Toward a Science of Morality

Over the past couple of months, I seem to have conducted a public experiment in the manufacture of philosophical and scientific ideas. In February, I spoke at the 2010 TED conference, where I briefly argued that morality should be considered an undeveloped branch of science. Normally, when one speaks at a conference the resulting feedback amounts to a few conversations in the lobby during a coffee break. I had these conversations at TED, of course, and they were useful. As luck would have it, however, my talk was broadcast on the internet just as I was finishing a book on the relationship between science and human values, and this produced a blizzard of criticism at a moment when criticism could actually do me some good. I made a few efforts to direct and focus this feedback, and the result has been that for the last few weeks I have had literally thousands of people commenting upon my work, more or less in real time. I can't say that the experience has been entirely pleasant, but there is no question that it has been useful.

If nothing else, the response to my TED talk proves that many smart people believe that something in the last few centuries of intellectual progress prevents us from making cross-cultural moral judgments -- or moral judgments at all. Thousands of highly educated men and women have now written to inform me that morality is a myth, that statements about human values are without truth conditions and, therefore, nonsensical, and that concepts like "well-being" and "misery" are so poorly defined, or so susceptible to personal whim and cultural influence, that it is impossible to know anything about them. Many people also claim that a scientific foundation for morality would serve no purpose, because we can combat human evil while knowing that our notions of "good" and "evil" are unwarranted. It is always amusing when these same people then hesitate to condemn specific instances of patently abominable behavior. I don't think one has fully enjoyed the life of the mind until one has seen a celebrated scholar defend the "contextual" legitimacy of the burqa, or a practice like female genital excision, a mere thirty seconds after announcing that his moral relativism does nothing to diminish his commitment to making the world a better place. Given my experience as a critic of religion, I must say that it has been disconcerting to see the caricature of the over-educated, atheistic moral nihilist regularly appearing in my inbox and on the blogs. I sincerely hope that people like Rick Warren have not been paying attention.
First, a disclaimer and non-apology: Many of my critics fault me for not engaging more directly with the academic literature on moral philosophy. There are two reasons why I haven't done this: First, while I have read a fair amount of this literature, I did not arrive at my position on the relationship between human values and the rest of human knowledge by reading the work of moral philosophers; I came to it by considering the logical implications of our making continued progress in the sciences of mind. Second, I am convinced that every appearance of terms like "metaethics," "deontology," "noncognitivism," "anti-realism," "emotivism," and the like, directly increases the amount of boredom in the universe. My goal, both in speaking at conferences like TED and in writing my book, is to start a conversation that a wider audience can engage with and find helpful. Few things would make this goal harder to achieve than for me to speak and write like an academic philosopher. Of course, some discussion of philosophy is unavoidable, but my approach is to generally make an end run around many of the views and conceptual distinctions that make academic discussions of human values so inaccessible. While this is guaranteed to annoy a few people, the prominent philosophers I've consulted seem to understand and support what I am doing.

Many people believe that the problem with talking about moral truth, or with asserting that there is a necessary connection between morality and well-being, is that concepts like "morality" and "well-being" must be defined with reference to specific goals and other criteria -- and nothing prevents people from disagreeing about these definitions. I might claim that morality is really about maximizing well-being and that well-being entails a wide range of cognitive/emotional virtues and wholesome pleasures, but someone else will be free to say that morality depends upon worshipping the gods of the Aztecs and that well-being entails always having a terrified person locked in one's basement, waiting to be sacrificed.

Of course, goals and conceptual definitions matter. But this holds for all phenomena and for every method we use to study them. My father, for instance, has been dead for 25 years. What do I mean by "dead"? Do I mean "dead" with reference to specific goals? Well, if you must, yes -- goals like respiration, energy metabolism, responsiveness to stimuli, etc. The definition of "life" remains, to this day, difficult to pin down. Does this mean we can't study life scientifically? No. The science of biology thrives despite such ambiguities. The concept of "health" is looser still: it, too, must be defined with reference to specific goals -- not suffering chronic pain, not always vomiting, etc. -- and these goals are continually changing. Our notion of "health" may one day be defined by goals that we cannot currently entertain with a straight face (like the goal of spontaneously regenerating a lost limb). Does this mean we can't study health scientifically?

I wonder if there is anyone on earth who would be tempted to attack the philosophical underpinnings of medicine with questions like: "What about all the people who don't share your goal of avoiding disease and early death? Who is to say that living a long life free of pain and debilitating illness is 'healthy'? What makes you think that you could convince a person suffering from fatal gangrene that he is not as healthy you are?" And yet, these are precisely the kinds of objections I face when I speak about morality in terms of human and animal well-being. Is it possible to voice such doubts in human speech? Yes. But that doesn't mean we should take them seriously.

The physicist Sean Carroll has written another essay in response to my TED talk, further arguing that one cannot derive "ought" from "is" and that a science of morality is impossible. Carroll's essay is worth reading on its own, but in the hopes of making the difference between our views as clear as possible, I have I excerpted his main points in their entirety, and followed them with my comments.

Carroll begins:
I want to start with a hopefully non-controversial statement about what science is. Namely: science deals with empirical reality -- with what happens in the world. (I.e. what "is.") Two scientific theories may disagree in some way -- "the observable universe began in a hot, dense state about 14 billion years ago" vs. "the universe has always existed at more or less the present temperature and density." Whenever that happens, we can always imagine some sort of experiment or observation that would let us decide which one is right. The observation might be difficult or even impossible to carry out, but we can always imagine what it would entail. (Statements about the contents of the Great Library of Alexandria are perfectly empirical, even if we can't actually go back in time to look at them.) If you have a dispute that cannot in principle be decided by recourse to observable facts about the world, your dispute is not one of science.

I agree with Carroll's definition of "science" here -- though some of his subsequent thinking seems to depend on a more restrictive definition. I especially like his point about the Library of Alexandria. Clearly, any claims we make about the contents of this library will be right or wrong, and the truth does not depend on our being able to verify such claims. We can also dismiss an infinite number of claims as obviously wrong without getting access to the relevant data. We know, for instance, that this library did not contain a copy of The Catcher in the Rye. When I speak about there being facts about human and animal well-being, this includes facts that are quantifiable and conventionally "scientific" (e.g., facts about human neurophysiology) as well as facts that we will never have access to (e.g., how happy would I have been if I had decided not to spend the evening responding to Carroll's essay?).

With that in mind, let's think about morality. What would it mean to have a science of morality? I think it would look have to look something like this:

Human beings seek to maximize something we choose to call "well-being" (although it might be called "utility" or "happiness" or "flourishing" or something else). The amount of well-being in a single person is a function of what is happening in that person's brain, or at least in their body as a whole. That function can in principle be empirically measured. The total amount of well-being is a function of what happens in all of the human brains in the world, which again can in principle be measured. The job of morality is to specify what that function is, measure it, and derive conditions in the world under which it is maximized.

Good enough. I would simply broaden picture to include animals and any other conscious systems that can experience gradations of happiness and suffering -- and weight them to the degree that they can experience such states. Do monkeys suffer more than mice from medical experiments? (The answer is almost surely "yes.") If so, all other things being equal, it is worse to run experiments on monkeys than on mice.

Skipping ahead a little, Carroll makes the following claims:

I want to argue that this program is simply not possible. I'm not saying it would be difficult -- I'm saying it's impossible in principle. Morality is not part of science, however much we would like it to be. There are a large number of arguments one could advance for in support of this claim, but I'll stick to three.
1. There's no single definition of well-being.

People disagree about what really constitutes "well-being" (or whatever it is you think they should be maximizing). This is so perfectly obvious, it's hard to know what to defend. Anyone who wants to argue that we can ground morality on a scientific basis has to jump through some hoops.

First, there are people who aren't that interested in universal well-being at all. There are serial killers, and sociopaths, and racial supremacists. We don't need to go to extremes, but the extremes certainly exist. The natural response is to simply separate out such people; "we need not worry about them," in Harris's formulation. Surely all right-thinking people agree on the primacy of well-being. But how do we draw the line between right-thinkers and the rest? Where precisely do we draw the line, in terms of measurable quantities? And why there? On which side of the line do we place people who believe that it's right to torture prisoners for the greater good, or who cherish the rituals of fraternity hazing? Most particularly, what experiment can we imagine doing that tells us where to draw the line?

This is where Carroll and I begin to diverge. He also seems to be conflating two separate issues: (1) He is asking how we can determine who is worth listening to. This is a reasonable question, but there is no way Carroll could answer it "precisely" and "in terms of measurable quantities" for his own field, much less for a nascent science of morality. How flakey can a Nobel laureate in physics become before he is no longer worth listening to -- indeed, how many crazy things could he say about matter and space-time before he would no longer even count as a "physicist"? Hard question. But I doubt Carroll means to suggest that we must answer such questions experimentally. I assume that he can make a reasonably principled decision about whom to put on a panel at the next conference on Dark Matter without finding a neuroscientist from the year 2075 to scan every candidate's brain and assess it for neurophysiological competence in the relevant physics. (2) Carroll also seems worried about how we can assess people's claims regarding their inner lives, given that questions about morality and well-being necessarily refer to the character subjective experience. He even asserts that there is no possible experiment that could allow us to define well-being or to resolve differences of opinion about it. Would he say this for other mental phenomena as well? Is it impossible to define or study this state of mind empirically? I'm not sure how deep Carroll's skepticism runs, but much of psychology now appears to hang in the balance. Of course, Carroll might want to say that the problem of access to the data of first-person experience is what makes psychology often seem to teeter at the margin of science. He might have a point -- but, if so, it would be a methodological point, not a point about the limits of scientific truth. Remember, the science of determining exactly which books were in the Library of Alexandria is stillborn and going absolutely nowhere, methodologically speaking. But this doesn't mean we can't be absolutely right or absolutely wrong about the relevant facts.

As for there being many people who "aren't interested in universal well-being," I would say that more or less everyone, myself included, is insufficiently interested in it. But we are seeking well-being in some form nonetheless, whatever we choose to call it and however narrowly we draw the circle of our moral concern. Clearly many of us (most? all?) are not doing as good a job of this as we might. In fact, if science did nothing more than help people align their own selfish priorities -- so that those who really wanted to lose weight, or spend more time with their kids, or learn another language, etc., could get what they most desired -- it would surely increase the well-being of humanity. And this is to say nothing of what would happen if science could reveal depths of well-being that most of us are unaware of, thereby changing our
priorities.

Carroll continues:

More importantly, it's equally obvious that even right-thinking people don't really agree about well-being, or how to maximize it. Here, the response is apparently that most people are simply confused (which is on the face of it perfectly plausible). Deep down they all want the same thing, but they misunderstand how to get there; hippies who believe in giving peace a chance and stern parents who believe in corporal punishment for their kids all want to maximize human flourishing, they simply haven't been given the proper scientific resources for attaining that goal.

While I'm happy to admit that people are morally confused, I see no evidence whatsoever that they all ultimately want the same thing. The position doesn't even seem coherent. Is it a priori necessary that people ultimately have the same idea about human well-being, or is it a contingent truth about actual human beings? Can we not even imagine people with fundamentally incompatible views of the good? (I think I can.) And if we can, what is the reason for the cosmic accident that we all happen to agree? And if that happy cosmic accident exists, it's still merely an empirical fact; by itself, the existence of universal agreement on what is good doesn't necessarily imply that it is good. We could all be mistaken, after all.

In the real world, right-thinking people have a lot of overlap in how they think of well-being. But the overlap isn't exact, nor is the lack of agreement wholly a matter of misunderstanding. When two people have different views about what constitutes real well-being, there is no experiment we can imagine doing that would prove one of them to be wrong. It doesn't mean that moral conversation is impossible, just that it's not science.

Imagine that we had a machine that could produce any possible brain state (this would be the ultimate virtual reality device, more or less like the Matrix). This machine would allow every human being to sample all available mental states (some would not be available without changing a person's brain, however). I think we can ignore most of the philosophical and scientific wrinkles here and simply stipulate that it is possible, or even likely, that given an infinite amount of time and perfect recall, we would agree about a range of brain states that qualify as good (as in, "Wow, that was so great, I can't imagine anything better") and bad (as in, "I'd rather die than experience that again.") There might be controversy over specific states -- after all, some people do like Marmite -- but being members of the same species with very similar brains, we are likely to converge to remarkable degree. I might find that brain state X242358B is my absolute favorite, and Carroll might prefer X979793L, but the fear that we will radically diverge in our judgments about what constitutes well-being seems pretty far-fetched. The possibility that my hell will be someone else's heaven, and vice versa, seems hardly worth considering. And yet, whatever divergence did occur must also depend on facts about the brains in question.

Even if there were ten thousand different ways for groups of human beings to maximally thrive (all trade-offs and personal idiosyncrasies considered), there will be many ways for them not to thrive -- and the difference between luxuriating on a peak of the moral landscape and languishing in a valley of internecine horror will translate into facts that can be scientifically understood.

2. It's not self-evident that maximizing well-being, however defined, is
the proper goal of morality.

Maximizing a hypothetical well-being function is an effective way of thinking about many possible approaches to morality. But not every possible approach. In particular, it's a manifestly consequentialist idea -- what matters is the outcome, in terms of particular mental states of conscious beings. There are certainly non-consequentialist ways of approaching morality; in deontological theories, the moral good inheres in actions themselves, not in their ultimate consequences. Now, you may think that you have good arguments in favor of consequentialism. But are those truly empirical arguments? You're going to get bored of me asking this, but: what is the experiment I could do that would distinguish which was true, consequentialism or deontological ethics?

It is true that many people believe that "there are non-consequentialist ways of approaching morality," but I think that they are wrong. In my experience, when you scratch the surface on any deontologist, you find a consequentialist just waiting to get out. For instance, I think that Kant's Categorical Imperative only qualifies as a rational standard of morality given the assumption that it will be generally beneficial (as J.S. Mill pointed out at the beginning of Utilitarianism). Ditto for religious morality. This is a logical point before it is an empirical one, but yes, I do think we might be able to design experiments to show that people are concerned about consequences, even when they say they aren't. While my view of the moral landscape can be classed as "consequentialist," this term comes with fair amount of philosophical baggage, and there are many traditional quibbles with consequentialism that do not apply to my account of morality.

The emphasis on the mental states of conscious beings, while seemingly natural, opens up many cans of worms that moral philosophers have tussled with for centuries. Imagine that we are able to quantify precisely some particular mental state that corresponds to a high level of well-being; the exact configuration of neuronal activity in which someone is healthy, in love, and enjoying a hot-fudge sundae. Clearly achieving such a state is a moral good. Now imagine that we achieve it by drugging a person so that they are unconscious, and then manipulating their central nervous system at a neuron-by-neuron level, until they share exactly the mental state of the conscious person in those conditions. Is that an equal moral good to the conditions in which they actually are healthy and in love etc.? If we make everyone happy by means of drugs or hypnosis or direct electronic stimulation of their pleasure centers, have we achieved moral perfection? If not, then clearly our definition of "well-being" is not simply a function of conscious mental states. And if not, what is it?

Clearly, we want our conscious states to track the reality of our lives. We want to be happy, but we want to be happy for the right reasons. And if we occasionally want to uncouple our mental state from our actual situation in the world (e.g. by taking powerful drugs, drinking great quantities of alcohol, etc.) we don't want this to render us permanently delusional, however pleasant such delusion might be. There are some obvious reasons for this: We need our conscious states to be well synched to their material context, otherwise we forget to eat, ramble incoherently, and step in front of speeding cars. And most of what we value in our lives, like our connection to other people, is predicated on our being in touch with external reality and with the probable consequences of our behavior. Yes, I might be able to take a drug that would make me feel good while watching my young daughter drown in the bathtub -- but I am perfectly capable of judging that I do not want to take such a drug out of concern for my (and her) well-being.
Such a judgment still takes place in my conscious mind, with reference to other conscious mental states (both real and imagined). For instance, my judgment that it would be wrong to take such a drug has a lot to do with the horror I would expect to feel upon discovering that I had happily let my daughter drown. Of course, I am also thinking about the potential happiness that my daughter's death would diminish -- her own, obviously, but also that of everyone who is now, and would have been, close to her. There is nothing mysterious about this: Morality still relates to consciousness and to its changes, both actual and potential. What else could it relate to?

3. There's no simple way to aggregate well-being over different individuals.

The big problems of morality, to state the obvious, come about because the interests of different individuals come into conflict. Even if we somehow agreed perfectly on what constituted the well-being of a single individual -- or, more properly, even if we somehow "objectively measured" well-being, whatever that is supposed to mean -- it would generically be the case that no achievable configuration of the world provided perfect happiness for everyone. People will typically have to sacrifice for the good of others; by paying taxes, if nothing else.

So how are we to decide how to balance one person's well-being against another's? To do this scientifically, we need to be able to make sense of statements like "this person's well-being is precisely 0.762 times the well-being of that person." What is that supposed to mean? Do we measure well-being on a linear scale, or is it logarithmic? Do we simply add up the well-beings of every individual person, or do we take the average? And would that be the arithmetic mean, or the geometric mean? Do more individuals with equal well-being each mean greater well-being overall? Who counts as an individual? Do embryos? What about dolphins? Artificially intelligent robots?

These are all good questions: Some admit of straightforward answers; others plunge us into moral paradox; none, however, proves that there are no right or wrong answers to questions of human and animal wellbeing. I discuss these issues at some length in my forthcoming book. For those who want to confront how difficult it can be to think about aggregating human well-being, I recommend Derek Parfit's masterpiece, Reasons and Persons. I do not claim to have solved all the puzzles raised by Parfit -- but I don't think we have to.

Practically speaking, I think we have some very useful intuitions on this front. We care more about creatures that can experience a greater range of suffering and happiness -- and we are right to, because suffering and happiness (defined in the widest possible sense) are all that can be cared about. Are all animal lives equivalent? No. Are all human lives equivalent? No. I have no problem admitting that certain people's lives are more valuable than mine -- I need only imagine a person whose death would create much greater suffering and foreclose much greater happiness. However, it also seems quite rational for us to collectively act as though all human lives were equally valuable. Hence, most of our laws and social institutions generally ignore differences between people. I suspect that this is a very good thing. Of course, I could be wrong about this -- and that is precisely the point. If we didn't behave this way, our world would be different, and these differences would either affect the totality of human well-being, or they wouldn't. Once again, there are answers to such questions, whether we can ever answer them in practice.

I believe that covers the heart of Carroll's argument. Skipping ahead to final point:
And finally: pointing out that people disagree about morality is not analogous to the fact that some people are radical epistemic skeptics who don't agree with ordinary science. That's mixing levels of description. It is true that the tools of science cannot be used to change the mind of a committed solipsist who believes they are a brain in a vat, manipulated by an evil demon; yet, those of us who accept the presuppositions of empirical science are able to make progress. But here we are concerned only with people who have agreed to buy into all the epistemic assumptions of reality-based science -- they still disagree about morality. That's the problem. If the project of deriving ought from is were realistic, disagreements about morality would be precisely analogous to disagreements about the state of the universe fourteen billion years ago. There would be things we could imagine observing about the universe that would enable us to decide which position was right. But as far as morality is concerned, there aren't.

The biologist P.Z. Myers has thrown his lot in with Carroll on a similar point:

I don't think Harris's criterion -- that we can use science to justify maximizing the well-being of individuals -- is valid. We can't... Harris is smuggling in an unscientific prior in his category of well-being.

It seems to me that these two quotations converge on the core issue. Of course, it is easy enough for Carroll to assert that moral skepticism isn't analogous to scientific skepticism, but I think he is simply wrong about this. To use Myer's formulation, we must smuggle in an "unscientific prior" to justify any branch of science. If this isn't a problem for physics, why should it be a problem of a science of morality? Can we prove, without recourse to any prior assumptions, that our definition of "physics" is the right one? No, because our standards of proof will be built into any definition we provide. We might observe that standard physics is better at predicting the behavior of matter than Voodoo "physics" is, but what could we say to a "physicist" whose only goal is to appease the spiritual hunger of his dead ancestors? Here, we seem to reach an impasse. And yet, no one thinks that the failure of standard physics to silence all possible dissent has any significance whatsoever; why should we demand more of a science of morality?

So, while it is possible to say that one can't move from "is" to "ought," we should be honest about how we get to "is" in the first place. Scientific "is" statements rest on implicit "oughts" all the way down. When I say, "Water is two parts hydrogen and one part oxygen," I have uttered a quintessential statement of scientific fact. But what if someone doubts this statement? I can appeal to data from chemistry, describing the outcome of simple experiments. But in so doing, I implicitly appeal to the values of empiricism and logic. What if my interlocutor doesn't share these values? What can I say then? What evidence could prove that we should value evidence? What logic could demonstrate the importance of logic? As it turns out, these are the wrong questions. The right question is, why should we care what such a person thinks in the first place?

So it is with the linkage between morality and well-being: To say that morality is arbitrary (or culturally constructed, or merely personal), because we must first assume that the well-being of conscious creatures is good, is exactly like saying that science is arbitrary (or culturally constructed, or merely personal), because we must first assume that a rational understanding of the universe is good. We need not enter either of these philosophical cul-de-sacs.

Carroll and Myers both believe nothing much turns on whether we find a universal
foundation for morality. I disagree. Granted, the practical effects cannot be our reason for linking morality and science -- we have to form our beliefs about reality based on what we think is actually true. But the consequences of moral relativism have been disastrous. And science's failure to address the most important questions in human life has made it seem like little more than an incubator for technology. It has also given faith-based religion -- that great engine of ignorance and bigotry -- a nearly uncontested claim to being the only source of moral wisdom. This has been bad for everyone. What is more, it has been unnecessary -- because we can speak about the well-being of conscious creatures rationally, and in the context of science. I think it is time we tried.

Follow Sam Harris on Twitter: www.twitter.com/SamHarrisOrg

Share
Facebook
Twitter
Buzz up!