morally good". Health has something to do with being physically fit and free of disease, and morality—on Harris' account—has something to do with being in a state of well-being and free of suffering. We don't know—and probably never will—how much "well-being" it is possible for human beings to experience; we don't know what our limits are. And we don't know how much suffering can be reduced.

Perhaps even more importantly, we don't know if it is possible for all humans to be as healthy as they can be at the same time. Perhaps it will turn out that the presence or level of a substance in the atmosphere will affect people's health differently, making some healthier and some less healthy. We don't, in other words, know what state of the universe we are aiming for when we practice medicine. This is also the case with well-being. We don't have any absolute, or ultimate, purpose to direct our actions towards. We don't know if a state where all humans experience as much well-being as possible is achievable. As a have argued, it is possible that X is involved in A's maximized well-being while being conducive to B's suffering. This leads Harris to believe that we can practice normative ethics without knowing exactly what state of the universe we are aiming to achieve, and that indeed this is the only way to go about it.

What is healthy and unhealthy to an organism will depend on what it is like, like what is valuable to an organism. Medicine is however the study of *human* health, and medical truths are however truths about *human* health. At this point we can start to ask how it is possible to have a working science of general human health, give that we don't know what ultimate state this science is aiming to bring about, and given the logical possibility of humans who are so constituted as to not be bound by general health advice. One of Harris most central

points is that the criticism which is typically applied to moral realism—as in the moral problem—can with seemingly equal force be applied to what Harris calls "medical realism".[22]

Medicine is not about discovering and attributing absolute value to things in the world. It is not about identifying a property of "healthiness" in objects or actions, at least not in any strict logical sense. Medical claims do not rule out that there are or will be humans to whom the conclusion "X is unhealthy" does not apply. All it takes is a small biological difference for a person to be immune to the inhibiting effects of a generally harmful cause. Similarly, a small biological difference can be responsible for a person being harmed by a generally harmless cause. As I have showed in the case of well-being, even if we in practice can't observe any exceptions to a rule we nevertheless can't rule out the logical possibility that exceptions might exist. If medicine purported to be about the identification of absolutely healthy and unhealthy practices and substances binding on all humans, then the enterprise would likely be in vein, just like a normative morality purporting to identify absolute moral value. What than is it that allows us to speak of medical truths?

We can view medicine as an effort to *support* people's health. It does so by prescribing treatments for disease and as well as methods for increasing ones fitness. Of course, there is not a separate scientific effort for each individual human, and the prescriptions of medicine are based on research on general human anatomy. This is very similar to what Harris imagines a science of morality will be about. At one point he even argues that we "must […] define "good" as that which *supports* well-being." (ML p. 12, *italics* added) A science of morality will produce prescriptions for how to increase personal well-being and avoid personal suffering, based on a scientific model of a general human type.

Any scientific model of human anatomy would necessarily allow for the variations that we observe in humans. We are thus dividing the notion of humanity into various traits. Some traits are shared by all humans: we are all carbon-based material beings and so on. We do however differ in many respects. The prescriptions of medicine will have different effect on the health of people with different traits. Prescription $P^1$ might for example increase the health of all humans with trait $T^1$, but reduce the health of people with trait $T^2$. If it can be determined that the health of $T^1$ humans will increase as a result of $P^1$ in all observed cases, we can make the induction that $P^1$ is healthy for $T^1$ humans. This is a normative claim. Any human with the trait $T^1$ will have a motivating reason to act in accordance with the prescription $P^1$. This indicative fact will be on par with all other scientific knowledge about

[22] This point makes up the bulk of Harris' response to the critics of ML, (Harris 2011).

the world. In case $P^1$ is the cure for a lethal disease, a $T^1$ human *will* die by not following the prescription.

The existence of various sub-types of humans, categorized by how their specific traits define how their health is affected by external causes is not necessary; it doesn't have to be this way. All humans could in principle be radically different, and be differently and arbitrarily affected by the same external cause. If this was the case then a science of medicine would be impossible, because we could not generate predictions of how a cause will affect more than one human at the time. However, as it happens, the universe is lawful, and similarly constituted structures are similarly affected by similar causes. Arsenic will cause a reduction in the health of every human who injects it.

How do we deal with prescriptions which affect only humans with a specific trait? Let's consider an everyday example. The judgment that milk is healthy for all rational agents is likely false, and so is the judgment that milk is healthy for all humans. The judgment that milk is healthy for (in relation to) all lactose tolerant humans is however plausible, and true given that milk is in fact healthy for lactose tolerant humans. We notice that no dilemma really arises at the sub-type level. On the level of all rational agents, and on the level of all humans, it would be impossible to determine a common norm for milk drinking, since milk has a radically different effect on members of each of these groups. The fact milk is conducive to the health of one large sub-group of humans and conducive to the suffering of others does not constitute a dilemma. In light of this scientific knowledge our choices are not to either ban milk (morally bad) or force people to drink it (morally good and binding). We can however establish that *access* to milk and information about milk drinking are good norms, because they are *supportive* of general health. By allowing access to milk we are producing and adding good consequences (increased total health) to the system, without subtracting from anyone's already existing health (because we are not forcing anyone to drink milk).

So what does this analogy tell us about moral truth? We have already established that moral judgments are cognitive, and therefore truth-apt. In the last few chapters I have been dealing with the problem of whether or not any moral judgments can be true. In this chapter I have argued, with Harris, that in case it is true that it is healthy to cure diseases and increase general health, it is also true that it is morally good to support practices which only increase personal well-being and to remove practices which are only conducive to personal suffering. I do think that the analogy between health and the relational definition of moral value holds,

and I think that it shows that at least some relational moral judgments are in fact true, by the same standard that other scientific facts are true. But can it address all moral questions?

Even in medicine there are disputes about what is healthy. Most treatments have so called "adverse effects", such that they affect health both positively and negatively. Chemotherapy for instance has a range of seriously unhealthy effects on the body. How unhealthy do the adverse effects have to get before the treatment as a whole can be unhealthy? The more serious problem is whether such conditions as grief or caffeine addiction are healthy or unhealthy. Such categorization seems to depend on what we mean by "healthy". Of course, Harris acknowledges such questions remains, precisely because we don't have an exact definition of health. It is the large amount of conditions and causes that medicine can successfully categorize as healthy and unhealthy which supports medicine as a domain of genuine truth claims. In order for the analogy between health and morality to be a good and useful analogy, we would expect most relevant moral propositions (claims about what is good and bad) to be categorizable as either true or false, in the same way that most medical propositions (claims about what is healthy and unhealthy) are categorizable as either true or false. It is however still not clear whether any but the most obvious moral propositions can be categorized as true or false on this account.

A science of morality modeled on medicine might have a very limited reach. What can be shown to be morally bad on this account might be limited to such obvious cases as showing that a forced mass injection of arsenic constitute a morally bad act, as it would cause the suffering of all humans. Nevertheless, if the theory allows us to establish any moral truths as all, it seems it can count as a minimal form of moral realism. The tasks I specified for Harris in the introduction was to (i) show that moral judgments are cognitive, and (ii) show that there exist empirically available truth-conditions capable of justifying *at least some* moral judgments. I think that the analogy with health helps us understand how Harris proposes to solve this second task. It is still unclear, however, whether or not we can develop a proper and useful normative theory unless it can help us resolve more than obvious cases of relational good and bad. It could even be argued that by "*moral* questions" we just *mean* the kind of questions which just can't be resolved in this way. Thomas Nagel seems to support such a view:

> "Harris rejects [fact/value dualism] in the only way it can be rejected—by pointing to evaluative truths so obvious that they need no defense. For example, a world in which everyone was maximally miserable would be

> worse than a world in which everyone was happy, and it would be wrong to try to move us toward the first world and away from the second. This is not true by definition, but it is obvious, just as it is obvious that elephants are larger than mice. If someone denied the truth of either of those propositions, we would have no reason to take him seriously." (Nagel 2010)

Even if Nagel acknowledges that some evaluative propositions are obviously true, he doesn't think that any substantive or "real" moral questions can be answered without the presumption of a principle of maximization. Defining moral truth by analogy to health is then clearly only the foundational element in a "real" moral theory. In the next chapter I start addressing moral disagreement and consider if this "minimal" moral realism supports the resolution of substantial moral questions.

# Chapter 9: Being Wrong about Value

This chapter and the next address two different kinds of moral disagreement, which can be seen to follow from the theory so far. The first kind of disagreement occurs when people who share the relevant traits are simply wrong about what is conducive to their well-being. This is the kind of disagreement that I have claimed is solvable on Harris' theory. Most disagreement about what is healthy falls into this category and is therefore solvable, such as "does X cause disease?" and "will Y increase my fitness?". As I have argued, knowledge of how people with various traits respond to X and Y allows us to establish norms which are supportive of health.

Part of Nagel's critique of Harris was that even if it could show that there are a few "obvious" moral truths, it failed to address substantial moral questions. What exactly is a "substantial" moral question? "Normal" moral questions are about what all humans have motivating reasons to do. The way I have defined it based on the contemporary moral debate, to claim that X is morally right is to claim that all humans has motivating reason to X, which means that X is conducive to the well-being of all humans. Given that Harris has successfully provided a model which establishes the factual moral value of certain kind of states and actions, he has shown that we can answer a few moral questions on this definition.

When Gilbert Harman (1996) argues for moral relativism he argues that the persistence of moral disagreement despite the availability of all facts to all sides indicates that such disagreement is unsolvable and that no one is morally right. Such disagreement on Harman's account traces back to the involved agents having different *basic* values. Harman argues that for each moral dilemma there will always be agents who value both sides. There will even be some agents whose basic values support detonating nuclear bombs. This is interesting, because it shows that given the right circumstances even the most obvious of truths about what is conducive to the well-being of humans can be disbelieved. Harman relies on empirical evidence to support his claim. We always find at least some people or cultures who—despite access to all the same facts—disagrees with even the most widely accepted moral norms. Harman's core argument is that since there is no difference in the factual basis for the varying values, then there is no objective standard to determine which is "right".

The point I wish to make is that all moral questions in Harman's sense are "substantial". As long as people in fact disagree and are differently motivated, despite access

to the same knowledge, it is not straightforward and "unsubstantial" to resolve the disagreement. In case this can be done, a substantial moral question has been answered.

On Harris' account—as we have seen—people can have different basic values, but this would be because they are differently constituted on the structural level. In case we can show—by way of science—that two or more disagreeing parties really are similar enough so that X has the same effect on their well-being, we have solved moral disagreement.

The other kind of moral disagreement is caused by people really having different basic values; when all parties—with full information—correctly judges different states to be the most valuable for them.

How can we know if a moral dispute belongs to the first or second category? The answer is that we can't really know that until we have either shown that the norm or the belief in question is conducive to the well-being of humans and that those who fails to become motivated by it are wrong about the value of the desire that motivates them, or shown that the disagreement is caused by a difference in traits. Both projects are incredibly difficult because they rely on knowledge about the functioning of the human nervous system and of consciousness. We still have a very limited understanding of these phenomena. Given that the universe is material and lawful, as it appears to be, we see that such knowledge must exist, much in the same way that knowledge of human anatomy exists, and existed in principle even before it became known. We see that both these categories apply to health as well as to well-being. It is possible to disagree about X's conduciveness to health, either by being wrong about the actual consequences X has on the human body, or because X actually have different consequences on different people. The science of medicine is currently much further along in its quest to generate medical truth then science of morality is in generating moral truth. Harris' point is that these are in principle the same projects. When people are arguing over whether or not cigarette smoke, wine or vaccines are healthy we don't automatically know if the disagreement is caused by people being wrong about the consequences of these particular substances or if the substances are in fact healthy for some and unhealthy for others. However once it has been established that cigarette smoke is unhealthy for *all* humans, the prescription "don't smoke" is normatively binding on all humans.[23] We notice that it is entirely possible to remain delusional about the health effects of smoking. Some people will perhaps be convinced by the propaganda of the tobacco industry. Even if one comes to believe the science, quitting smoking can be close to impossible. Coming to believe that smoking is

[23] Normative for achieving optimal health. In moral terms it seems possible to value unhealthiness, or particular causes of unhealthiness, in case they are conducive of well-being in the current circumstance.

unhealthy provides a reason to not smoke, but it doesn't necessarily override the motivation caused by the habit of smoking. We see that the argument by which Harman wanted to establish moral relativism also seems to establish medical relativism. Every time medicine establishes that X is unhealthy there will be some people who will continue to believe that it is healthy, and many people will continue to practice it.

*

Harris' spends a good amount of space discussing real life example of norms which he thinks qualify as morally wrong. As we have seen he means by this that those people who adopt the norm as an action-guiding belief are wrong about the norm's effect on their well-being and the behavior prescribed by the norm is not instrumental in achieving or maintaining homeostasis or well-being, but has the opposite effect. A particularly striking example in the society created by the Dobu islanders:

> "Every Dobuan's primary interest was to cast spells on other members of the tribe in an effort to sicken or kill them and in the hopes of magically appropriating their crops. [...] The conscious application of magic was believed necessary for the most mundane tasks. Even the work of gravity had to be supplemented by relentless wizardry. [...] To make matters worse, the Dobu imagined that good fortune conformed to a rigid law of thermodynamics: if one man succeeded in growing more yams than his neighbor, his surplus crop must have been pilfered through sorcery. As all Dobu continuously endeavored to steal one another's crops by such methods, the lucky gardener is likely to have viewed his surplus in precisely these terms. A good harvest, therefore, was tantamount to "a confession of theft." This strange marriage of covetousness and magical thinking created a perfect obsession with secrecy in Dobu society. Whatever possibility of love and real friendship remained seems to have been fully extinguished by a final doctrine: the power of sorcery was believed to grow in proportion to one's intimacy with the intended victim. This belief gave every Dobuan an incandescent mistrust of all others, which burned brightest on those closest. Therefore, if a man fell seriously ill or died, his misfortune was immediately blamed on his wife, and vice versa. The picture is of a society completely in thrall to antisocial delusions. (ML p. 60-62)

The judgment of the practices and norms of the Dobu society as morally wrong, does of course not rely on an evaluative *premise*. The judgment is based on facts about human beings and what type of external causes result in well-being. In so far as we can prove that living alone and in constant fear—as is the state Harris attributes to the Dobus—will systematically cause humans to suffer, then the Dobu norms are really bad.

Harris' favorite example seems to be the society based on the norms of the Taliban, which for years now has been a prominent part of the modern world, and of which Harris has the following to say:

> "I think it is quite clear that members of the Taliban are seeking well-being in this world (as well as hoping for it in the next). But their religious beliefs have led them to create a culture that is almost perfectly hostile to human flourishing. Whatever they think they want out of life—like keeping all women and girls subjugated and illiterate—they simply do not understand how much better life would be for them if they had different priorities." (ML p. 36-37)

As we saw in chapter 2, Harris' agenda includes criticizing religion, and his moral theory does provide him with a tool for doing so. Religious beliefs that contradict scientific knowledge about how to improve well-being—perhaps by false promises of even greater well-being in the next life—are on Harris' account morally bad.

*

As mentioned humans do have the ability to form believes about what is conducive to their well-being, an ability which has freed us from a reliance on only being motivated by our predispositions. We have of course also evolved biases in belief formation. That is, it is easier for us to form some beliefs than other beliefs. Michael Shermer explores the issue of belief formation in "The Believing Brain, *How We Construct Beliefs and Reinforce Them as Truths*" (2011) where he for example identifies a strong tendency to attribute intentionality and agency to process which has none. Shermer points to the evolutionary advantage of making such attributions. Those who tend to form the belief that the sound from the bushes is only the wind will be correct most of the time but will be eaten when then sound is a lion. Whereas those who always think it is a lion, will never be eaten. Shermer then uses this point to tentatively explain various modern version of such attribution, such as gods, ghosts and

conspiracy theories. On the other end, forming the belief that it is correct to switch doors in the Monty Hall problem, even when the obvious fact is carefully explained, turns out to be very difficult for the human brain.[24] Shermer's theory seems coherent enough, if a little speculative and difficult to prove. The larger point however is that such biases are real, regardless of how we got them. An extensive overview of cognitive biases from a psychological perspective is provided in Daniel Kahneman's "Thinking, Fast and Slow" (2011), which sums up the research tradition on heuristics and cognitive biases which Kahneman himself, together with his colleague Amos Tversky, has had a central role both in founding and developing over the years. This research strongly corroborates the claim that the beliefs human beings form in all areas of life are systematically and predictably biased.

From the point of view of science—which we are currently occupying—the truth of a proposition does not depend on being believed. The fact that we systematically fail to form certain true beliefs and systematically succeed at forming certain false believes does not chance the truth-value of the propositions. Moral psychologists such as Joshua Greene—currently the director of the moral cognition lab at Harvard university—has shown in several essays that our moral judgments are subject to similar biases. We are for example very quick to form certain believes about consensual incest and pushing fat people onto railroad tracks, even in scenarios where such actions are obviously conducive to well-being (see for example Greene's doctoral dissertation (2002), which Harris also references in ML).

One interesting theory which seeks to explain the moral judgments we do make comes from physiologists Jonathan Haidt and Selin Kesebir (2010), and has been discussed by Haidt in other works as well. They have studied the moral judgments actually made by people, and identified five broad categories of moral concerns:

> 1. Harm/care: Concerns for the suffering of others, including virtues of caring and compassion.
>
> 2. Fairness/reciprocity: Concerns about unfair treatment, cheating, and more abstract notions of justice and rights.

[24] In the "Monty Hall problem" you are asked by a game-show host to choose one out of three doors. Hidden behind one of the doors is a car which you can win, and hidden behind the other two doors are goats. After having made your choice the host opens one of the doors you didn't choose and reveals a goat. He then asks if you want to switch to the other of the now two remaining doors. Initially you had 1/3 chance of hitting the right door, but if you switch you have 2/3 chance of hitting. Few people form this belief that it is right to switch doors without having it explained to them, and many people fail to form the belief even when they have all the relevant information. It is that difficult for the human brain to form this true belief.

> 3. In-group/loyalty: Concerns related to obligations of group membership, such as loyalty, self- sacrifice, and vigilance against betrayal.
>
> 4. Authority/respect: Concerns related to social order and the obligations of hierarchical relationships, such as obedience, respect, and the fulfillment of role – based duties.
>
> 5. Purity/sanctity: Concerns about physical and spiritual contagion, including virtues of chastity, wholesomeness, and control of desires.

These five categories are predispositions, shared by all humans. This doesn't mean that all humans come to value these five things, rather, Haidt suggests that these five things make up what we can value, and what we actually end up valuing depends on the particular circumstances in which we grow up. The first three categories are nevertheless close to universal, in that everyone values them to some degree, whereas the last two categories are only valued by some. These predispositions have evolved to become a part of human nature. The environments in which the human species has evolved were quite different from the social and technological environment we are facing today. It is therefore no necessary connection between the conduciveness of these values to human well-being today. This is precisely the point. As long as our well-being is the goal of our actions, we need to figure out what is conducive of this state in our *current* environment. In order to do so we must see past our predispositions and the things we are thought to value. Our failure to convince the Taliban that they are wrong then, is irrelevant.

Note that in order to establish this conclusion about the particular norms of the Taliban and the Dobus, we are assuming some scientific conclusions about well-being that are not a part of Harris' meta-ethical theory itself. But what if these norms turned out to be conducive to the well-being of some of the Dobus or some of the Taliban? What do we do when we discover that a set of norms really is conducive of the well-being of some and to the suffering of others?

# Chapter 10: Moral Disagreement

This chapter is about how to handle the second category of moral disagreement, the kind caused by groups of people having different traits, thus giving them a different relation to X.

Consider the following simple thought experiment: Two types of people—$T^1$ and $T^2$—live in a world governed by the set of norms "X", under which everyone suffers slightly but equally. We learn through science that the set of norms "Y" is conducive of the increased well-being of everyone, and is the alternative which generates the most total well-being. We also learn that the set of norms "Z" is conducive to the maximized well-being of $T^1$, but leaves $T^2$ in the state of slight suffering.

If we follow the objection we saw Blackford make earlier, $T^1$ humans will be fully justified in valuing Z over Y, because Z is conducive to what for them is the most desirable state possible, as a matter of fact. Based on the theory of natural value, we would have to agree. We have already claimed that it is a fact of conscious agents that they are motivated to do what is in its best interest, so if a conscious agent is so constituted as to really desire a state which involved the suffering of others, then that is the state it will be motivated to achieve. There seems to be no way around this. If this thought experiment accurately represents the nature of moral dilemmas belonging to the second category, then I think we simply have to concede that science can't solve them.

But why would we think that most moral questions belong to this second and unsolvable category? Isn't it rather reasonable to believe—based on what we currently know—that most moral dilemmas are caused by people being wrong about what is conducive to human (including their own) well-being, and as such are solvable by science?

What I think Harris wants to suggest is that almost all moral questions we can ask belong to the first and solvable category. When we carefully analyze our reasons for actions, the cases where a group of people are genuinely motivated to act so as to produce a state where another group suffers—and as such form an unsolvable moral dilemma—are very rare, if they exist at all. I think Harris' point is that there is a finite, and in fact very limited, range of possible norms which contribute to human well-being. Based on what we currently know about human nature and human circumstances we can effectively rule out norms which leave a group of people in suffering.

> "It seems clear that what we are really asking when we wonder whether a certain state of pleasure is "good," is whether it is conducive to, or obstructive of, some deeper form of well-being. This question is perfectly coherent; it surely has an answer (whether or not we are in a position to answer it); and yet, it keeps notions of goodness anchored to the experience of sentient beings. Defining goodness in this way does not resolve all questions of value; it merely directs our attention to what values actually are—the set of attitudes, choices, and behaviors that potentially affect our well-being, as well as that of other conscious minds. While this leaves the question of what constitutes well-being genuinely open, there is every reason to think that this question has a finite range of answers. Given that change in the well-being of conscious creatures is bound to be a product of natural laws, we must expect that this space of possibilities—the moral landscape—will increasingly be illuminated by science." (ML p. 12-13)

In this paragraph Harris' reminds us that values *are* "the set of attitudes, choices, and behaviors that potentially affect our well-being, as well as that of other conscious minds.". Nothing else *can* be valuable. Harris goes on to say that "while this leaves the question of what constitutes well-being genuinely open, there is every reason to think that this question has a finite range of answers.". Even if this formulation is a bit vague, it illuminates the central part of Harris' model. I take this, and other similar statements, to mean that the question of what constitutes an actual material state of ultimate moral value has a finite and limited range of possible answers.

Like the case is with milk-drinking, we are typically not limited to two possible norms—do or don't do—for each type of action. Many things can be done with milk which will typically be true of each commodity. Consider that we could select any norm for what to do with milk to be binding on all humans. We could of course ban milk or we could force everyone to drink it. We could allow milk, but not produce any. We could offer milk to everyone but not inform them of the consequences of drinking it. We could produce milk, and empty it all into the ocean. The list is close to endless. We could come up with many different norms for the act of murder as well. Banning it or demanding it of people are possibilities. We could allow murder, but randomly incinerate murderers with sky-lasers. We could allow people to murder each other using sky-lasers. We could murder every fifth child born. We cold murder criminals convicted of horrible crimes and deemed unredeemable. The point is that any state where people are allowed to incinerate each other with sky lasers, or where

every fifth child is murdered, are most likely not included in this range of states where it is possible for *any* human to flourish, based on the actual consequences of such norms.

*

Let's take a step back and start at the beginning of the method which I think Harris wants to claim leads us to the solution of moral disagreement. At first we observe different groups of people making different moral judgments. As a model of what actually happens when most people make moral judgment, we have said that the moral judgment that "X is good" amounts to the judgment that "X is good for me and similar organisms including all similarly situated humans". We then try to figure out what cause the disagreement. Is the disagreement caused by tradition, established cultural norms, genetic predisposition or any other contingent factor which "trick" us into being motivated away from what really is good for us? Alternatively, is there a real biological, or rather neurological, difference between the disagreeing groups (as it is in the example that opened this chapter), such that none of the groups are wrong about what is conducive to their well-being? In the latter case the disagreeing groups can legitimately be said to have different and opposing values, as in the example. So, what makes us think that this disagreement can be resolved?

The point I am going to make now might sound trivial, but I think it is important, and I believe that when we think about morality in absolutistic terms it is easy to overlook this simple point. If we accept that a moral judgment is the judgment that "X is good for all humans", or that "all humans has a motivating reason to X in the context S", we can see rather plainly that in case X really is conducive to the well-being of one group and the suffering of another, then the judgment that "X is morally good" is false. X is thus not included in the range of possible norms which is conducive to human well-being. We realize however that this false belief is constitutive of the moral disagreement. Consider that $T^1$ humans believe that the set of norms "Z" is conducive of their maximized well-being and that Z is morally valuable. By wrongly believing that $T^2$ humans also has motivating reason to Z, the $T^1$ humans fail to recognize all the consequences of Z, where it to be enforced as a moral norm. $T^1$ humans are in this case simply not aware of the fact that they will not be able to convince the $T^2$ humans that Z is conducive of their well-being.

The belief that "Z is conducive of the $T^1$ humans' maximized well-being" depended on the false belief that "everyone has a motivating reason to Z". Thus, in order for the value-disagreement to remain, we have to assume that $T^1$ humans remain motivated to Z, even after

realizing that Z causes the suffering of others. It will remain true that the perfect state of the universe for $T^1$ humans is the state where everyone has a motivating reason to Z and acts accordingly, however this state as it turned out is not a *real possibility*. Genuine disagreement about what ought to be done then can only exist as long as a group of people are so constituted as to be genuinely motivated to bring about a state which leaves another group suffering. I am not claiming that no such situations can occur, and I don't think Harris argues that either.

We can see that if we accept the evaluative premise that the most total well-being is valuable to all humans—which is what Nagel takes Harris to presume—then all such dilemmas would be solvable. That is the point of making such an evaluative premise. If we merely presume that other people's well-being is valuable to all humans—which is what Blackford takes Harris to presume—we are not thereby saying how much the well-being of others matter compared to one's own or that of one's kin. Such an evaluative premise is nevertheless sufficient to make the above thought experiment incoherent: It could not be the case that Z was conducive to the maximized well-being of $T^1$, because $T^1$ would be concerned about the well-being of $T^2$ to which Z is not conducive. We might instead find that as a matter of fact, when calculating in the unsatisfied concern for other people, that the compromise Y is the norm which is more conducive of the well-being of $T^1$. This would, as mentioned, require $T^1$ humans to relinquish their false belief that Z is morally valuable; the belief that *everyone* has a motivating reason to act in accordance with Z.

I have been very persistent in denying that Harris' moral theory is founded on an evaluative premise, even when Harris himself seems somewhat ambiguous about the question. Allowing for the acceptance of such a principle reduced the theory to something like plain utilitarianism. As mentioned in chapter 5, I do think that the value of the well-being of others plays a central part in Harris' theory, for the reason I have just explained. By accepting that the well-being of others is in fact valuable to humans, we can actually rule out norms which are conducive to the suffering of some, and in a sense "force" compromises. We are thereby able to solve many of these apparently "unsolvable" moral dilemmas.

I don't however think that the value of the well-being of others has the role of an evaluative premise in Harris' theory. I think that Harris considers it an obvious inductive fact that human beings to some extent value the well-being of others. This amounts to just the kind of fact that I have argued his meta-ethical theory can establish. I think Harris considers it equally obvious that all humans have a motivating reason to secure the well-being of others, as they have to not detonate a nuclear bomb, and that elephants are larger than mice. As such

this value is not a presumption but follows from the theory. The existence of such a value does not change the theory itself, but it changes its reach. The possibility that human beings are really capable of valuing states which include the unnecessary suffering of others diminishes, increasing the plausibility that moral questions fall under the solvable category. Thus, if the well-being of others can really be said to be a value, that might be the contingent fact about our nature that enables us to answer more than obvious moral questions. In the next and final chapter I turn to discussing whether or not we can justify such a value.

# Chapter 11: Reconsidering a "Principle" of Maximization

In the last two chapters I have discussed two different forms of moral disagreement. The first kind of disagreement is caused by people being wrong about what is conducive to well-being, including their own. This can cause people who are practically identical to disagree about what is conducive of their well-being. The other kind of disagreement is caused by people who are different, and are right about which hypothetical state their maximized well-being consists in. I have argued that in case people are actually different in this way, the ideal state for one type of people will necessarily be less than ideal for people with different values. In the last chapter I suggested that we have very little reason to believe that the maximized well-being of any group of humans includes living side by side with other people whose motivations goes unfulfilled. It is very important to realize that this is not a presumption or a required premise for Harris' theory to work, even though Harris himself sometimes gives the impression that it is. That the well-being of others is a value which is shared by all human beings is supposed to be a fact with a similar claim to truth as the fact that cancer is unhealthy. Neither of these facts are absolute, since we can't rule out the logical possibility of a being that is so constituted as to not value the well-being of others or a being to which cancer is conducive of health. These facts are inductive, like all scientific facts. We are therefore equally obligated to take seriously those who claim that cancer is healthy, fire is chilly and the suffering of others is morally good.

In order for this contingent fact to be true, we do of course have to show that there are no exceptions to this value. Harris does argue this case in "The Moral Landscape". His argument is very simplistic and as I have mentioned it does at times seem like Harris is simply presuming this value. In a very obvious way, all humans do have to care about other people. Everyone have to care what everyone else is motivated to do and what they actually do. This is because other people's actions and their reasons for them have an effect of everyone's personal well-being. This is fairly standard consequentialist reasoning. In his books on religion (Harris 2004 & Harris 2008) Harris argues the case that people's religious believes help shape their actions. We share a world with people of different religious beliefs and their actions and beliefs concern us because they affect our well-being. Harris attacks the notion that having faith and believing in God is a virtue, constitutive of good character and worthy of our praise. He attacks the notion that believing substantial truths about the nature of the

universe for which there is no evidence is innocent and morally unproblematic. For example, the belief shared by several religions about reward and punishment being distributed in an afterlife distorts our motivation in this life. The kind of rewards offered to Muslims for inflicting pain, suffering and death upon infidels—if genuinely believed—is guiding the actions of the believer through the maze of practical possibilities towards a miserable end for everyone. These books also attack the norm by which religious and other non-scientific beliefs about the world are allowed to flourish. Harris asks, what happens when Muslim extremists (or other people with extreme false beliefs about well-being, such as many North-Koreans) gain the ability to eradicate their enemies with the push of a button, or enforce their views upon the rest of the world? Why would they not use this opportunity to ensure (what they believe to be) the well-being of everyone?

All false beliefs on Harris' account are dangerous to various degrees if only because they prevent us from acquiring true beliefs about the world and about well-being. Harris' argument must not be seen as an argument against the freedom of speech or belief, which is an important tool in our quest to form true beliefs about the world. He is calling for serious criticism of beliefs that are harmful, to the believer and to others. Most such beliefs are condemned by societies, and the behavior they propagate is regulated. The belief that speed limits for cars should not exist is rightfully not thought to children and the resulting behavior is banned. One of the main points of "The End of Faith", which is also one of the main points of Dawkins' "The God Delusion" (2006), is that religious believes are currently enjoying an immunity to criticism, a norm which is allowing dangerous beliefs to flourish. This norm is thus not supportive of well-being, and therefore morally bad on Harris' account. The point for the current discussion is that other people's actions and believes concern us. We are concerned that other people's actions and beliefs are conducive of our personal well-being.

This point alone doesn't answer our question. How is it that the well-being of others is conducive to our well-being? Harris does surprisingly little to answer this question directly, and he seems to assume that everyone is better off if the people they surround themselves with are satisfied. There are however several good reasons to support such an assumption, and I will now briefly consider a few. In chapter 9 I mentioned Haidt's study of moral judgments, which revealed that it is a close to universal concern for (i) "the suffering of others, including virtues of caring and compassion", (ii) "unfair treatment, cheating, and more abstract notions of justice and rights", (iii) "obligations of group membership, such as loyalty, self- sacrifice, and vigilance against betrayal". While Hadith's study reveals that we are in fact concerned about avoiding the suffering of others, this is merely a predisposition and, as I have argued,

does not show that this concern is conducive of our well-being. We are interested to know if this concern, that we happen to have evolved, is true.

Why is it that almost all of us are concerned about harm/care, fairness/reciprocity and in-group/loyalty, as well as being predisposed to care about authority/respect and purity/sanctity? It seems that all these innate concerns function so as to generate behavior which again contributes to social stability. They ensure that we are motivated to take care of each other, to trust each other and to cooperate within our social group. Importantly we are also motivated to be vigilant against betrayal and to reciprocate by punishing those who betray and rewarding those who are loyal. We are also disposed to be motivated to follow leaders and to respect a hierarchical order, and we are disposed to be motivated to be civilized and control desires. So it seems that the moral judgments we do make all serve the purpose of creating and enforcing stable and reliable social conditions.

I have talked about Damasio's concept of homeostasis, which is the internal state an organism "wants" to be in. By "want" I am talking about the organism's natural "interest" in continued existence. These are the kind of "interest" that an organism must have in order to be selected for by evolution, and which we can observe in some form in all evolved organisms. A homeostatic state lays within a very narrow range of parameters, including such factors as body temperature, the presence of exactly the right amount of certain chemicals, no damage to the intricate physical structure of the organism and so on. External causes changes these parameters, and the all organisms depend on very specific environmental factors, such as the pretense of nutrition, specific weather conditions and so on. Damasio suggests, as we saw, that conscious organisms experience being in homeostasis as well-being. Naturally, other humans are a part of our environment, and can be a huge threat to our homeostasis. Other people can deprive us of food, damage us or simply kill us. Our well-being therefore depends on securing conditions where other people don't do these things, and don't threaten to do these things. Damasio introduces the term "sociocultural homeostasis" (2010 p. 308-315). Just as we will be motivated to secure the right internal parameters and the right environmental parameters, we will be motivated to secure the right social parameters. Like most internal states of the body and most environmental states, most states of society will not be conducive to the well-being of any human. We can't help but experience various degrees of suffering when our internal state and our natural and social environments changes and moves outside our parameters for homeostasis and well-being. Damasio also employs the term "homeostatic range" to describe the range of possible states—internal, environmental and social—which doesn't threaten the organism's homeostasis. Haidt's five basic human concerns complement

Damasio's theory very well. These traits have been selected for because they motivate us to act so at to secure our sociocultural homeostasis, at the expense of those traits that did not.

Observing that people judge it to be good to increase the well-being of others does not rule out that everyone is delusional when doing so. However, observing at the physical level that people's personal well-being increases a result of helping others does rule out delusions. In order to make such observation we have to have some indication of what human bodies in homeostasis look like, so that when someone reports that he suffers greatly after having helped someone, we can observe his neural and other bodily processes and catch his lie. We are beginning to understand the neural underpinnings of well-being. In "Braintrust" (2010) Patricia Churchland argues that sociability is a basic value, not only for humans, but for all social mammals. She provides much data in support of this hypothesis, but I won't rehearse everything here. Much of her argument is centered on the hormones oxytocin and vasopressin, which in humans correlate with experienced well-being. The research Churchland presents shows that these hormones are released into the human brain following acts of caring for others as well as physical contact, childbirth and various other social interactions. Oxytocin can also be administered directly and has some interesting effects while active in the brain. There is a measurable increase in trust towards others and a measurable decrease in aggression. It is also experienced as pleasurable. If oxytocin hadn't taken on the function is has in humans the kind of cooperation and trust which is characteristic of humans wouldn't have evolved. "Sociability" is a human trait which inclines us towards pro-social behavior, and friendly social interactions. Such interactions reward all parties with pleasurable experiences and reduce hostility and aggression in those involved. I think Churchland is right to argue that sociability is a basic human value. Being sociable seems to be a central component in securing one's own sociocultural homeostasis, and one's own well-being.

I don't wish to read too much into this research, or draw too broad conclusions based on the available results. I think there is ample research in support of the conclusion that the well-being of others is generally valuable to humans. I am simply pointing out the very basics. Harris' theory is in any case not dependent on any final conclusions about the actual moral value of the well-being of others. If we can show that the well-being of others is valuable *to most people to some degree*, and predict that it will continue to be so, we have come a long way towards showing that much moral disagreement can be solved. This again helps strengthen the analogy with health—and Harris' account as a whole—by showing that a science of morality too can successfully categorize enough condition and causes as good or bad to really be considered a domain of genuine truth claims. We saw in the last chapter that

unsolvable moral disagreement was between two groups of differently constituted people who both genuinely preferred a state where their particular interests would be prioritized at the expense of the interest of the other group. I argued that such unsolvable disagreement only seems likely to occur when both groups are so constituted as to not be concerned with the consequences of leaving the other group in a state suffering. The neglected group, we can speculate, is likely to feel unjustly treated, feel a need to reciprocate, deny the other group the benefits of their social cooperation, and so on. In effect the neglected group will to some degree threaten the sociocultural homeostasis, and the well-being, of the group that wish this suffering upon them. The stronger this threat turns out to be, the likelier it is that *a compromise really is the most valuable alternative for everyone*, that there really never was an unsolvable moral disagreement and that the disagreeing parties were just wrong.

One group of human beings at least nevertheless seems to provide an exception to this general valuing of the well-being of others. Sadistic psychopaths certainly share traits which distinguish them significantly from the rest of the human population. It seems fitting to end my treatment of Harris' moral theory by discussing what may be the most prominent threat to any kind of moral realism, and see if the theory we have arrived at can deal with the case of sadistic psychopaths.[25]

When sadists claim that they are motivated to torture and kill innocent people, it seems we have every reason to believe them. But is the state of the world where sadists get to kill and torture at will actually best for them? We can perhaps assume that an arrangement where the sadist was free to do whatever he liked without there being any consequences would be preferable to the sadist. Of course, this would probably be true for everyone else as well. None of these are actual possibilities. No one is free to do what they like, because actions do have consequences. In order to determine if torturing and killing innocent people really is valuable to the sadist we have to take into account the reactions this behavior will create in other people. I would like to make the uncontroversial observation that when a sadist tortures and kills someone, and threatens to repeat such behavior, all affected humans—arguably also including other psychopaths—*must* become motivated to remove the threat by some means, and likely also *will* become motivated to execute revenge on the sadist. Thus, by torturing and killing innocents the sadistic psychopath is not *only* causing himself pleasure, but he is also risking "everything" else that he cares about, so to speak. As I have said, in case people are

[25] By this I mean people who are so constituted as to lack empathy and to derive pleasure from cruelty and the infliction of pain. This is primarily a thought experiment and doesn't address the deep complexities of sadism or psychopathy. For an interesting discussion of these traits and their relation to our concept of "evil" see for example Simon Baron-Cohen's "The Science of Evil" (2011).

concerned with avoiding torture, death, imprisonment and other social sanctions it is fair to say that actively provoking others is not the best way to secure such a state, sadists are no exception. It is possible that some sadists realize this, they are after all not lacking in cognitive functions. The urge to kill might be like a drug addiction, which the sadist can't resist but which causes him suffering because he knows what he is forsaking.

This doesn't rule out the possibility that the sadist's desire to torture and kill is so strong, and the received pleasure so intense, that it is *worth* forsaking the benefits of living in peace with society. The sadist's deeper well-being could in principle consist in knowing that he has tortured and killed many people. It might even consist in being on the run from punishment and retribution. Again, this will depend on what the individual is like. Would such an individual constitute an exception to the fact that all humans to some degree value the well-being of others?

We can't, according to Harris, think of maximized personal well-being in terms of what the individual is actually like at any particular time. Sadistic psychopaths are human beings, but they have a trait (the lack of empathy) which is determinative of what is conducive of their well-being. Because of this trait, this group of people are unable to experience the deeper well-being associated with being in sociocultural homeostasis, available to other humans. They are as such prevented from reaching the peaks of human well-being. An extreme, but appropriate, comparison could be a person in coma. Could we really say that since this person is so constituted as to be in a coma—perhaps his state happens to be caused by a brain tumor—he is in a state of maximized well-being and health? Because of his physical constitution this person is prevented from being motivated to change his state. It seems uncontroversial to say that it is in this person's interest to have the tumor removed. We are determining how healthy this person is and how much he flourishes, not by what his current physical constitution allows for, but by how healthy he can *potentially* be and how much well-being he can *potentially* experience as a human being. As long as we *know* that his health or his well-being will increase as the result of even a physical alteration of the body, we can say that he has a motivating reason do so. His comprehension of the benefits of such a procedure is irrelevant to his reason.

It is a small point that we can't—at least not yet—cure psychopathy, but the fact that we can't cure a physical disease doesn't prevent us from truthfully stating that the sick individual is not healthy. We can compare psychopathy and sadism to physical disease as long as we can prove that people who have these traits aren't experiencing as much well-being as *it is possible for humans to experience*. Psychopathy can be seen as the name for the condition

one has when ones' empathic ability is working incorrectly, provided we acknowledge that the correct way is to secure homeostasis and maximize well-being.

The purpose of this chapter has been to consider the role that the value of other people's well-being plays in Harris theory. Since Harris' focus on the value of well-being in general has been misinterpreted as an evaluative premise, I feel that this was necessary to do. We know that the well-being of others is valuable to many people and I have tried to explain why this is, and how this helps us to solve moral disagreement, even in the case of sadistic psychopaths.

# Conclusion

"The Moral Landscape" is a different approach to meta-ethical questions. Sam Harris argues from a naturalist standpoint that there are objective answers to moral questions; questions that it is up to *science* to answer. My main goal in this thesis has been to *consider whether Harris' theory has the resources to face up to the explanatory challenges found in contemporary meta-ethical literature, and whether it constitutes a viable naturalist explanation of morality*. I haven't compared Harris' particular view to other alternative meta-ethical explanations of morality, or attempted argue that Harris' explanation is the best available. My focus has been on Harris' theory itself as I have attempted to show that it is a serious candidate for the explanation of morality, which can rise up to challenge the various existing explanations. More specifically, I set up two explanatory challenges for Harris: (i) Showing that moral judgments are truth-apt mental states, like beliefs, and (ii) showing that there exist empirically available truth-conditions capable of justifying at least some moral judgments. I also said that Harris must show both these things without compromising a fundamental feature of moral judgments, namely that they provide us with a reason for acting in accordance with them. To be clear, my main goal has not been to show that Harris actually and straightforwardly meets these challenges, but that the *kind of* naturalistic and scientific explanation he suggests is *capable* of meeting these challenges, and thus can't be dismissed as a non-starter, the way naturalistic versions of moral realism typically are.

The nature of these challenges and the conceptual problems involved in meeting them were elaborated in chapter 2: "The Moral Problem". I then went on to show how Harris can meet the first of these challenges in chapters 3 and 4, by way of what I called *the theory of natural value*. In establishing this view I emphasized the rejection of value absolutism in favor of a form of *value relationalism*. Accepting value relationalism allowed us to reject—or at least seriously challenge—*fact/value dualism*, the view that facts and values are distinct existences. By doing so we could claim that values are reducible to—or at least determined by—observable physical and structural properties of conscious organisms, a move that is typically unavailable without the rejection of fact/value dualism. Harris' particular view—as we saw—is that claims about value are really claims about the *well-being of conscious organisms*, and further that the well-being of conscious organisms is caused by external objects and events according to lawful and observable natural processes. This means that what

is valuable to a given conscious organism is determined by the organism's *relation* to the world; the particular way that the material structure which constitutes the organism is affected by external events.

I think that what is said in chapters 3 and 4 succeeds at showing that it is possible to establish cognitivism without compromising the motivational feature of morality within an entirely naturalistic framework, as long as value absolutism is rejected. An important conclusion from these chapters is that what is established is cognitivism about *judgments of relational value*, which as a category doesn't necessarily include (what we usually mean by) *moral judgments*. Success at establishing cognitivism about *moral judgments* will thus depend on whether we can accept that moral judgments are a form of relational judgments. In chapters 5 and 6 I went on to argue—what I take to be Harris' view—that judgments of relational value include moral judgments, when moral judgments are taken to be judgments about what is valuable to all humans, or in Harris' terms, what is conducive to the well-being of humans in general. I think that this relational definition of moral judgments accurately captures what we mean by moral judgments as defined in chapter 2. I therefore think that Harris' theory is capable of meeting the first explanatory challenge. This means that on Harris' theory, moral judgments are truth-apt.

The second challenge—which I consider the more difficult of the two—is discussed in chapters 7 through 11. As shown in chapter 7, it appears as if the only way that a relational moral judgment can be *true*, is if we make a non-scientific assumption about the nature of this relation. We could assume, for example, that all human beings—arguably also including future humans—are so constituted as to value the well-being of other people, or maximized total well-being. Science is incapable of showing us that this is the case. It is objected that if no such evaluative premise is assumed, there can be no truth conditions which could justify the truth of relational moral judgments. This is because human beings aren't identical, and as such don't share an identical relation to external events. A moral judgment proposes that, and is true if, all humans have a reason to act in accordance with it. But as there is no guarantee that any relational moral judgment—for example that caring for each other is good—will hold for all humans, it seems such judgments fail to have the appropriate binding force.

Harris proposes to solve this problem by allowing moral truth to be inductive, like all other scientific truth. By doing so we can say that at least some actions—like raising the temperature on earth to 200 degrees centigrade—are morally bad, because such actions are systematically and reliably conducive to the suffering of *all* humans. There is of course no *guarantee* that there will not be some humans who flourish in such temperatures, but based on

what we know of biology the emergence of such a person seems as unlikely as air suddenly becoming heavier then water. Harris suggests that no one has any real reason to act against such judgments as "blowing up the earth is conducive to your suffering". You could wrongly *believe* that blowing up the earth is *conducive of well-being*, but as the theory of natural value tells us, the well-being of conscious organisms is caused by external objects and events according to lawful and observable natural processes, so your beliefs don't change the fact that exploding is conducive of suffering (if this is the case). Harris thus suggests that if we can observe that an action reliably causes humans beings to suffer, we can say that the action *is conducive to the suffering of all humans*, and therefore bad. Further we can predict that the action will continue to case suffering in all future instances, and this prediction will be as binding on future humans as the law of gravity.

Does Harris by this show that he *can* meet the second challenge (showing that there exist empirically available truth-conditions capable of justifying at least some moral judgments)? I think it does, even though there remains much doubt about to what *degree* the challenge is *actually* met by "The Moral Landscape", as I have attempted to show. Even if his theory allows for only a few, and very obvious, moral truths, it can count as a minimal form of moral realism. This conclusion is still open for debate though, because it might be reasonable to demand that meta-ethical moral realism should allow for the development of methods for solving substantial moral questions. Such methods include accepting evaluative principles—such as the principle of utility—on rational grounds, but as we have seen, Harris denies that this is possible.

In chapters 9 through 11, I suggest how Harris can approach the hard questions of ethics. My primary strategy has been to argue that one of the obvious inductive moral facts that come out of science on Harris' model is that it is valuable to all humans to secure their own "sociocultural homeostasis", which basically involves at least some degree of caring about and cooperating with other humans. As such it is possible to say that the well-being of others has some (relational) value for everyone. I think Harris supports this view, and he does provide some arguments for it. I mostly seems like he just accepts that the well-being of others is valuable to everyone, and I have suggested that this is what causes many of his critic to attribute this evaluation to him as a premise, and not as a contingent fact.

All successful actions change the state of the universe as a consequence. A good way to understand the essence of Harris' moral theory is that among the almost endless states the universe can be changed into as a result of our actions, very few include humans in states of well-being. Most states of the universe don't even support human life, and most of those that

do contain too extreme conditions to be conducive of well-being; from lack of food and water to oppressive dictatorships and violent conflicts. Mapping the consequences of events in the world on human nervous systems helps us to narrow down the field of actions we can be said to have a reason to take. How far down the field can be narrowed will depend on how homogenous human nervous systems are. If brains turn out to be as homogenous as bodies (and the mental lawfully depends on the physical), it seems we have much reason to suspect that there are very narrow parameters where brains can be as happy and satisfied as they possibly can, in which case it seems plausible that a science of morality could be developed to become as factual and rational as medicine.

In "The Moral Landscape" we find Harris making the following concluding remarks:

> "If we are not able to perfectly reconcile the tension between personal and collective well-being, there is still no reason to think that they are generally in conflict. Most boats will surely rise with the same tide. It is not at all difficult to envision the global changes that would improve life for everyone: We would all be better off in a world where we devoted fewer of our resources to preparing to kill one another. Finding clean sources of energy, cures for disease, improvements in agriculture, and new ways to facilitate human cooperation are general goals that are obviously worth striving for. What does such a claim mean? It means that we have every reason to believe that the pursuit of such goals will lead upward on the slopes of the moral landscape." (ML p. 188)

*

Do we now think that science can determine human values? Maybe. Perhaps at least some values, in some sense. Do we now understand how scientific knowledge can be said to be constitutive of human value? I hope so, at least given the naturalistic framework on which this theory is based. We must however acknowledge that there remains genuine disagreement about whether or not Harris' theory deserves to be categorized as moral realism. If the conceptual framework of meta-ethics requires realist theories to give an evaluative premise in order to secure the resolution of moral disagreement, then Harris' theory fails, because it can't make such a non-empirical assumption.

Harris himself does on occasions seemingly suggest that it might be unnecessary to talk about a science of *morality* and of *moral* truth. We could instead simply talk about a science

of well-being and of truths about well-being, and reserve the moral terminology for absolutistic theories. The normativity of these truths about well-being would not change as a result of such semantics. It is a fairly uncontroversial proposition that even if we reject the reality of (absolute) moral truth, we can still criticize various actions and practices. Russell Blackford—who has followed us through this thesis—views Harris theory as such a non-realist, but normative, position:

> "It is quite open to us to condemn traditional systems of morality to the extent that they are harsh or cruel, rather than providing what most of us (quite rationally) want from a moral tradition: for example that it ameliorate suffering, regulate conflict, and provide personal security and social cooperation, yet allow individuals a substantial degree of discretion to live their lives as they wish." (Blackford 2010)

Harris responds to Blackford's view by saying that:

> "I'm afraid I have seen too much evidence to the contrary to accept Blackford's happy talk on this point. I consistently find that people who hold this view are far less clear-eyed and committed than (I believe) they should be when confronted with moral pathologies—especially those of other cultures—precisely because they believe there is no deep sense in which any behavior or system of thought can be considered pathological in the first place." (Harris 2011)

By referring to this exchange I wish to support the conclusion that there is remaining disagreement about whether or not non-absolutistic and normative theories really deserves to be categorizes under "moral realism".

Harris' insistence that the kind of moral truth he argues for *really* is moral truth, seems to me justified by his claim that these truths are on par with all other scientific truth, and thereby deserving of the same recognition. Based on Harris rejection of fact/value dualism and value absolutism I think that his claim to moral realism makes sense. Without value absolutism there could be no absolute moral truth in any case. I think that Harris theory amounts to a very good and well supported explanation of both the specified explanatory challenges—in its own currently limited and non-absolutistic way—as I have attempted to show. I will therefore conclude that Harris' theory of "The Moral Landscape" constitutes an appealing naturalist alternative to the other available explanations of morality. I do however

think that many aspects of the theory—specifically its normative implications, its reach, and its status as moral realism—needs to clarified and developed further, before we can start to seriously compare this theory to the rest of the alternatives. I think this is worth doing.

# References


Baron-Cohen, S. 2011, *The Science of Evil: On Empathy and the Origins of Cruelty*, Basic Books, New York.

Bentham, J. 1780, An Introduction to the Principles of Morals and Legislation, Dover Publications, Mineola.

Blackford, R. 2010, "Book review: Sam Harris' The Moral Landscape", *Journal of Evolution and Technology*, Vol. 21, Issue 2 (December 2010), p. 53-62.

Churchland, P. 2010, *Braintrust: What Neuroscience Tells us about Morality*, Princeton University Press, Princeton.

Damasio, A. 2010, *The Self Comes to Mind: Constructing the Conscious Brain*, Vintage, New York.

Darwall, S., Gibbard, A. and Railton, P. 1992, "Toward Fin de siècle Ethics: Some Trends", *The Philosophical Review*, Vol. 101, No. 1, p. 115-189.

Dawkins, R. 1976. *The Selfish Gene*, 30th Anniversary edition (2006), Oxford University Press, Oxford.

Dawkins, R. 2006, *The God Delusion*, Black Swan, London.

Dennett, D. C. 1995, *Darwin's Dangerous Idea: Evolution and the Meaning of Life*, Simon & Schuster Paperbacks, New York.

Dennett, D. C. 2007, *Breaking the Spell: Religion as a Natural Phenomenon*, Penguin Books, London.

Greene, J. D. 2002, *The terrible, horrible, no good, very bad truth about morality and what to do about it*, Princeton University, Princeton. Available at URL: < http://www.wjh.harvard.edu/~jgreene/GreeneWJH/Greene-Dissertation.pdf > (Viewed May 4th 2013)

Haidt, J., & Kesebir, S. 2010, "Morality", in S. Fiske, D. Gilbert & G. Lindsky (Eds.), *Handbook of Social Phsycology*, 5th Edition, p. 797-832, Wiley & Sons, Hoboken. Available by request from Haidt's homepage <http://people.stern.nyu.edu/jhaidt/publications.html> (Viewed May 4th 2013).

Harman, G. and Thompson, J. J. 1996, *Moral Relativism and Moral Objectivity*, Blackwell Publishing, Cambridge.

Harris, S. 2004, *The End of Faith: Religion, Terror and the Future of Reason*, W. W. Norton, New York.

Harris, S. 2007, *Letter to a Christian Nation*, Transworld Publishers, London.

Harris, S., Sheth, S. A. and Cohen, M. S. 2008, "Functional Neuroimaging of Belief, Disbelief, and Uncertainty", *Annals of Neurology*, Vol. 63, Issue 2, p. 141-147.

Harris, S., Kaplan, J. T., Curiel, A., Bookheimer, S. Y., Iacoboni, M., and Cohen, M. S. 2009, "The neural correlates of religious and nonreligious belief", *PLoS ONE*, 4 (10), e7272.

Harris, S. 2010, *The Moral Landscape: How Science can Determine Human Values*, Bantam Press, London.

Harris, S. 2011, "A Response to Critics", on *Huffintonpost.com*, URL: <http://www.huffingtonpost.com/sam-harris/a-response-to-critics_b_815742.html> (Published January 29. 2011, viewed May 4th 2013)

Harris, S. 2012, *Free Will*, Free Press, New York.

Hawking, S. and Mlodinow, L. 2010, *The Grand Design*, Bantam Books, New York.

Hitchens, C. 2007, *God is Not Great: How religion Poisons Everything*, Atlantic Books, London.

Hookway, C. "Pragmatism", *The Stanford Encyclopedia of Philosophy* (Spring 2010 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2010/entries/pragmatism/>.

Hume, D. 1739-1740, *A Treatise of Human Nature: Being an Attempt to Introduce the Experimental Method of Reasoning into Moral Subjects*, Penguin Books, London.

Kahneman, D. 2011, *Thinking, Fast and Slow*, Penguin Books Ltd, London.

Kant, I. 1785, "Grunnlegging til Moralens Metafysikk", in Storheim, E. (red.), *Immanuel Kant: Morallov og Frihet*, 1970, Gyldendal, Oslo.

Jollimore, T. 2010, Review of *The Moral Landscape*, by Sam Harris, on *barnesandnoblerwview.com*, URL: <http://bnreview.barnesandnoble.com/t5/Reviews-Essays/The-Moral-Landscape/ba-p/3477> (Published October 22th 2012, viewed May 4. 2013)

Krauss, L. M. 2012, *A Universe from Nothing: Why There is Something Rather than Nothing*, Free Press, New York.

Mill, J. S. 1863, "Utilitarianism", in Crisp, R. (Ed.) 1998, *Oxford Philosophical Texts: J. S. Mill Utilitarianism*, Oxford university Press, Oxford.

Moore, G. E. 1903, *Principia Ethica*, Barnes & Noble Books, New York.

Nagel, T. 2010, "The Facts Fetish", *New Republic*, November 11, 2010 Issue. Available online at URL: < http://www.newrepublic.com/article/books-and-arts/magazine/78546/the-facts-fetish-morality-science#> (viewed May 4th 2013).

Papineau, D. 2007, “Naturalism”, *The Stanford Encyclopedia of Philosophy* (Spring 2009 Edition), Edward N. Zalta (ed.), URL: <http://plato.stanford.edu/archives/spr2009/entries/naturalism> (viewed May 4th 2013).

Rawls, J. 1971, *A Theory of Justice*, Revised Edition (1999), Oxford University Press, Oxford.

Railton, P. 1986a, “Moral Realism”, *The Philosophical Review*, XCV, No.2 (April 1986), in Railton, P. (red.) 2003, *Facts, Values and Norms: Essays Toward a Morality of Consequence*, Cambridge University Press, Cambridge, p. 3-42.

Railton, P. 1986b, “Facts and Values”, *Philosophical Topics*, XIV, No.2 (Fall 1986), in Railton, P. (red.) 2003, *Facts, Values and Norms: Essays Toward a Morality of Consequence*, Cambridge University Press, Cambridge, p. 34-68.

Shermer, M. 2011, *The Believing Brain: From Ghosts and Gods to politics and Conspiracies—How We Construct Beliefs and reinforce them as truth*, Times Books, New York.

Singer, P. 1981, *The Expanding Circle: Ethics, Evolution, and Moral Progress*, Princeton University Press, Princeton.

Smith, M. 1994, *The Moral Problem*, Blackwell Publishing, Malden.

Stenger, V. 2009, *The New Atheism: Taking a Stand for Science and Reason*, Prometheus Books, New York.

Stenger, V. 2010, “What’s New about the New Atheism?”, *Philosophy Now*, Issue 78 (April/May 2010).