No-Nonsense Metaethics

Heading Toward: No-Nonsense Metaethics

Part of the sequence: No-Nonsense Metaethics

A few months ago, I predicted that we could solve metaethics in 15 years. To most people, that was outrageously optimistic. But I've updated since then. I think much of metaethics can be solved *now* (depending on where you draw the boundary around the term 'metaethics'.) My upcoming sequence 'No-Nonsense Metaethics' will solve the part that can be solved, and make headway on the parts of metaethics that aren't yet solved. Solving the easier problems of metaethics will give us a clear and stable platform from which to solve the *hard* questions of morality.

Metaethics has been my target for a while now, but first I had to explain the neuroscience of pleasure and desire, and how to use intuitions for philosophy.

Luckily, Eliezer laid *most* of the groundwork when he explained couldness, terminal and instrumental values, the complexity of human desire and happiness, how to dissolve philosophical problems, how to taboo words and replace them with their substance, how to avoid definitional disputes, how to carve reality at its joints with our words, how an algorithm feels from the inside, the mind projection fallacy, how probability is in the mind, reductionism, determinism, free will, evolutionary psychology, how to grasp slippery things, and what you would do without morality.

Of course, Eliezer wrote his own metaethics sequence. Eliezer and I seem to have similar views on morality, but I'll be approaching the subject from a different angle, I'll be phrasing my solution differently, and I'll be covering a different spread of topics.

Why do I think much of metaethics can be solved now? We have enormous resources not available just a few years ago. The neuroscience of pleasure and desire didn't exist two decades ago. (Well, we thought dopamine was 'the pleasure chemical', but we were wrong.) Detailed models of reductionistic meta-ethics weren't developed until the 1980s and 90s (by Peter Railton and Frank Jackson). Reductionism has been around for a while, but there are few philosophers who relentlessly play Rationalist's Taboo. Eliezer didn't write How an Algorithm Feels from the Inside until 2008.

Our methods will be familiar ones, already used to dissolve problems ranging from free will to disease. We will play Taboo with our terms, reducing philosophical questions into scientific ones. Then we will examine the cognitive algorithms that make it *feel* like open questions remain.

Along the way, we will solve or dissolve the traditional problems of metaethics: moral epistemology, the role of moral intuition, the is-ought gap, matters of moral psychology, the open question argument, moral realism vs. moral anti-realism, moral cognitivism vs. non-cognitivism, and more.

You might respond, "Sure, Luke, we can do the reduce-to-algorithm thing with free will or disease, but morality is different. Morality is *fundamentally normative*. You can't just dissolve moral questions with Taboo-playing and reductionism and cognitive science."

Well, we're going to examine the cognitive algorithms that generate *that* intuition, too.

And at the end, we will see what this all means for the problem of Friendly AI.

I must note that I didn't exactly *invent* the position I'll be defending. After sharing my views on metaethics with many scientifically-minded people in private conversation, many have said something like "Yeah, that's basically what I think about metaethics, I've just never thought it through in so much detail and cited so much of the relevant science [e.g. recent work in neuroeconomics and the science of intuition]."

But for convenience I do need to invent a *name* for my theory of metaethics. I call it *pluralistic moral reductionism*.

Next post: What is Metaethics?

What is Metaethics?

When I say I think I can solve (some of) metaethics, what exactly is it that I think I can solve?

First, we must distinguish the study of *ethics* or *morality* from the *anthropology* of moral belief and practice. The first one asks: "What is right?" The second one asks: "What do people think is right?" Of course, one can inform the other, but it's important not to confuse the two. One can correctly say that different cultures *have* different 'morals' in that they have different moral beliefs and practices, but this may not answer the question of whether or not they are behaving in morally right ways.

My focus is metaethics, so I'll discuss the anthropology of moral belief and practice only when it is relevant for making points about metaethics.

So what is metaethics? Many people break the field of ethics into three sub-fields: applied ethics, normative ethics, and metaethics.

Applied ethics: Is abortion morally right? How should we treat animals? What political and economic systems are most moral? What are the moral responsibilities of businesses? How should doctors respond to complex and uncertain situations? When is lying acceptable? What kinds of sex are right or wrong? Is euthanasia acceptable?

Normative ethics: What moral principles should we use in order to *decide* how to treat animals, when lying is acceptable, and so on? Is morality decided by what produces the greatest good for the greatest number? Is it decided by a list of unbreakable rules? Is it decided by a list of character virtues? Is it decided by a hypothetical social contract drafted under ideal circumstances?

Metaethics: What does moral language mean? Do moral facts exist? If so, what are they like, and are they reducible to natural facts? How can we know whether moral judgments are true or false? Is there a connection between making a moral judgment and being motivated to abide by it? Are moral judgments objective or subjective, relative or absolute? Does it make sense to talk about moral progress?

Others prefer to combine applied ethics and normative ethics so that the breakdown becomes: normative ethics vs. metaethics, or 'first order' moral questions (normative ethics) vs. 'second order' questions (metaethics).

Mainstream views in metaethics

To illustrate how people can give different answers to the questions of metaethics, let me summarize some of the mainstream philosophical positions in metaethics.

Cognitivism vs. non-cognitivism: This is a debate about what is happening when people engage in moral discourse. When someone says "Murder is wrong," are they trying to state a fact about murder, that it has the property of *being wrong*? Or are they merely expressing a negative emotion toward murder, as if they had gasped aloud and said "Murder!" with a disapproving tone?

Another way of saying this is that cognitivists think moral discourse is 'truth-apt' - that is, moral statements are the kinds of things that can be true or false. Some cognitivists think that all moral claims are in fact false (error theory), just as the atheist thinks that claims about gods are usually meant to be fact-stating but in fact are all false because gods don't exist.1 Other cognitivists think that at least some moral claims are true. *Naturalism* holds that moral judgments are true or false because of natural facts,2 while *non-naturalism* holds that moral judgments are true or false because of non-natural facts.3 *Weak cognitivism* holds that moral judgments can be true or false not because they agree with certain (natural or non-natural) opinion-independent facts, but because our considered opinions *determine* the moral facts.4

Non-cognitivists, in contrast, tend to think that moral discourse is *not* truth-apt. Ayer (1936) held that moral sentences express our emotions ("Murder? Yuck!") about certain actions. This is called *emotivism* or *expressivism*. Another theory is *prescriptivism*, the idea that moral sentences express commands ("Don't murder!").5 Or perhaps moral judgments express our acceptance of certain norms (*norm expressivism*).6 Or maybe our moral judgments express our dispositions to form sentiments of approval or disapproval (*quasi-realism*).7

Moral psychology: One major debate in moral psychology concerns whether moral judgments require some (defeasible) motivation to adhere to the moral judgment (*motivational internalism*), or whether one can make a moral judgment without being motivated to adhere to it (*motivational externalism*). Another debate concerns whether motivation depends on both beliefs *and* desires (the *Humean theory of motivation*), or whether some beliefs are by themselves intrinsically motivating (*non-Humean theories of motivation*).

More recently, researchers have run a number of experiments to test the mechanisms by which people make moral judgments. I will list a few of the most surprising and famous results:

- Whether we judge an action as 'intentional' or not often depends on the judged goodness or badness of the action, not the internal states of the agent.8
- Our moral judgments are significantly affected by whether we are in the presence of freshly baked bread or a low concentration of fart spray that only the subconscious mind can detect.9
- Our moral judgments are greatly affected by pointing magnets at the point in our brain that processes theory of mind.10
- People tend to insist that certain things are right or wrong even when a hypothetical situation is constructed such that they admit they can give no reason for their judgment.11
- We use our recently-evolved neocortex to make utilitarian judgments, and deontological judgments tend to come from evolutionarily older parts of our brains.12
- People give harsher moral judgments when they feel clean.13

Moral epistemology: Different views on cognitivism vs. non-cognitivism and moral psychology suggest different views of moral epistemology. How can we know moral facts? Non-cognitivists and error theorists think there *are* no moral facts to be known. Those who believe moral facts answer to non-natural facts tend to think that moral knowledge comes from intuition, which somehow has access to non-natural facts. Moral naturalists tend to think that moral facts can be accessed simply by doing science.

Tying it all together

I will not be trying very hard to fit my *pluralistic moral reductionism* into these categories. I'll be arguing about the substance, not the symbols. But it still helps to have a concept of the subject matter by way of such examples.

Maybe mainstream metaethics will make more sense in flowchart form. Here's a flowchart I adapted from Miller (2003). If you don't understand the bottom-most branching, read chapter 9 of Miller's book or else just don't worry about it. (Click through for full size.)

Next post: Conceptual Analysis and Moral Theory

Previous post: Heading Toward: No-Nonsense Metaethics

Notes

1 This is not quite correct. The error theorist can hold that a statement like "Murder is not wrong" is true, for he thinks that murder is not wrong *or* right. Rather, the error theorist claims that all moral statements *which presuppose the existence of a moral property* are false, because no such moral properties exist. See Joyce (2004). Mackie (1977) is the classic statement of error theory.

2 Sturgeon (1988); Boyd (1988); Brink (1989); Brandt (1979); Railton (1986); Jackson (1998). I have written introductions to the three major versions of moral naturalism: Cornell realism, Railton's moral reductionism (1, 2), and Jackson's moral functionalism.

3 Moore (1903); McDowell (1998); Wiggins (1987).

4 For an overview of such theories, see Miller (2003), chapter 7.

5 See Carnap (1937), p. 23-25; Hare (1952).

6 Gibbard (1990).

7 Blackburn (1984).

8 The Knobe Effect. See Knobe (2003).

9 Schnall et al. (2008); Baron & Thomley (1994).

10 Young et al. (2010). I interviewed the author of this study here.

11 This is moral dumfounding. See Haidt (2001).

12 Greene (2007).

13 Zhong et al. (2010).

References

Baron & Thomley (1994). A Whiff of Reality: Positive Affect as a Potential Mediator of the Effects of Pleasant Fragrances on Task Performance and Helping. *Environment and*

Behavior, 26(6): 766-784.

Blackburn (1984). *Spreading the Word*. Oxford University Press.

Brandt (1979). *A Theory of the Good and the Right*. Oxford University Press.

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

Boyd (1988). How to be a Moral Realist. In Sayre-McCord (ed.), *Essays in Moral Realism* (pp. 181-122). Cornell University Press.

Carnap (1937). *Philosophy and Logical Syntax*. Kegan Paul, Trench, Trubner & Co.

Gibbard (1990). *Wise Choices, Apt Feelings*. Clarendon Press.

Greene (2007). The secret joke of Kant's soul. In Sinnott-Armstrong (ed.), *Moral Psychology, Vol. 3: The Neuroscience of Morality: Emotion, Disease, and Development*. MIT Press.

Haidt (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*: 814-834

Hare (1952). *The Language of Morals*. Oxford University Press.

Jackson (1998). *From Metaphysics to Ethics*. Oxford UniversityPress.

Joyce (2001). *The Myth of Morality*. Cambridge University Press.

Knobe (2003). Intentional Action and Side Effects in Ordinary Language. *Analysis, 63*: 190-193.

Mackie (1977). *Ethics: Inventing Right and Wrong*. Penguin.

McDowell (1998). *Mind, Value, and Reality*. Harvard University Press.

Miller (2003). *An Introduction to Contemporary Metaethics*. Polity.

Moore (1903). *Principia Ethica*. Cambridge University Press.

Schnall, Haidt, Clore, & Jordan (2008). Disgust as embodied moral judgment. *Personality and Social Psychology Bulletin, 34(8)*: 1096-1109.

Sturgeon (1988). Moral explanations. In Sayre-McCord (ed.), *Essays in Moral Realism* (pp. 229-255). Cornell University Press.

Railton (1986). Moral realism. *Philosophical Review, 95*: 163-207.

Wiggins (1987). A sensible subjectivism. In *Needs, Values, Truth* (pp. 185-214). Blackwell.

Young, Camprodon, Hauser, Pascual-Leone, & Saxe (2010). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proceedings of the National Academy of Sciences, 107*: 6753-6758.

Zhong, Strejcek, & Sivanathan (2010). A clean self can render harsh moral judgment. *Journal of Experimental Social Psychology, 46 (5)*: 859-862

Conceptual Analysis and Moral Theory

Part of the sequence: No-Nonsense Metaethics. Also see: A Human's Guide to Words.

If a tree falls in the forest, and no one hears it, does it make a sound?

> Albert: "Of course it does. What kind of silly question is that? Every time I've listened to a tree fall, it made a sound, so I'll guess that other trees falling also make sounds. I don't believe the world changes around when I'm not looking."

Barry: "Wait a minute. If no one hears it, how can it be a sound?"

Albert and Barry are not arguing about facts, but about definitions:

> ...the first person is speaking as if 'sound' means acoustic vibrations in the air; the second person is speaking as if 'sound' means an auditory experience in a brain. If you ask "Are there acoustic vibrations?" or "Are there auditory experiences?", the answer is at once obvious. And so the argument is really about the definition of the word 'sound'.

Of course, Albert and Barry *could* argue back and forth about which definition best fits their intuitions about the meaning of the word. Albert could offer this argument in favor of using his definition of sound:

> My computer's microphone can record a sound without anyone being around to hear it, store it as a file, and it's called a 'sound file'. And what's stored in the file is the pattern of vibrations in air, not the pattern of neural firings in anyone's brain. 'Sound' means a pattern of vibrations.

Barry might retort:

> Imagine some aliens on a distant planet. They haven't evolved any organ that translates vibrations into neural signals, but they still hear sounds inside their own head (as an evolutionary biproduct of some other evolved cognitive mechanism). If these creatures seem metaphysically possible to you, then this shows that our concept of 'sound' is not dependent on patterns of vibrations.

If their debate seems silly to you, I have sad news. A large chunk of moral philosophy looks like this. What Albert and Barry are doing is what philosophers call conceptual analysis.1

The trouble with conceptual analysis

I won't argue that *everything* that has ever been called 'conceptual analysis' is misguided.2 Instead, I'll give *examples* of common kinds of conceptual analysis that corrupt discussions of morality and other subjects.

The following paragraph explains succinctly what is wrong with much conceptual analysis:

> Analysis [had] one of two reputations. On the one hand, there was sterile cataloging of pointless folk wisdom - such as articles analyzing the concept

> VEHICLE, wondering whether something could be a vehicle without wheels. This seemed like trivial lexicography. On the other hand, there was metaphysically loaded analysis, in which ontological conclusions were established by holding fixed pieces of folk wisdom - such as attempts to refute general relativity by holding fixed allegedly conceptual truths, such as the idea that motion is intrinsic to moving things, or that there is an objective present.3

Consider even the 'naturalistic' kind of conceptual analysis practiced by Timothy Schroeder in *Three Faces of Desire*. In private correspondance, I tried to clarify Schroeder's project:

> As I see it, [your book] seeks the cleanest reduction of the folk psychological term 'desire' to a natural kind, ala the reduction of the folk chemical term 'water' to H2O. To do this, you employ a naturalism-flavored method of conceptual analysis according to which the best theory of desire is one that is logically consistent, fits the empirical facts, and captures how we use the term and our intuitions about its meaning.

Schroeder confirmed this, and it's not hard to see the motivation for his project. We have this concept 'desire', and we might like to know: "Is there anything in the world similar to what we mean by 'desire'?" Science can answer the "is there anything" part, and intuition (supposedly) can answer the "what we mean by" part.

The trouble is that philosophers often take this "what we mean by" question so seriously that thousands of pages of debate concern *which definition to use* rather than *which facts are true* and what to anticipate.

In one chapter, Schroeder offers 8 objections4 to a popular conceptual analysis of 'desire' called the 'action-based theory of desire'. Seven of these objections concern our intuitions about the meaning of the word 'desire', including one which asks us to imagine the existence of alien life forms that have desires about the weather but have no dispositions to act to affect the weather. If our intuitions tell us that such creatures are metaphysically possible, goes the argument, then our concept of 'desire' need not be linked to dispositions to act.

Contrast this with a conversation you might have with someone from the Singularity Institute. Within *20 seconds* of arguing about the definition of 'desire', someone will say, "Screw it. Taboo 'desire' so we can argue about facts and anticipations, not definitions."5

Disputing definitions

Arguing about definitions is not *always* misguided. Words can be wrong:

> When the philosophers of Plato's Academy claimed that the best definition of a human was a "featherless biped", Diogenes the Cynic is said to have exhibited a plucked chicken and declared "Here is Plato's Man." The Platonists promptly changed their definition to "a featherless biped with broad nails."

Likewise, if I give a lecture on correlations between income and subjective well-being and I conclude by saying, "And *that*, ladies and gentlemen, is my theory of the atom," then you have some reason to object. Nobody else uses the term 'atom' to mean

anything remotely like what I've just discussed. If I ever do that, I hope you *will* argue that my definition of 'morality' is 'wrong' (or unhelpful, or confusing, or something).

Some unfortunate words are used in a *wide* variety of vague and ambiguous ways.6 Moral terms are among these. As one example, consider some commonly used definitions for 'morally good':

- that which produces the most pleasure for the most people
- that which is in accord with the divine will
- that which adheres to a certain list of rules
- that which the speaker's intuitions approve of in a state of reflective equilibrium
- that which the speaker generally approves of
- that which our culture generally approves of
- that which our species generally approves of
- that which we would approve of if we were fully informed and perfectly rational
- that which adheres to the policies we would vote to enact from behind a veil of ignorance
- that which does not violate the concept of our personhood
- that which resists entropy for as long as possible

Often, people can't tell you what they mean by moral terms when you question them. There is little hope of taking a survey to decide what moral terms 'typically mean' or '*really* mean'. The problem may be worse for moral terms than for (say) art terms. Moral terms have more powerful connotations than art terms, and are thus a greater attractor for sneaking in connotations. Moral terms are used to persuade. "It's just *wrong*!" the moralist cries, "I don't *care* what definition you're using right now. It's *just wrong*: don't do it."

Moral discourse is rife with motivated cognition. This is part of why, I suspect, people resist dissolving moral debates even while they have no trouble dissolving the 'tree falling in a forest' debate.

Disputing the definitions of moral terms

So much moral philosophy is consumed by debates over definitions that I will skip to an example from someone you might hope would know better: reductionist Frank Jackson7:

> ...if Tom tells us that what he means by a right action is one in accord with God's will, rightness according to Tom is being in accord with God's will. If Jack tells us that what he means by a right action is maximizing expected value as measured in hedons, then, for Jack, rightness is maximizing expected value...

But if we wish to address the concerns of our fellows when we discuss the matter - and if we don't, we will not have much of an audience - we had better mean what they mean. We had better, that is, identify our subject via the folk theory of rightness, wrongness, goodness, badness, and so on. We need to identify rightness as the property that satisfies, or near enough satisfies, the folk theory of rightness - and likewise for the other moral properties. It is, thus, folk theory that will be our guide in identifying rightness, goodness, and so on.8

The meanings of moral terms, says Jackson, are given by their place in a network of platitudes ('clauses') from folk moral discourse:

> The input clauses of folk morality tell us what kinds of situations described in descriptive, non-moral terms warrant what kinds of description in ethical terms: if an act is an intentional killing, then normally it is wrong; pain is bad; 'I cut, you choose' is a fair procedure; and so on.
>
> The internal role clauses of folk morality articulate the interconnections between matters described in ethical, normative language: courageous people are more likely to do what is right than cowardly people; the best option is the right option; rights impose duties of respect; and so on.
>
> The output clauses of folk morality take us from ethical judgements to facts about motivation and thus behaviour: the judgement that an act is right is normally accompanied by at least some desire to perform the act in question; the realization that an act would be dishonest typically dissuades an agent from performing it; properties that make something good are the properties we typically have some kind of pro-attitude towards, and so on.
>
> Moral functionalism, then, is the view that the meanings of the moral terms are given by their place in this network of input, output, and internal clauses that makes up folk morality.9

And thus, Jackson tosses his lot into the definitions debate. Jackson supposes that we can pick out *which* platitudes of moral discourse matter, and how *much* they matter, for determining the meaning of moral terms - despite the fact that individual humans, and especially groups of humans, are *themselves* confused about the meanings of moral terms, and which platitudes of moral discourse should 'matter' in fixing their meaning.

This is a debate about definitions that will never end.

Austere Metaethics vs. Empathic Metaethics

In the next post, we'll dissolve standard moral debates the same way Albert and Barry should have dissolved their debate about sound.

But that is only the first step. It is important to *not stop* after sweeping away the confusions of mainstream moral philosophy to arrive at mere *correct answers*. We must stare directly into the heart of the problem and do the impossible.

Consider Alex, who wants to do the 'right' thing. But she doesn't know what 'right' means. Her question is: "How do I do what is right if I don't know exactly what 'right' means?"

The Austere Metaethicist might cross his arms and say:

> Tell me what you mean by 'right', and I will tell you what is the right thing to do. If by 'right' you mean X, then Y is the right thing to do. If by 'right' you mean P, then Z is the right thing to do. But if you can't tell me what you mean by 'right', then you have failed to ask a coherent question, and no one can answer an incoherent question.

The Empathic Metaethicist takes up a greater burden. The Empathic Metaethicist says to Alex:

> You may not know what you mean by 'right.' You haven't asked a coherent question. But let's not stop there. Here, let me come alongside you and help decode the cognitive algorithms that generated your question in the first place, and then we'll be able to answer your question. Then not only can we tell you what the right thing to do is, but also we can help bring your *emotions* into alignment with that truth... as you go on to (say) help save the world rather than being filled with pointless existential angst about the universe being made of math.

Austere metaethics is easy. Empathic metaethics is hard. But empathic metaethics is what needs to be done to answer Alex's question, and it's what needs to be done to build a Friendly AI. We'll get there in the next few posts.

Next post: Pluralistic Moral Reductionism

Previous post: What is Metaethics?

Notes

1 Eliezer advises against reading mainstream philosophy because he thinks it will "teach very bad habits of thought that will lead people to be unable to do real work." Conceptual analysis is, I think, exactly that: a very bad habit of thought that renders many people unable to do real work. Also: My thanks to Eliezer for his helpful comments on an early draft of this post.

2 For example: Jackson (1998), p. 28, has a different view of conceptual analysis: "conceptual analysis is the very business of addressing when and whether a story told in one vocabulary is made true by one told in some allegedly more fundamental vocabulary." For an overview of Jackson's kind of conceptual analysis, see here. Also, Alonzo Fyfe reminded me that those who interpret the law must do a kind of conceptual analysis. If a law has been passed declaring that vehicles are not allowed on playgrounds, a judge must figure out whether 'vehicle' includes or excludes rollerskates. More recent papers on conceptual analysis are available at Philpapers. Finally, read Chalmers on verbal disputes.

3 Braddon-Mitchell (2008). A famous example of the first kind lies at the heart of 20th century epistemology: the definition of 'knowledge.' Knowledge had long been defined as 'justified true belief', but then Gettier (1963) presented some hypothetical examples of justified true belief that many of us would *intuitively* not label as 'knowledge.' Philosophers launched a cottage industry around new definitions of 'knowledge' and new counterexamples to *those* definitions. Brian Weatherson called this the "analysis of knowledge merry-go-round." Tyrrell McAllister called it the 'Gettier rabbit-hole.'

4 Schroeder (2004), pp. 15-27. Schroeder lists them as 7 objections, but I count his 'trying without desiring' and 'intending without desiring' objections separately.

5 Tabooing one's words is similar to what Chalmers (2009) calls the 'method of elimination'. In an earlier post, Yudkowsky used what Chalmers (2009) calls the 'subscript gambit', except Yudkowsky used underscores instead of subscripts.

6 See also Gallie (1956).

7 Eliezer said that the closest thing to his metaethics from mainstream philosophy is Jackson's 'moral functionalism', but of course moral functionalism is not quite right.

8 Jackson (1998), p. 118.

9 Jackson (1998), pp. 130-131.

References

Braddon-Mitchell (2008). Naturalistic analysis and the a priori. In Braddon-Mitchell & Nola (eds.), *Conceptual Analysis and Philosophical Naturalism* (pp. 23-43). MIT Press.

Chalmers (2009). Verbal disputes. Unpublished.

Gallie (1956). Essentially contested concepts. *Proceedings of the Aristotelean Society, 56*: 167-198.

Gettier (1963). Is justified true belief knowledge? *Analysis, 23*: 121-123.

Jackson (1998). *From Metaphysics to Ethics: A Defense of Conceptual Analysis*. Oxford University Press.

Schroeder (2004). *Three Faces of Desire*. Oxford University Press.

Pluralistic Moral Reductionism

Part of the sequence: No-Nonsense Metaethics

> Disputes over the definition of morality... are disputes over words which raise no really significant issues. [Of course,] lack of clarity about the meaning of words is an important source of error… My complaint is that what should be regarded as something to be got out of the way in the introduction to a work of moral philosophy has become the subject matter of almost the whole of moral philosophy...
>
> Peter Singer

If a tree falls in the forest, and no one hears it, does it make a sound? If by 'sound' you mean 'acoustic vibrations in the air', the answer is 'Yes.' But if by 'sound' you mean an auditory experience in the brain, the answer is 'No.'

We might call this straightforward solution *pluralistic sound reductionism*. If people use the word 'sound' to mean different things, and people have different intuitions about the meaning of the word 'sound', then we needn't endlessly debate which definition is 'correct'.1 We can be pluralists about the meanings of 'sound'.

To facilitate communication, we can taboo and reduce: we can replace the symbol with the substance and talk about facts and anticipations, not definitions. We can avoid using the word 'sound' and instead talk about 'acoustic vibrations' or 'auditory brain experiences.'

Still, some definitions can be *wrong*:

> Alex: If a tree falls in the forest, and no one hears it, does it make a sound?

Austere MetaAcousticist: Tell me what you mean by 'sound', and I will tell you the answer.

Alex: By 'sound' I mean 'acoustic messenger fairies flying through the ether'.

Austere MetaAcousticist: There's no such thing. Now, if you had asked me about this other definition of 'sound'...

There are other ways for words to be wrong, too. But once we admit to multiple potentially useful reductions of 'sound', it is not hard to see how we could admit to multiple useful reductions of *moral* terms.

Many Moral Reductionisms

Moral terms are used in a greater variety of ways than sound terms are. There is little hope of arriving at the One True Theory of Morality by analyzing common usage or by triangulating from the platitudes of folk moral discourse. But we can use stipulation, and we can taboo and reduce. We can use *pluralistic moral reductionism*2 (for austere metaethics, not for empathic metaethics).

Example #1:

> Neuroscientist Sam Harris: Which is better? Religious totalitarianism or the Northern European welfare state?

Austere Metaethicist: What do you mean by 'better'?

Harris: By 'better' I mean 'that which tends to maximize the well-being of conscious creatures'.

Austere Metaethicist: Assuming we have similar reductions of 'well-being' and 'conscious creatures' in mind, the evidence I know of suggests that the Northern European welfare state is more likely to maximize the well-being of conscious creatures than religious totalitarianism.

Example #2:

> Philosopher Peter Railton: Is capitalism the best economic system?
>
> Austere Metaethicist: What do you mean by 'best'?
>
> Railton: By 'best' I mean 'would be approved of by an ideally instrumentally rational and fully informed agent considering the question ‘How best to maximize the amount of non-moral goodness?' from a social point of view in which the interests of all potentially affected individuals are counted equally.
>
> Austere Metaethicist: Assuming we agree on the meaning of 'ideally instrumentally rational' and 'fully informed' and 'agent' and 'non-moral goodness' and a few other things, the evidence I know of suggests that capitalism would not be approved of by an ideally instrumentally rational and fully informed agent considering the question ‘How best to maximize the amount of non-moral goodness?' from a social point of view in which the interests of all potentially affected individuals were counted equally.

Example #3:

> Theologian Bill Craig: Ought we to give 50% of our income to efficient charities?
>
> Austere Metaethicist: What do you mean by 'ought'?
>
> Craig: By 'ought' I mean 'approved of by an essentially just and loving God'.
>
> Austere Metaethicist: Your definition doesn't connect to reality. It's like talking about atom-for-atom 'indexical identity' even though the world is made of configurations and amplitudes instead of Newtonian billiard balls. Gods don't exist.

But before we get to empathic metaethics, let's examine the standard problems of metaethics using the framework of pluralistic moral reductionism.

Cognitivism vs. Noncognitivism

One standard debate in metaethics is cognitivism vs. noncognitivism. Alexander Miller explains:

> Consider a particular moral judgement, such as the judgement that murder is wrong. What sort of psychological state does this express? Some philosophers,

called cognitivists, think that a moral judgement such as this expresses a belief.

Beliefs can be true or false: they are truth-apt, or apt to be assessed in terms of truth and falsity. So cognitivists think that moral judgements are capable of being true or false.

On the other hand, non-cognitivists think that moral judgements express non-cognitive states such as emotions or desires. Desires and emotions are not truth-apt. So moral judgements are not capable of being true or false.3

But why should we expect all people to use moral judgments like "Stealing is wrong" to express the same thing?4

Some people who say "Stealing is wrong" are *really* just trying to express emotions: "Stealing? Yuck!" Others use moral judgments like "Stealing is wrong" to express commands: "Don't steal!" Still others use moral judgments like "Stealing is wrong" to assert factual claims, such as "stealing is against the will of God" or "stealing is a practice that usually adds pain rather than pleasure to the world."

It may be interesting to study all such uses of moral discourse, but this post focuses on addressing *cognitivists*, who use moral judgments to assert factual claims. We ask: Are those claims true or false? What are their implications?

Objective vs. Subjective Morality

Is morality objective or subjective? It depends which moral reductionism you have in mind, and what you mean by 'objective' and 'subjective'.

Here are some common5 uses of the objective/subjective distinction in ethics:

- Moral facts are objective1 if they are made true or false by mind-independent facts, otherwise they are subjective1.
- Moral facts are objective2 if they are made true or false by facts independent of the opinions of sentient beings, otherwise they are subjective2.
- Moral facts are objective3 if they are made true or false by facts independent of the opinions of humans, otherwise they are subjective3.

Now, consider Harris' reduction of morality to facts about the well-being of conscious creatures. His theory of morality is objective3 and objective2, because facts about well-being are independent of anyone's opinion. Even if the Nazis had won WWII and brainwashed everybody to have the opinion that torturing Jews was moral, it would remain true that torturing Jews does not increase the average well-being of conscious creatures. But Harris' theory of morality is not objective1, because facts about the well-being of conscious creatures are mind-*dependent* facts.

Or, consider Craig's theory of morality in terms of divine approval. His theory doesn't connect to reality, but still: is it objective or subjective? Craig's theory says that moral facts are objective3, because they don't depend on human opinion (God isn't human). But his theory *doesn't* say that morality is objective2 or objective1, because for him, moral facts depend on the opinion of a sentient being: God.

A warning: ambiguous terms like 'objective' and 'subjective' are attractors for sneaking in connotations. Craig himself provides an example. In his writings and

public appearances, Craig insists that only God-based morality can be objective.6 What does he mean by 'objective'? On a single page,7 he uses 'objective' to mean "independent of *people*'*s* opinions" (objective2) and also to mean "independent of *human* opinion" (objective3). I'll assume he means that only God-based morality can be objective3, because God-based morality is clearly not objective2 (Craig's God is a *person*, a sentient being).

And yet, Craig says that we need God in order to have objective3 morality as if this should be a *big deal*. But hold on. Even a moral code defined in terms of the preferences of Washoe the chimpanzee is objective3. So not only is Bill's claim that only God-based morality can be objective3 *false* (because Harris' moral theory is also objective3), but also it's trivially easy to come up with a moral theory that is 'objective' in *Craig*'*s* (apparent) sense of the term (that is, objective3).8

Moreover, Harris' theory of morality is objective in a 'stronger' sense than Craig's theory of morality is. Harris' theory is objective3 and objective2, while Craig's theory is merely objective3. Whether he's doing it consciously or not, I wonder if Craig is using the word 'objective' to try to sneak in connotations that don't actually apply to his claims once you pay attention to what Craig *actually* means by the word 'objective'. If Craig told his audience that we need God for morality to be 'objective' in the same sense that morality defined in terms of the preferences of a chimpanzee is 'objective', would this still still have his desired effect on his audience? I doubt it.

Once you've stipulated your use of 'objective' and 'subjective', it is often trivial to determine whether a given moral reductionism is 'objective' or 'subjective'. But what of it? What force should those words carry after you've tabooed them? Be careful not to sneak in connotations that don't belong.

Relative vs. Absolute Morality

Is morality relative or absolute? Again, it depends which moral reductionism you have in mind, and what you mean by 'relative' and 'absolute'. Again, we must be careful about sneaking in connotations.

Moore's Open Question Argument

"He's an unmarried man, but is he a bachelor?" This is a 'closed' question. The answer is obviously "Yes."

In contrast, said G.E. Moore, all questions of the type "Such and such is X, but is it good?" are *open* questions. It feels like you can always ask, "Yes, but is it good?" In this way, Moore resists the identification of 'morally good' with any set of natural facts. This is Moore's Open Question Argument. Because some moral reductionisms *do* identify 'good' or 'right' with a particular X, those reductionisms had better have an answer to Moore.

The Yudkowskian response is to point out that when cognitivists use the term 'good', their intuitive notion of 'good' is captured by a massive logical function that can't be expressed in simple statements like "maximize pleasure" or "act only in accordance with maxims you could wish to be a universal law without contradiction." Even if you *think* everything you want (or rather, *want* to want) can be realized by (say) maximizing the well-being of conscious creatures, you're wrong. Your values are more

complex than that, and we can't see the structure of our values. *That* is why it feels like an open question remains no matter which simplistic identification of "Good = X" you choose.

The problem is not that there *is* no way to identify 'good' or 'right' (as used intuitively, without tabooing) with a certain X. The problem is that X is huge and complicated and we don't (yet) have access to its structure.

But that's the response to Moore *after* righting a wrong question - that is, when doing empathic metaethics. When doing mere *pluralistic moral reductionism*, Moore's argument doesn't apply. If we taboo and reduce, then the question of "...but is it good?" is out of place. The reply is: "Yes it is, because I just told you *that's what I mean to communicate* when I use the word-tool 'good' for this discussion. I'm not here to debate definitions; I'm here to get something done."9

The Is-Ought Gap

(This section rewritten for clarity.)

Many claim that you cannot infer an 'ought' statement from a series of 'is' statements. The objection comes from Hume, who said he was surprised whenever an argument made of *is* and *is not* propositions suddenly shifted to an *ought* or *ought not* claim, without explanation.10

The solution is to make explicit the bridge from 'ought' statements to 'is' statements.

Perhaps the arguer means something non-natural by 'ought', such as 'commanded by God' or 'in accord with irreducible, non-natural facts about goodness' (see Moore). If so, I would reject that premise of the argument, because I'm a reductionist. At this point, our discussion might need to shift to a debate over the merits of reductionism.

Or perhaps by 'you ought to X' the arguer means something fully natural, such as:

- "X is obligatory (by deontic logic) if you assume axiomatic imperatives Y and Z."
- Or: "X tends to maximizes reward signals in agents exhibiting multiple-drafts consciousness" (or, as Sam Harris more broadly puts it, "X tends to maximize well-being in conscious creatures").
- Or: "X is what a Bayes-rational and Hubble-volume-omniscient agent would do if it was motivated to maximize the amount of non-moral goodness from a view in which the interests of all potentially affected individuals were counted equally, where 'non-moral goodness' refers to what an agent would want if it were he to contemplate its present situation from a standpoint fully and vividly informed about itself and its circumstances, and entirely free of cognitive error or lapses of instrumental rationality" (see Railton's metaethics).
- Or: "X maximizes the complicated function that can be computed by extrapolating (in a particular way) the motivations encoded by my brain" (see CEV).
- Or: "[insert here whatever statement, if believed, would motivate one to do X]" (see Will Sawin).

Or, the speaker may have in mind a common ought-reductionism known as the *hypothetical imperative*. This is an ought of the kind: "*If* you desire to lose weight,

then you *ought* to consume fewer calories than your burn." (But usually, people leave off the implied *if* statement, and simply say "You should eat less and exercise more.")

A hypothetical imperative (as some use it) can be translated from 'ought' to 'is' in a straightforward way: "If you desire to lose weight, then you *ought* to consume fewer calories than you burn" translates to the claim "If you consume fewer calories than you burn, then you *will* (or are, *ceteris paribus*, more likely to) fulfill your desire to lose weight."11

Or, the speaker may be using 'ought' to communicate something only about other symbols (example: Bayes' Rule), leaving the bridge from 'ought' to 'is' to be built when the logical function represented by his use of 'ought' is plugged into a theory that refers to the world.

But one must not fall into the trap of thinking that a definition you've stipulated (aloud or in your head) for 'ought' must match up to your intended meaning of 'ought' (to which you don't have introspective access). In fact, I suspect it *never* does, which is why the conceptual analysis of 'ought' language can go in circles for centuries, and why any stipulated meaning of 'ought' is a fake utility function. To see clearly to our intuitive concept of ought, we'll have to try empathic metaethics (see below).

But whatever our intended meaning of 'ought' is, the same reasoning applies. Either our intended meaning of 'ought' refers (eventually) to the world of math and physics (in which case the is-ought gap is bridged), or else it doesn't (in which case it fails to refer).12

Moral realism vs. Anti-realism

So, does all this mean that we can embrace moral realism, or does it doom us to moral anti-realism? Again, it depends on what you mean by 'realism' and 'anti-realism'.

In a sense, pluralistic moral reductionism can be considered a robust form of moral 'realism', in the same way that pluralistic sound reductionism is a robust form of *sound realism*. "Yes, there really *is* sound, and we can locate it in reality — either as vibrations in the air or as mental auditory experiences, however you are using the term." In the same way: "Yes, there really *is* morality, and we can locate it in reality — either as a set of facts about the well-being of conscious creatures, or as a set of facts about what an ideally rational and perfectly informed agent would prefer, or as some other set of natural facts."

But in another sense, pluralistic moral reductionism is 'anti-realist'. It suggests that there is no One True Theory of Morality. (We use moral terms in a variety of ways, and some of those ways refer to different sets of natural facts.) And as a reductionist approach to morality, it might also leave no room for moral theories which say there are *universally binding* moral rules for which the universe (e.g. via a God) will hold us accountable.

What matters are the facts, not whether labels like 'realism' or 'anti-realism' apply to 'morality'.

Toward Empathic Metaethics

But pluralistic moral reductionism satisfies only a would-be austere metaethicist, not an empathic metaethicist.

Recall that when Alex asks how she can do what is right, the Austere Metaethicist replies:

> Tell me what you mean by 'right', and I will tell you what is the right thing to do. If by 'right' you mean X, then Y is the right thing to do. If by 'right' you mean P, then Z is the right thing to do. But if you can't tell me what you mean by 'right', then you have failed to ask a coherent question, and no one can answer an incoherent question.

Alex may reply to the Austere Metaethicist:

> Okay, I'm not sure exactly what I mean by 'right'. So how do I do what is right if I'm not sure what I mean by 'right'?

The Austere Metaethicist refuses to answer this question. The Empathic Metaethicist, however, is willing to go the extra mile. He says to Alex:

> You may not know what you mean by 'right.' But let's not stop there. Here, let me come alongside you and help decode the cognitive algorithms that generated your question in the first place, and then we'll be able to answer your question. Then we can tell you what the right thing to do is.

This may seem like too much work. Would we be motivated to decode the cognitive algorithms producing Albert and Barry's use of the word 'sound'? Would we try to solve 'empathic meta-acoustics'? Probably not. We can simply taboo and reduce 'sound' and then get some work done.

But moral terms and value terms are about what we *want*. And unfortunately, we often don't *know* what we want. As such, we're unlikely to get what we *really* want if the world is re-engineered in accordance with our current best *guess* as to what we want. That's why we need to decode the cognitive algorithms that generate our questions about value and morality.

So how can the Empathic Metaethicist answer Alex's question? We don't know the details yet. For example, we don't have a completed cognitive neuroscience. But we have some ideas, and we know of some open problems that may admit of progress once more people understand them. In the next few posts, we'll take our first look at empathic metaethics.13

Previous post: Conceptual Analysis and Moral Theory

Notes

1 Some have objected that the conceptual analysis argued against in Conceptual Analysis and Moral Theory is not just a battle over definitions. But a definition is "the formal statement of the meaning or significance of a word, phrase, etc.", and a conceptual analysis is (usually) a "formal statement of the meaning or significance of a word, phrase, etc." in terms of necessary and sufficient conditions. The goal of a conceptual analysis is to arrive at a definition for a term that captures our intuitions about its meaning. The process is to bash our intuitions against others' intuitions until we converge upon a set of necessary and sufficient conditions that captures them all.

But consider Barry and Albert's debate over the definition of 'sound'. Why think Albert and Barry have the same concept in mind? Words mean slightly different things in different cultures, subcultures, and small communities. We develop different intuitions about their meaning based on divergent life experiences. Our intuitions differ from each other's due to the specifics of unconscious associative learning and attribution substitution heuristics. What is the point of bashing our intuitions about the meaning of terms against each other for thousands of pages, in the hopes that we'll converge on a precise set of necessary and sufficient conditions? Even if we can get Albert and Barry to agree, what happens when Susan wants to use the same term, but has slightly differing intuitions about its meaning? And, let's say we arrive at a messy set of 6 necessary and sufficient conditions for the intuitive meaning of the term. Is that going to be as useful for communication as one we consciously chose because it carved-up thingspace well? I doubt it. The IAU's definition of 'planet' is more useful than the folk-intuitions definition of 'planet'. Folk intuitions about 'planet' evolved over thousands of years and different people have different intuitions which may not always converge. In 2006, the IAU used modern astronomical knowledge to carve up thingspace in a more useful and informed way than our intuitions do.

A passage from Bertrand Russell (1953) is appropriate. Russell said that many philosophers reminded him of

> the shopkeeper of whom I once asked the shortest way to Winchester. He called to a man in the back premises:

"Gentleman wants to know the shortest way to Winchester."

"Winchester?" an unseen voice replied.

"Aye."

"Way to Winchester?"

"Aye."

"Shortest way?"

"Aye."

"Dunno."

He wanted to get the nature of the question clear, but took no interest in answering it. This is exactly what modern philosophy does for the earnest seeker after truth. Is it surprising that young people turn to other studies?

2 Compare also to the biologist's 'species concept pluralism' and the philosopher's 'art concept pluralism.' See Uidhir & Magnus (2011). Also see 'causal pluralism' (Godfrey-Smith, 2009; Cartwright, 2007), 'theory concept pluralism' (Magnus, 2009) and, especially, 'metaethical contextualism' (Bjornsson & Finlay, 2010) or 'metaethical pluralism' or 'metaethical ambivalence' (Joyce, 2011). Joyce quotes Lewis (1989), who wrote that some concepts of value refer to things that really exist, and some concepts don't, and what you make of this situation is largely a matter of temperament:

> What to make of the situation is mainly a matter of temperament. You can bang the drum about how philosophy has uncovered a terrible secret: there are no values! ... Or you can think it better for public safety to keep quiet and hope

> people will go on as before. Or you can declare that there are no values, but that nevertheless it is legitimate—and not just expedient—for us to carry on with value-talk, since we can make it all go smoothly if we just give the name of value to claimants that don't quite deserve it... Or you can think it an empty question whether there are values: say what you please, speak strictly or loosely. When it comes to deserving a name, there's better and worse but who's to say how good is good enough? Or you can think it clear that the imperfect deservers of the name are good enough, but only just, and say that although there are values we are still terribly wrong about them. Or you can calmly say that value (like simultaneity) is not quite as some of us sometimes thought. Myself, I prefer the calm and conservative responses. But as far as the analysis of value goes, they're all much of a muchness.

Joyce concludes that, for example, the moral naturalist and the moral error theorist may agree with each other (when adopting each other's own language):

> [Metaethical ambivalence] begins with a kind of metametaethical enlightenment. The moral naturalist espouses moral naturalism, but this espousal reflects a mature decision, by which I mean that the moral naturalist doesn't claim to have latched on to an incontrovertiblerealm of moral facts of which the skeptic is foolishly ignorant, but rather acknowledges that this moral naturalism has been achieved only via a non-mandatory piece of conceptual precisification. Likewise, the moral skeptic champions moral skepticism, but this too is a sophisticated verdict: not the simple declaration that there are no moral values and that the naturalist is gullibly uncritical, but rather a decision that recognizes that this skepticism has been earned only by making certain non-obligatory but permissible conceptual clarifications.

...The enlightened moral naturalist doesn't merely (grudgingly) admit that the skeptic is warranted in his or her views, but is able to adopt the skeptical position in order to gain the insights that come from recognizing that we live in a world without values. And the enlightened moral skeptic goes beyond (grudgingly) conceding that moral naturalism is reasonable, but is capable of assuming that perspective in order to gain whatever benefits come from enjoying epistemic access to a realm of moral facts.

3 Miller (2003), p. 3.

4 I changed the example moral judgment from "murder is wrong" to "stealing is wrong" because the former invites confusion. 'Murder' often means *wrongful* killing.

5 Also see Jacobs (2002), starting on p. 2.

6 The first premise of one of his favorite arguments for God's existence is "If God does not exist, objective moral values and duties do not exist."

7 Craig (2010), p. 11.

8 It's also possible that Craig intended a different sense of objective than the ones explicitly given in his article. Perhaps he meant objective4: "morality is objective4 if it is not grounded in the opinion of non-divine persons."

9 Also see Moral Reductionism and Moore's Open Question Argument.

10 Hume (1739), p. 469. The famous paragraph is:

> In every system of morality, which I have hitherto met with, I have always remarked, that the author proceeds for some time in the ordinary ways of reasoning, and establishes the being of a God, or makes observations concerning human affairs; when all of a sudden I am surprised to find, that instead of the usual copulations of propositions, is, and is not, I meet with no proposition that is not connected with an *ought*, or an ought not. This change is imperceptible; but is however, of the last consequence. For as this ought, or ought not, expresses some new relation or affirmation, 'tis necessary that it should be observed and explained; and at the same time that a reason should be given; for what seems altogether inconceivable, how this new relation can be a deduction from others, which are entirely different from it.

11 For more on reducing certain kinds of normative statements, see Finlay (2010).

12 Assuming reductionism is true. If reductionism is false, then of course there are problems for pluralistic moral reductionism as a theory of austere (but not empathic) metaethics. The clarifications in the last three paragraphs of this section are due to discussions with Wei Dai and Vladimir Nesov.

13 My thanks to Steve Rayhawk and Will Newsome for their feedback on early drafts of this post.

References

Bjornsson & Finlay (2010). Metaethical contextualism defended. *Ethics, 121*: 7-36.

Craig (2010). Five Arguments for God. The Gospel Coalition.

Cartwright (2007).*Hunting Causes and Using Them: Approaches in Philosophy and Economics*. Cambridge University Press.

Godfrey-Smith (2009). Causal pluralism. In Beebee, Hitchcock, & Menzies (eds.), *The Oxford Handbook of Causation* (pp. 326-337). Oxford University Press.

Hume (1739). *A Treatise on Human Nature*. John Noon.

Finlay (2010). Normativity, Necessity and Tense: A Recipe for Homebaked Normativity. In Shafer-Landau (ed.), *Oxford Studies in Metaethics 5* (pp. 57-85). Oxford University Press.

Jacobs (2002). *Dimensions of Moral Theory*. Wiley-Blackwell.

Joyce (2011).Metaethical pluralism: How both moral naturalism and moral skepticism may be permissible positions. In Nuccetelli & Seay (eds.), *Ethical Naturalism: Current Debates*. Cambridge University Press.

Lewis (1989). Dispositional theories of value. Part II. *Proceedings of the Aristotelian Society, supplementary vol. 63*: 113-137.

Magnus (2009). What species can teach us about theory.

Miller (2003). *An Introduction to Contemporary Metaethics*. Polity.

Russell (1953). The cult of common usage. *British Journal for the Philosophy of Science, 12*: 305-306.

Uidhir & Magnus (2011). Art concept pluralism. *Metaphilosophy, 42*: 83-97.

Quick thoughts on empathic metaethics

Years ago, I wrote an unfinished sequence of posts called "No-Nonsense Metaethics." My last post, Pluralistic Moral Reductionism, said I would next explore "empathic metaethics," but I never got around to writing those posts. Recently, I wrote a high-level summary of some initial thoughts on "empathic metaethics" in section 6.1.2 of a report prepared for my employer, the Open Philanthropy Project. With my employer's permission, I've adapted that section for publication here, so that it can serve as the long-overdue concluding post in my sequence on metaethics.

In my previous post, I distinguished "austere metaethics" and "empathic metaethics," where austere metaethics confronts moral questions roughly like this:

> Tell me what you mean by 'right', and I will tell you what is the right thing to do. If by 'right' you mean X, then Y is the right thing to do. If by 'right' you mean P, then Z is the right thing to do. But if you can't tell me what you mean by 'right', then you have failed to ask a coherent question, and no one can answer an incoherent question.

Meanwhile, empathic metaethics says instead:

> You may not know what you mean by 'right.' But let's not stop there. Here, let me come alongside you and help decode the cognitive algorithms that generated your question in the first place, and then we'll be able to answer your question. Then we can tell you what the right thing to do is.

Below, I provide a high-level summary of some of my initial thoughts on what one approach to "empathic metaethics" could look like.

Given my metaethical approach, when I make a “moral judgment” about something (e.g. about which kinds of beings are moral patients), I don’t conceive of myself as perceiving an objective moral truth, or coming to know an objective moral truth via a series of arguments. Nor do I conceive of myself as merely expressing my moral feelings as they stand today. Rather, I conceive of myself as making a conditional forecast about what my values would be if I underwent a certain “idealization” or “extrapolation” procedure (coming to know more true facts, having more time to consider moral arguments, etc.).[1]

Thus, in a (hypothetical) "extreme effort" attempt to engage in empathic metaethics (for thinking about *my own* moral judgments), I would do something like the following:

1. I would try to make the scenario I'm aiming to forecast as concrete as possible, so that my brain is able to treat it as a genuine forecasting challenge, akin to participating in a prediction market or forecasting tournament, rather than as a fantasy about which my brain feels "allowed" to make up whatever story feels nice, or signals my values to others, or achieves something else that isn't *forecasting accuracy*.[2] In my case, I concretize the extrapolation procedure as one involving a large population of copies of me who learn many true facts, consider many moral arguments, and undergo various other experiences, and then collectively advise me about what I should value and why.[3]

2. However, I would also try to make forecasts I can actually check for accuracy, e.g. about what my moral judgment about various cases will be 2 months in the future.
3. When making these forecasts, I would try to draw on the best research I've seen concerning how to make accurate estimates and forecasts. For example I would try to "think like a fox, not like a hedgehog," and I've already done several hours of probability calibration training, and some amount of forecasting training.[4]
4. Clearly, my current moral intuitions would serve as one important source of evidence about what my extrapolated values might be. However, recent findings in moral psychology and related fields lead me to assign more evidential weight to some moral intuitions than to others. More generally, I interpret my current moral intuitions as data generated partly by my moral principles, and partly by various "error processes" (e.g. a hard-wired disgust reaction to spiders, which I don't endorse upon reflection). Doing so allows me to make use of some standard lessons from statistical curve-fitting when thinking about how much evidential weight to assign to particular moral intuitions.[5]
5. As part of forecasting what my extrapolated values might be, I would try to consider different processes and contexts that could generate alternate moral intuitions in moral reasoners both similar and dissimilar to my current self, and I would try to consider how I feel about the the "legitimacy" of those mechanisms as producers of moral intuitions. For example I might ask myself questions such as "How might I feel about that practice if I was born into a world in which it was already commonplace?" and "How might I feel about that case if my built-in (and largely unconscious) processes for associative learning and imitative learning had been exposed to different life histories than my own?" and "How might I feel about that case if I had been born in a different century, or a different country, or with a greater propensity for clinical depression?" and "How might a moral reasoner on another planet feel about that case if it belonged to a more strongly r-selected species (compared to humans) but had roughly human-like general reasoning ability?"[6]
6. Observable patterns in how people's values change (seemingly) in response to components of my proposed extrapolation procedure (learning more facts, considering moral arguments, etc.) would serve as another source of evidence about what my extrapolated values might be. For example, the correlation between aggregate human knowledge and our "expanding circle of moral concern" (Singer 2011) might (very weakly) suggest that, if I continued to learn more true facts, my circle of moral concern would continue to expand. Unfortunately, such correlations are badly confounded, and might not provide much evidence.[7]
7. Personal facts about how my own values have evolved as I've learned more, considered moral arguments, and so on, would serve as yet another source of evidence about what my extrapolated values might be. Of course, these relations are likely confounded as well, and need to be interpreted with care.[8]

---

1. This general approach sometimes goes by names such as "ideal advisor theory" or, arguably, "reflective equilibrium." Diverse sources explicating various extrapolation procedures (or fragments of extrapolation procedures) include: Rosati (1995); Daniels (2016); Campbell (2013); chapter 9 of Miller (2013); Muehlhauser & Williamson (2013); Trout (2014); Yudkowsky's "Extrapolated volition (normative moral theory)" (2016); Baker (2016); Stanovich (2004), pp. 224-275; Stanovich (2013).

2. For more on forecasting accuracy, see this blog post. My use of research on the psychological predictors of forecasting accuracy for the purposes of doing moral

philosophy is one example of my support for the use of "ameliorative psychology" in philosophical practice — see e.g. Bishop & Trout (2004, 2008).

3. Specifically, the scenario I try to imagine (and make conditional forecasts about) looks something like this:

1. In the distant future, I am non-destructively "uploaded." In other words, my brain and some supporting cells are scanned (non-destructively) at a fine enough spatial and chemical resolution that, when this scan is combined with accurate models of how different cell types carry out their information-processing functions, one can create an executable computer model of my brain that matches my biological brain's input-output behavior almost exactly. This whole brain emulation ("em") is then connected to a virtual world: computed inputs are fed to the em's (now virtual) signal transduction neurons for sight, sound, etc., and computed outputs from the em's virtual arm movements, speech, etc. are received by the virtual world, which computes appropriate changes to the virtual world in response. (I don't think anything remotely like this will ever happen, but as far as I know it is a *physically possible* world that can be described in some detail; for one attempt, see Hanson 2016.) Given functionalism, this "em" has the same memories, personality, and conscious experience that I have, though it experiences quite a shock when it awakens to a virtual world that might look and feel somewhat different from the "real" world.
2. This initial em is copied thousands of times. Some of the copies interact inside the same virtual world, other copies are placed inside isolated virtual worlds.
3. Then, these ems spend a very long time (a) collecting and generating arguments and evidence about morality and related topics, (b) undergoing various experiences, in varying orders, and reflecting on those experiences, (c) dialoguing with ems sourced from other biological humans who have different values than I do, and perhaps with sophisticated chat-bots meant to simulate the plausible reasoning of other types of people (from the past, or from other worlds) who were not available to be uploaded, and so on. They are able to do these things for a very long time because they and their virtual worlds are run at speeds thousands of times faster than my biological brain runs, allowing subjective eons to pass in mere months of "objective" time.
4. Finally, at some time, the ems dialogue with each other about which values seem "best," they engage in moral trade (Ord 2015), and they try to explain to me what values they think I should have and why. In the end, I am not forced to accept any of the values they then hold (collectively or individually), but I am able to come to much better-informed moral judgments than I could have without their input.

For more context on this sort of values extrapolation procedure, see Muehlhauser & Williamson (2013).

4. For more on forecasting "best practices," see this blog post.

5. Following Hanson (2002) and ch. 2 of Beckstead (2013), I consider my moral intuitions in the context of Bayesian curve-fitting. To explain, I'll quote Beckstead (2013) at some length:

> Curve fitting is a problem frequently discussed in the philosophy of science. In the standard presentation, a scientist is given some data points, usually with an independent variable and a dependent variable, and is asked to predict the values of the dependent variable given other values of the independent variable.

> Typically, the data points are *observations*, such as "measured height" on a scale or "reported income" on a survey, rather than true values, such as height or income. Thus, in making predictions about additional data points, the scientist has to account for the possibility of error in the observations. By an error process I mean anything that makes the observed values of the data points differ from their true values. Error processes could arise from a faulty scale, failures of memory on the part of survey participants, bias on the part of the experimenter, or any number of other sources. While some treatments of this problem focus on predicting observations (such as measured height), I'm going to focus on predicting the true values (such as true height).
>
> …For any consistent data set, it is possible to construct a curve that fits the data exactly… If the scientist chooses one of these polynomial curves for predictive purposes, the result will usually be *overfitting*, and the scientist will make worse predictions than he would have if he had chosen a curve that did not fit the data as well, but had other virtues, such as a straight line. On the other hand, always going with the simplest curve and giving no weight to the data leads to *underfitting*…
>
> I intend to carry over our thinking about curve fitting in science to reflective equilibrium in moral philosophy, so I should note immediately that curve fitting is not limited to the case of two variables. When we must understand relationships between multiple variables, we can turn to multiple-dimensional spaces and fit planes (or hyperplanes) to our data points. Different axes might correspond to different considerations which seem relevant (such as total well-being, equality, number of people, fairness, etc.), and another axis could correspond to the value of the alternative, which we can assume is a function of the relevant considerations. Direct Bayesian updating on such data points would be impractical, but the philosophical issues will not be affected by these difficulties.
>
> …On a Bayesian approach to this problem, the scientist would consider a number of different hypotheses about the relationship between the two variables, including both hypotheses about the phenomena (the relationship between X and Y) and hypotheses about the error process (the relationship between observed values of Y and true values of Y) that produces the observations…
>
> …Lessons from the Bayesian approach to curve fitting apply to moral philosophy. Our moral intuitions are the data, and there are error processes that make our moral intuitions deviate from the truth. The complete moral theories under consideration are the hypotheses about the phenomena. (Here, I use "theory" broadly to include any complete set of possibilities about the moral truth. My use of the word "theory" does not assume that the truth about morality is simple, systematic, and neat rather than complex, circumstantial, and messy.) If we expect the error processes to be widespread and significant, we must rely on our priors more. If we expect the error processes to be, in addition, biased and correlated, then we will have to rely significantly on our priors even when we have a lot of intuitive data.

Beckstead then summarizes the framework with a table (p. 32), edited to fit into LessWrong's formatting:

- Hypotheses about phenomena
    - *(Science)* Different trajectories of a ball that has been dropped

    - *(Moral Philosophy)* Moral theories (specific versions of utilitarianism, Kantianism, contractualism, pluralistic deontology, etc.)
- Hypotheses about error processes
    - *(Science)* Our position measurements are accurate on average, and are within 1 inch 95% of the time (with normally distributed error)
    - *(Moral Philosophy)* Different hypotheses about the causes of error in historical cases; cognitive and moral biases; different hypotheses about the biases that cause inconsistent judgments in important philosophical cases
- Observations
    - *(Science)* Recorded position of a ball at different times recorded with a certain clock
    - *(Moral Philosophy)* Intuitions about particular cases or general principles, and any other relevant observations
- Background theory
    - *(Science)* The ball never bounces higher than the height it started at. The ball always moves along a continuous trajectory.
    - *(Moral Philosophy)* Meta-ethical or normative background theory (or theories)

6. For more on this, see my conversation with Carl Shulman, O'Neill (2015), the literature on the evolution of moral values (e.g. de Waal et al. 2014; Sinnott-Armstrong & Miller 2007; Joyce 2005), the literature on moral psychology more generally (e.g. Graham et al. 2013; Doris 2010; Liao 2016; Christen et al. 2014; Sunstein 2005), the literature on how moral values vary between cultures and eras (e.g. see Flanagan 2016; Inglehart & Welzel 2010; Pinker 2011; Morris 2015; Friedman 2005; Prinz 2007, pp. 187-195), and the literature on moral thought experiments (e.g. Tittle 2004, ch. 7). See also Wilson (2016)'s comments on internal and external validity in ethical thought experiments, and Bakker (2017) on "alien philosophy."

I do not read much fiction, but I suspect that some types of fiction — e.g. historical fiction, fantasy, and science fiction — can help readers to temporarily transport themselves into fully-realized alternate realities, in which readers can test how their moral intuitions differ when they are temporarily "lost" in an alternate world.

7. There are many sources which discuss how people's values seem to change along with (and perhaps in response to) components of my proposed extrapolation procedure, such as learning more facts, reasoning through more moral arguments, and dialoguing with others who have different values. See e.g. Inglehart & Welzel (2010), Pinker (2011), Shermer (2015), and Buchanan & Powell (2016). See also the literatures on "enlightened preferences" (Althaus 2003, chs. 4-6) and on "deliberative polling."

8. For example, as I've learned more, considered more moral arguments, and dialogued more with people who don't share my values, my moral values have become more "secular-rational" and "self-expressive" (Inglehart & Welzel 2010), more geographically global, more extensive (e.g. throughout more of the animal kingdom), less person-affecting, and subject to greater moral uncertainty (Bykvist 2017).