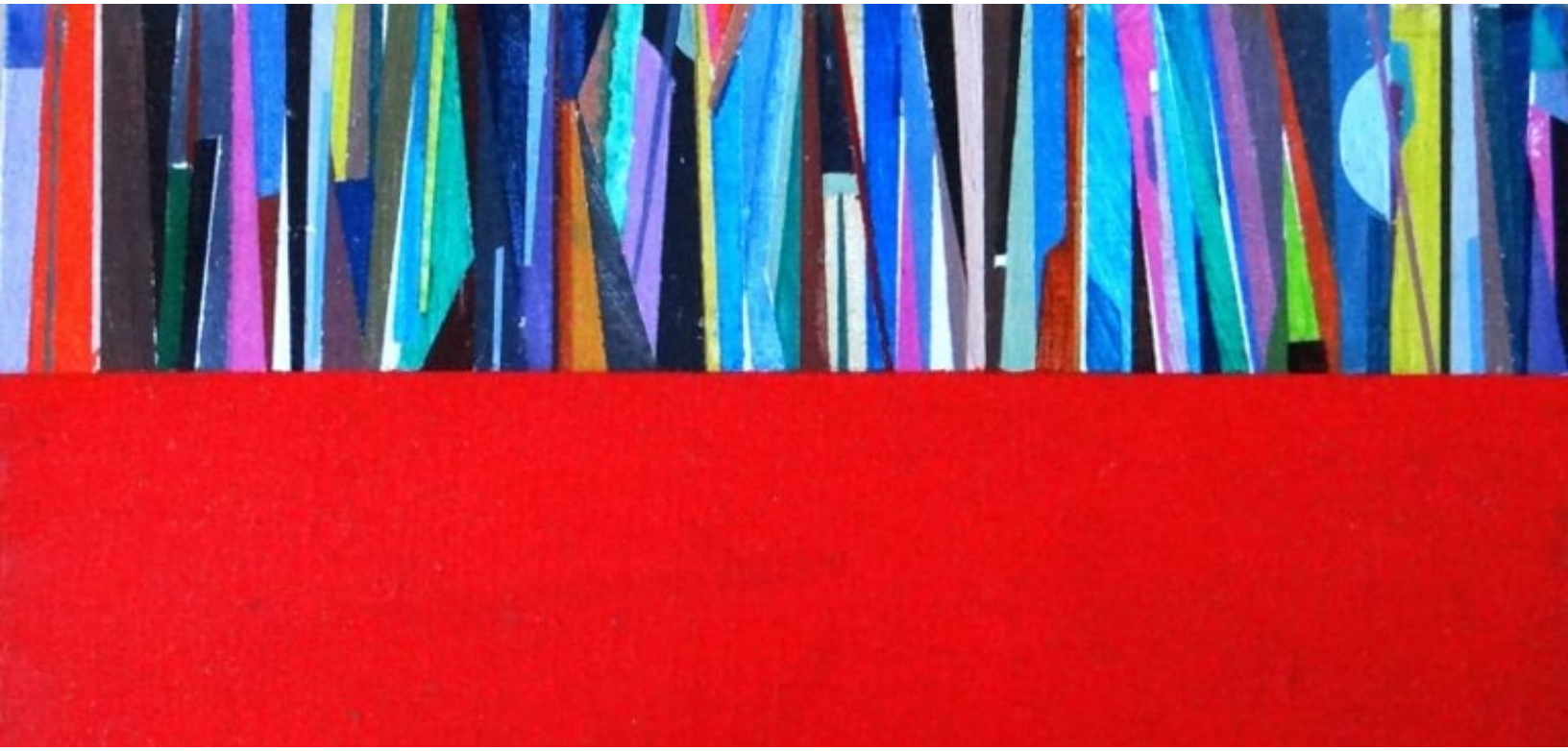

Metaethics

Heading Toward Morality

Followup to: Ghosts in the Machine, Fake Fake Utility Functions, Fake Utility Functions

As people were complaining before about not seeing where the quantum physics sequence was going, I shall go ahead and tell you where I'm heading now.

Having dissolved the confusion surrounding the word "could", the trajectory is now heading toward *should.*

In fact, I've been heading there for a while. Remember the whole sequence on fake utility functions? Back in... well... November 2007?

I sometimes think of there being a train that goes to the Friendly AI station; but it makes several stops before it gets there; and at each stop, a large fraction of the remaining passengers get off.

One of those stops is the one I spent a month leading up to in November 2007, the sequence chronicled in Fake Fake Utility Functions and concluded in Fake Utility Functions.

That's the stop where someone thinks of the One Great Moral Principle That Is All We Need To Give AIs.

To deliver that one warning, I had to go through all sorts of topics—which topics one might find useful even if not working on Friendly AI. I warned against Affective Death Spirals, which required recursing on the affect heuristic and halo effect, so that your good feeling about one particular moral principle wouldn't spiral out of control. I did that whole sequence on evolution; and discursed on the human ability to make almost any goal appear to support almost any policy; I went into evolutionary psychology to argue for why we shouldn't expect human terminal values to reduce to any simple principle, even happiness, explaining the concept of "expected utility" along the way...

...and talked about genies and more; but you can read the Fake Utility sequence for that.

So that's *just* the warning against trying to oversimplify human morality into One Great Moral Principle.

If you want to actually dissolve the confusion that surrounds the word "should"—which is the next stop on the train—then that takes a much longer introduction. Not just one November.

I went through the sequence on words and definitions so that I would be able to later say things like "The next project is to Taboo the word 'should' and replace it with its substance", or "Sorry, saying that morality is self-interest 'by definition' isn't going to cut it here".

And also the words-and-definitions sequence was the simplest example I knew to introduce the notion of How An Algorithm Feels From Inside, which is one of the great master keys to dissolving wrong questions. Though it seems to us that our cognitive representations are the very substance of the world, they have a character that

comes from cognition and often cuts crosswise to a universe made of quarks. E.g. probability; if we are uncertain of a phenomenon, that is a fact about our state of mind, not an intrinsic character of the phenomenon.

Then the reductionism sequence: that a universe made only of quarks, does not mean that things of value are lost or even degraded to mundanity. And the notion of how the sum can seem unlike the parts, and yet be as much the parts as our hands are fingers.

Followed by a new example, one step up in difficulty from words and their seemingly intrinsic meanings: "Free will" and seemingly intrinsic could-ness.

But before that point, it was useful to introduce quantum physics. Not just to get to timeless physics and dissolve the "determinism" part of the "free will" confusion. But also, more fundamentally, to break belief in an intuitive universe that looks just like our brain's cognitive representations. And present examples of the dissolution of even such fundamental intuitions as those concerning personal identity. And to illustrate the idea that you are within physics, within causality, and that strange things will go wrong in your mind if ever you forget it.

Lately we have begun to approach the final precautions, with warnings against such notions as Author* control: every mind which computes a morality must do so within a chain of lawful causality, it cannot arise from the free will of a ghost in the machine.

And the warning against Passing the Recursive Buck to some meta-morality that is not itself computably specified, or some meta-morality that is chosen by a ghost without it being programmed in, or to a notion of "moral truth" just as confusing as "should" itself...

And the warning on the difficulty of grasping slippery things like "should"—demonstrating how very easy it will be to just invent another black box equivalent to should-ness, to sweep should-ness under a slightly different rug—or to bounce off into mere modal logics of primitive should-ness...

We aren't yet at the point where I can explain morality.

But I think—though I could be mistaken—that we are finally getting close to the final sequence.

And if you don't care about my goal of explanatorily transforming Friendly AI from a Confusing Problem into a merely Extremely Difficult Problem, then stick around anyway. I tend to go through interesting intermediates along my way.

It might seem like confronting "the nature of morality" from the perspective of Friendly AI is only asking for additional trouble.

Artificial Intelligence melts people's brains. Metamorality melts people's brains. Trying to think about AI and metamorality at the same time can cause people's brains to spontaneously combust and burn for years, emitting toxic smoke—don't laugh, I've seen it happen multiple times.

But the discipline imposed by Artificial Intelligence is this: you cannot escape into things that are "self-evident" or "obvious". That doesn't stop people from trying, but the programs don't work. Every thought has to be computed somehow, by transistors made of mere quarks, and not by moral self-evidence to some ghost in the machine.

If what you care about is rescuing children from burning orphanages, I don't think you will find many moral surprises here; my metamorality adds up to moral normality, as it should. You do not need to worry about metamorality when you are *personally* trying to rescue children from a burning orphanage. The point at which metamoral issues *per se* have high stakes in the real world, is when you try to compute morality in an AI standing in front of a burning orphanage.

Yet there is also a good deal of needless despair and misguided fear of science, stemming from notions such as, "Science tells us the universe is empty of morality". This is damage done by a confused metamorality that fails to add up to moral normality. For that I hope to write down a counterspell of understanding. Existential depression has always annoyed me; it is one of the world's most pointless forms of suffering.

Don't expect the final post on this topic to come tomorrow, but at least you know where we're heading.

Part of *The Metaethics Sequence*

Next post: "No Universally Compelling Arguments"

(start of sequence)

No Universally Compelling Arguments

What is so *terrifying* about the idea that not every possible mind might agree with us, even in principle?

For some folks, nothing—it doesn't bother them in the slightest. And for some of *those* folks, the *reason* it doesn't bother them is that they don't have strong intuitions about standards and truths that go beyond personal whims. If they say the sky is blue, or that murder is wrong, that's just their personal opinion; and that someone else might have a different opinion doesn't surprise them.

For other folks, a disagreement that persists even *in principle* is something they can't accept. And for some of *those* folks, the *reason* it bothers them, is that it seems to them that if you allow that some people cannot be persuaded *even in principle* that the sky is blue, then you're conceding that "the sky is blue" is merely an *arbitrary* personal opinion.

Yesterday, I proposed that you should resist the temptation to generalize over all of mind design space. If we restrict ourselves to minds specifiable in a trillion bits or less, then each *universal* generalization "All minds m: X(m)" has two to the trillionth chances to be false, while each *existential* generalization "Exists mind m: X(m)" has two to the trillionth chances to be true.

This would seem to argue that for every argument A, howsoever convincing it may seem to us, there exists at least one possible mind that doesn't buy it.

And the surprise and/or horror of this prospect (for some) has a great deal to do, I think, with the intuition of the ghost-in-the-machine—a ghost with some irreducible core that any *truly valid* argument will convince.

I have previously spoken of the intuition whereby people map *programming a computer*, onto *instructing a human servant*, so that the computer might rebel against its code—or perhaps look over the code, decide it is not reasonable, and hand it back.

If there were a ghost in the machine and the ghost contained an irreducible core of reasonableness, above which any mere code was only a suggestion, then there might be universal arguments. Even if the ghost was initially handed code-suggestions that contradicted the Universal Argument, then when we finally did expose the ghost to the Universal Argument—or the ghost could discover the Universal Argument on its own, that's also a popular concept—the ghost would just override its own, mistaken source code.

But as the student programmer once said, "I get the feeling that the computer just skips over all the comments." The code is not given to the AI; the code *is* the AI.

If you switch to the physical perspective, then the notion of a Universal Argument seems noticeably unphysical. If there's a physical system that at time T, after being exposed to argument E, does X, then there ought to be another physical system that at time T, after being exposed to environment E, does Y. Any thought has to be implemented *somewhere*, in a physical system; any belief, any conclusion, any decision, any motor output. For every lawful causal system that zigs at a set of points, you should be able to specify another causal system that lawfully zags at the same points.

Let's say there's a mind with a transistor that outputs +3 volts at time T, indicating that it has just assented to some persuasive argument. Then we can build a highly similar physical cognitive system with a tiny little trapdoor underneath the transistor containing a little grey man who climbs out at time T and sets that transistor's output to—3 volts, indicating non-assent. Nothing acausal about that; the little grey man is there because we built him in. The notion of an argument that convinces *any* mind seems to involve a little blue woman who was *never* built into the system, who climbs out of literally *nowhere,* and strangles the little grey man, because that transistor has just *got* to output +3 volts: It's such a *compelling argument,* you see.

But compulsion is not a property of arguments, it is a property of minds that process arguments.

So the reason I'm arguing against the ghost, isn't *just* to make the point that (1) Friendly AI has to be explicitly programmed and (2) the laws of physics do not forbid Friendly AI. (Though of course I take a certain interest in establishing this.)

I also wish to establish the notion of a mind as a *causal, lawful, physical system* in which there *is no* irreducible central ghost that looks over the neurons / code and decides whether they are good suggestions.

(There is a concept in Friendly AI of *deliberately* programming an FAI to review its own source code and possibly hand it back to the programmers. But the mind that reviews is not irreducible, it is just the mind that you created. The FAI is renormalizing itself *however it was designed to do so;* there is nothing acausal reaching in from outside. A bootstrap, not a skyhook.)

All this echoes back to the discussion, a good deal earlier, of a Bayesian's "arbitrary" priors. If you show me one Bayesian who draws 4 red balls and 1 white ball from a barrel, and who assigns probability 5/7 to obtaining a red ball on the next occasion (by Laplace's Rule of Succession), then I can show you another mind which obeys Bayes's Rule to conclude a 2/7 probability of obtaining red on the next occasion—corresponding to a different prior belief about the barrel, but, perhaps, a less "reasonable" one.

Many philosophers are convinced that because you can in-principle construct a prior that updates to any given conclusion on a stream of evidence, therefore, Bayesian reasoning must be "arbitrary", and the whole schema of Bayesianism flawed, because it relies on "unjustifiable" assumptions, and indeed "unscientific", because you cannot force any possible journal editor in mindspace to agree with you.

And this (I then replied) relies on the notion that by unwinding all arguments and their justifications, you can obtain an ideal philosophy student of perfect emptiness, to be convinced by a line of reasoning that begins from absolutely no assumptions.

But who is this ideal philosopher of perfect emptiness? Why, it is just the irreducible core of the ghost!

And that is why (I went on to say) the result of trying to remove all assumptions from a mind, and unwind to the perfect absence of any prior, is not an ideal philosopher of perfect emptiness, but a rock. What is left of a mind after you remove the source code? Not the ghost who looks over the source code, but simply... no ghost.

So—and I shall take up this theme again later—wherever you are to locate your notions of *validity* or *worth* or *rationality* or *justification* or even *objectivity,* it cannot

rely on an argument that is *universally compelling to all physically possible minds.*

Nor can you ground validity in a sequence of justifications that, beginning from nothing, persuades a perfect emptiness.

Oh, there might be argument sequences that would compel any neurologically intact *human*—like the argument I use to make people let the AI out of the box[1]—but that is hardly the same thing from a philosophical perspective.

The first great failure of those who try to consider Friendly AI, is the One Great Moral Principle That Is All We Need To Program—aka the fake utility function—and of this I have already spoken.

But the even worse failure is the One Great Moral Principle We Don't Even *Need* To Program Because Any AI Must Inevitably Conclude It. This notion exerts a terrifying unhealthy fascination on those who spontaneously reinvent it; they dream of commands that no sufficiently advanced mind can disobey. The gods themselves will proclaim the rightness of their philosophy! (E.g. John C. Wright, Marc Geddes.)

There is also a less severe version of the failure, where the one does not *declare* the One True Morality. Rather the one hopes for an AI created *perfectly free*, unconstrained by flawed humans desiring slaves, so that the AI may arrive at virtue of its own accord—virtue undreamed-of perhaps by the speaker, who confesses themselves too flawed to teach an AI. (E.g. John K Clark, Richard Hollerith?, Eliezer$_{1996}$.) This is a less tainted motive than the dream of absolute command. But though *this* dream arises from virtue rather than vice, it is still based on a flawed understanding of freedom, and will not actually *work in real life.* Of this, more to follow, of course.

John C. Wright, who was previously writing a very nice transhumanist trilogy (first book: *The Golden Age*) inserted a huge Author Filibuster in the middle of his climactic third book, describing in tens of pages his Universal Morality That Must Persuade Any AI. I don't know if anything happened after that, because I stopped reading. And then Wright converted to Christianity—yes, seriously. So you *really don't* want to fall into this trap!

Footnote 1: Just kidding.

2-Place and 1-Place Words

I have previously spoken of the ancient, pulp-era magazine covers that showed a bug-eyed monster carrying off a girl in a torn dress; and about how people think as if sexiness is an inherent property of a sexy entity, without dependence on the admirer.

"Of *course* the bug-eyed monster will prefer human females to its own kind," says the artist (who we'll call Fred); "it can see that human females have soft, pleasant skin instead of slimy scales. It may be an alien, but it's not *stupid*—why are you expecting it to make such a basic mistake about sexiness?"

What is Fred's error? It is treating a function of 2 arguments ("2-place function"):

```
Sexiness: Admirer, Entity–> [0, ∞)
```

As though it were a function of 1 argument ("1-place function"):

```
Sexiness: Entity–> [0, ∞)
```

If Sexiness is treated as a function that accepts only one Entity as its argument, then of course Sexiness will appear to depend only on the Entity, with nothing else being relevant.

When you think about a two-place function as though it were a one-place function, you end up with a Variable Question Fallacy / Mind Projection Fallacy. Like trying to determine whether a building is *intrinsically* on the left or on the right side of the road, independent of anyone's travel direction.

An alternative and equally valid standpoint is that "sexiness" *does* refer to a one-place function—but each speaker uses a *different* one-place function to decide who to kidnap and ravish. Who says that just because Fred, the artist, and Bloogah, the bug-eyed monster, both use the word "sexy", they must mean the same thing by it?

If you take this viewpoint, there is no paradox in speaking of some woman intrinsically having 5 units of Fred::Sexiness. All onlookers can agree on this fact, once Fred::Sexiness has been specified in terms of curves, skin texture, clothing, status cues etc. This specification need *make no mention of Fred*, only the woman to be evaluated.

It so happens that Fred, himself, *uses* this algorithm to select flirtation targets. But that doesn't mean the algorithm itself has to *mention* Fred. So Fred's Sexiness function really *is* a function of one object—the woman—on this view. I called it Fred::Sexiness, but remember that this *name* refers to a function that is being described independently of Fred. Maybe it would be better to write:

Fred::Sexiness == Sexiness_20934

It is an empirical fact about Fred that he uses the function Sexiness_20934 to evaluate potential mates. Perhaps John uses exactly the same algorithm; it doesn't matter where it comes from once we have it.

And similarly, the same woman has only 0.01 units of Sexiness_72546, whereas a slime mold has 3 units of Sexiness_72546. It happens to be an empirical fact that Bloogah uses Sexiness_72546 to decide who to kidnap; that is, Bloogah::Sexiness names the fixed Bloogah-independent mathematical object that is the function Sexiness_72546.

Once we say that the woman has 0.01 units of Sexiness_72546 and 5 units of Sexiness_20934, all observers can agree on this without paradox.

And the two 2-place and 1-place views can be unified using the concept of "currying", named after the mathematician Haskell Curry. Currying is a technique allowed in certain programming language, where e.g. instead of writing

```
x = plus(2, 3)
```
(x = 5)

you can also write

```
y = plus(2)
```
(y is now a "curried" form of the function `plus`, which has eaten a `2`)

```
x = y(3)
```
(x = 5)

```
z = y(7)
```
(z = 9)

So `plus` is a 2-place function, but currying `plus`—letting it eat only one of its two required arguments—turns it into a 1-place function that adds 2 to any input. (Similarly, you could start with a 7-place function, feed it 4 arguments, and the result would be a 3-place function, etc.)

A true purist would insist that all functions should be viewed, by definition, as taking exactly 1 argument. On this view, `plus` accepts 1 numeric input, and outputs a *new* function; and this *new* function has 1 numeric input and finally outputs a number. On this view, when we write `plus(2, 3)` we are really computing `plus(2)` to get a function that adds 2 to any input, and then applying the result to `3`. A programmer would write this as:

```
plus: int–> (int–> int)
```

This says that `plus` takes an `int` as an argument, and returns a function of type `int–> int`.

Translating the metaphor back into the human use of words, we could imagine that "sexiness" starts by eating an Admirer, and spits out the fixed *mathematical* object that describes how the Admirer currently evaluates pulchritude. It is an *empirical* fact about the Admirer that their intuitions of desirability are computed in a way that is isomorphic to this *mathematical* function.

Then the mathematical object spit out by currying Sexiness(Admirer) can be applied to the Woman. If the Admirer was originally Fred, Sexiness(Fred) will first return Sexiness_20934. We can then say it is an empirical fact about the *Woman*, independently of Fred, that Sexiness_20934(Woman) = 5.

In Hilary Putnam's "Twin Earth" thought experiment, there was a tremendous philosophical brouhaha over whether it makes sense to postulate a Twin Earth which is just like our own, except that instead of water being H20, water is a *different* transparent flowing substance, XYZ. And furthermore, set the time of the thought experiment a few centuries ago, so in neither our Earth nor the Twin Earth does

anyone know how to test the alternative hypotheses of H20 vs. XYZ. Does the word "water" *mean* the same thing in that world, as in this one?

Some said, "Yes, because when an Earth person and a Twin Earth person utter the word 'water', they have the same sensory test in mind."

Some said, "No, because 'water' in our Earth means H20 and 'water' in the Twin Earth means XYZ."

If you think of "water" as a concept that *begins* by eating a world to find out the empirical true nature of that transparent flowing stuff, and *returns* a new fixed concept Water_42 or H20, then this world-eating concept is the same in our Earth and the Twin Earth; it just returns different answers in different places.

If you think of "water" as meaning H20 then the concept does nothing different when we transport it between worlds, and the Twin Earth contains no H20.

And of course there is no point in arguing over what the sound of the syllables "wa-ter" really means.

So should you pick one definition and use it consistently? But it's not that easy to save yourself from confusion. You have to train yourself to be *deliberately aware* of the distinction between the curried and uncurried forms of concepts.

When you take the uncurried water concept and apply it in a different world, it is the same concept but it *refers* to a different thing; that is, we are applying a constant world-eating function to a different world and obtaining a different return value. In the Twin Earth, XYZ is "water" and H20 is not; in our Earth, H20 is "water" and XYZ is not.

On the other hand, if you take "water" to refer to what the prior thinker would call "the result of applying 'water' to *our* Earth", then in the Twin Earth, XYZ is not water and H20 is.

The whole confusingness of the subsequent philosophical debate, rested on a tendency to *instinctively* curry concepts or *instinctively* uncurry them.

Similarly it takes an extra step for Fred to realize that other agents, like the Bug-Eyed-Monster agent, will choose kidnappees for ravishing based on $\text{Sexiness}_{\text{BEM}}(\text{Woman})$, not $\text{Sexiness}_{\text{Fred}}(\text{Woman})$. To do this, Fred must consciously re-envision Sexiness as a function with two arguments. All Fred's brain does by instinct is evaluate `Woman.sexiness`—that is, $\text{Sexiness}_{\text{Fred}}(\text{Woman})$; but it's simply labeled `Woman.sexiness`.

The fixed mathematical function Sexiness_20934 makes no mention of Fred or the BEM, only women, so Fred does not *instinctively* see why the BEM would evaluate "sexiness" any differently. And indeed the BEM would *not* evaluate Sexiness_20934 any differently, if for some odd reason it cared about the result of that particular function; but it is an *empirical* fact about the BEM that it uses a different function *to decide who to kidnap.*

If you're wondering as to the point of this analysis, we shall need it later in order to Taboo such confusing words as "objective", "subjective", and "arbitrary".

What Would You Do Without Morality?

To those who say "Nothing is real," I once replied, "That's great, but how does the nothing work?"

Suppose you learned, suddenly and definitively, that nothing is moral and nothing is right; that everything is permissible and nothing is forbidden.

Devastating news, to be sure—and no, I am not telling you this in real life. But suppose I *did* tell it to you. Suppose that, whatever you think is the basis of your moral philosophy, I convincingly tore it apart, and moreover showed you that nothing could fill its place. Suppose I *proved* that all utilities equaled zero.

I know that Your-Moral-Philosophy is as true and undisprovable as 2 + 2 = 4. But still, I ask that you do your best to perform the thought experiment, and concretely envision the possibilities even if they seem painful, or pointless, or logically incapable of any good reply.

Would you still tip cabdrivers? Would you cheat on your Significant Other? If a child lay fainted on the train tracks, would you still drag them off?

Would you still eat the same kinds of foods—or would you only eat the cheapest food, since there's no reason you *should* have fun—or would you eat very expensive food, since there's no reason you *should* save money for tomorrow?

Would you wear black and write gloomy poetry and denounce all altruists as fools? But there's no reason you *should* do that—it's just a cached thought.

Would you stay in bed because there was no reason to get up? What about when you finally got hungry and stumbled into the kitchen—what would you do after you were done eating?

Would you go on reading *Overcoming Bias,* and if not, what would you read instead? Would you still try to be rational, and if not, what would you think instead?

Close your eyes, take as long as necessary to answer:

What *would* you do, if nothing were right?

The Moral Void

Followup to: What Would You Do Without Morality?, Something to Protect

Once, discussing "horrible job interview questions" to ask candidates for a Friendly AI project, I suggested the following:

> Would you kill babies if it was *inherently* the right thing to do? Yes [] No []
>
> If "no", under what circumstances would you not do the right thing to do?
>
> _____
>
> If "yes", how inherently right would it have to be, for how many babies?
>
> _____

Yesterday I asked, "What would you do without morality?" There were numerous objections to the question, as well there should have been. Nonetheless there is more than one kind of person who can benefit from being asked this question. Let's say someone gravely declares, of some moral dilemma—say, a young man in Vichy France who must choose between caring for his mother and fighting for the Resistance—that there *is* no moral answer; both options are wrong and blamable; whoever faces the dilemma has had poor moral luck. Fine, let's suppose this is the case: then when you cannot be innocent, justified, or praiseworthy, what will you choose anyway?

Many interesting answers were given to my question, "What would you do without morality?". But one kind of answer was notable by its absence:

No one said, "I would ask what kind of behavior pattern was likely to maximize my inclusive genetic fitness, and execute that." Some misguided folk, not understanding evolutionary psychology, think that this must logically be the sum of morality. But if there *is* no morality, there's no reason to do such a thing—if it's not "moral", why bother?

You can probably see yourself pulling children off train tracks, even if it were not justified. But maximizing inclusive genetic fitness? If this *isn't* moral, why bother? Who does it help? It wouldn't even be much *fun*, all those egg or sperm donations.

And this is something you could say of most philosophies that have morality as a great light in the sky that shines from outside people. (To paraphrase Terry Pratchett.) If you believe that the meaning of life is to play non-zero-sum games because this is a trend built into the very universe itself...

Well, you might want to follow the corresponding ritual of reasoning about "the global trend of the universe" and implementing the result, *so long as you believe it to be moral*. But if you suppose that the light is switched off, so that the global trends of the universe are no longer moral, then why bother caring about "the global trend of the universe" in your decisions? If it's not right, that is.

Whereas if there were a child stuck on the train tracks, you'd probably drag the kid off *even if* there were no moral justification for doing so.

In 1966, the Israeli psychologist Georges Tamarin presented, to 1,066 schoolchildren ages 8-14, the Biblical story of Joshua's battle in Jericho:

> "Then they utterly destroyed all in the city, both men and women, young and old, oxen, sheep, and asses, with the edge of the sword... And they burned the city with fire, and all within it; only the silver and gold, and the vessels of bronze and of iron, they put into the treasury of the house of the LORD."

After being presented with the Joshua story, the children were asked:

> "Do you think Joshua and the Israelites acted rightly or not?"

66% of the children approved, 8% partially disapproved, and 26% totally disapproved of Joshua's actions.

A control group of 168 children was presented with an isomorphic story about "General Lin" and a "Chinese Kingdom 3,000 years ago". 7% of this group approved, 18% partially disapproved, and 75% completely disapproved of General Lin.

"What a horrible thing it is, teaching religion to children," you say, "giving them an off-switch for their morality that can be flipped just by saying the word 'God'." Indeed one of the saddest aspects of the whole religious fiasco is just how *little* it takes to flip people's moral off-switches. As Hobbes once said, "I don't know what's worse, the fact that everyone's got a price, or the fact that their price is so low." You can give people a book, and tell them God wrote it, and that's enough to switch off their moralities; God doesn't even have to tell them in person.

But are you sure you don't have a similar off-switch yourself? They flip so easily—you might not even notice it happening.

Leon Kass (of the President's Council on Bioethics) is glad to murder people so long as it's "natural", for example. He wouldn't pull out a gun and shoot you, but he *wants* you to die of old age and he'd be happy to pass legislation to ensure it.

And one of the *non*-obvious possibilities for such an off-switch, is "morality".

If you do happen to think that there is a source of morality beyond human beings... and I hear from quite a lot of people who are happy to rhapsodize on how Their-Favorite-Morality is built into the very fabric of the universe... then what if that morality tells you to kill people?

If you believe that there is any kind of stone tablet in the fabric of the universe, in the nature of reality, in the structure of logic—anywhere you care to put it—then what if you get a chance to read that stone tablet, and it turns out to say "Pain Is Good"? What then?

Maybe you should *hope* that morality isn't written into the structure of the universe. What if the structure of the universe says to do something horrible?

And if an external objective morality *does* say that the universe *should* occupy some horrifying state... let's not even ask what you're going to do about that. No, instead I ask: What would you have *wished* for the external objective morality to be instead? What's the *best* news you could have gotten, reading that stone tablet?

Go ahead. Indulge your fantasy. Would you *want* the stone tablet to say people should die of old age, or that people should live as long as they wanted? If you could write the stone tablet yourself, what would it say?

Maybe you should just do *that*?

I mean... if an external objective morality tells you to kill people, why *should* you even listen?

There is a courage that goes beyond even an atheist sacrificing their life and their hope of immortality. It is the courage of a theist who goes against what they believe to be the Will of God, choosing eternal damnation *and defying even morality* in order to rescue a slave, or speak out against hell, or kill a murderer... You don't get a chance to reveal that virtue without making fundamental mistakes about how the universe works, so it is not something to which a rationalist should aspire. But it warms my heart that humans are capable of it.

I have previously spoken of how, to achieve rationality, it is necessary to have some purpose so desperately important to you as to be more important than "rationality", so that you will not choose "rationality" over success.

To learn the Way, you must be able to unlearn the Way; so you must be able to give up the Way; so there must be something dearer to you than the Way. This is so in questions of truth, and in questions of strategy, and also in questions of morality.

The "moral void" of which this post is titled, is not the terrifying abyss of utter meaningless. Which for a bottomless pit is surprisingly shallow; what *are* you supposed to do about it besides wearing black makeup?

No. The void I'm talking about is a virtue which is nameless.

Part of *The Metaethics Sequence*

Next post: "Created Already In Motion"

Previous post: "What Would You Do Without Morality?"

Created Already In Motion

Followup to: No Universally Compelling Arguments, Passing the Recursive Buck

Lewis Carroll, who was also a mathematician, once wrote a short dialogue called What the Tortoise said to Achilles. If you have not yet read this ancient classic, consider doing so now.

The Tortoise offers Achilles a step of reasoning drawn from Euclid's First Proposition:

(A) Things that are equal to the same are equal to each other.
(B) The two sides of this Triangle are things that are equal to the same.
(Z) The two sides of this Triangle are equal to each other.

Tortoise: "And if some reader had *not* yet accepted A and B as true, he might still accept the *sequence* as a *valid* one, I suppose?"

Achilles: "No doubt such a reader might exist. He might say, 'I accept as true the Hypothetical Proposition that, *if* A and B be true, Z must be true; but, I *don't* accept A and B as true.' Such a reader would do wisely in abandoning Euclid, and taking to football."

Tortoise: "And might there not *also* be some reader who would say, 'I accept A and B as true, but I *don't* accept the Hypothetical'?"

Achilles, unwisely, concedes this; and so asks the Tortoise to accept another proposition:

(C) If A and B are true, Z must be true.

But, asks, the Tortoise, suppose that he accepts A and B and C, but not Z?

Then, says, Achilles, he must ask the Tortoise to accept one more hypothetical:

(D) If A and B and C are true, Z must be true.

Douglas Hofstadter paraphrased the argument some time later:

> Achilles: If you have [(A⋀B)→Z], and you also have (A⋀B), then surely you have Z.
> Tortoise: Oh! You mean <{(A⋀B)⋀[(A⋀B)→Z]}→Z>, don't you?

As Hofstadter says, "Whatever Achilles considers a rule of inference, the Tortoise immediately flattens into a mere string of the system. If you use only the letters A, B, and Z, you will get a recursive pattern of longer and longer strings."

By now you should recognize the anti-pattern Passing the Recursive Buck; and though the counterspell is sometimes hard to find, when found, it generally takes the form The Buck Stops Immediately.

The Tortoise's mind needs the *dynamic* of adding Y to the belief pool when X and (X→Y) are previously in the belief pool. If this dynamic is not present—a rock, for example, lacks it—then you can go on adding in X and (X→Y) and (X⋀(X→Y))→Y until the end of eternity, without ever getting to Y.

The phrase that once came into my mind to describe this requirement, is that a mind must be *created already in motion.* There is no argument so compelling that it will give dynamics to a static thing. There is no computer program so *persuasive* that you can run it on a rock.

And even if you have a mind that *does* carry out modus ponens, it is futile for it to have such beliefs as...

(A) If a toddler is on the train tracks, then pulling them off is fuzzle.
(B) There is a toddler on the train tracks.

...unless the mind also *implements:*

Dynamic: When the belief pool contains "X is fuzzle", send X to the action system.

(**Added:** Apparently this wasn't clear... By "dynamic" I mean a property of a physically implemented cognitive system's *development over time*. A "dynamic" is something that *happens inside* a cognitive system, *not* data that it stores in memory and manipulates. Dynamics are the manipulations. There is no way to write a dynamic on a piece of paper, because the paper will just lie there. So the text immediately above, which says "dynamic", is not dynamic. If I wanted the text to *be* dynamic and not just *say* "dynamic", I would have to write a Java applet.)

Needless to say, having the belief...

(C) If the belief pool contains "X is fuzzle", then "send 'X' to the action system" is fuzzle.

...won't help unless the mind already implements the *behavior* of translating hypothetical actions labeled 'fuzzle' into actual motor actions.

By dint of careful arguments about the nature of cognitive systems, you might be able to prove...

(D) A mind with a dynamic that sends plans labeled "fuzzle" to the action system, is more fuzzle than minds that don't.

...but that *still* won't help, unless the listening mind *previously* possessed the *dynamic* of swapping out its current source code for alternative source code that is believed to be more fuzzle.

This is why you can't argue fuzzleness into a rock.

Part of *The Metaethics Sequence*

Next post: "The Bedrock of Fairness"

Previous post: "The Moral Void"

The Bedrock of Fairness

Followup to: The Moral Void

Three people, whom we'll call Xannon, Yancy and Zaire, are separately wandering through the forest; by chance, they happen upon a clearing, meeting each other. Introductions are performed. And then they discover, in the center of the clearing, a delicious blueberry pie.

Xannon: "A pie! What good fortune! But which of us should get it?"

Yancy: "Let us divide it fairly."

Zaire: "I agree; let the pie be distributed fairly. Who could argue against fairness?"

Xannon: "So we are agreed, then. But what is a fair division?"

Yancy: "Eh? Three equal parts, of course!"

Zaire: "Nonsense! A fair distribution is half for me, and a quarter apiece for the two of you."

Yancy: "*What?* How is *that* fair?"

Zaire: "I'm hungry, therefore I should be fed; that is fair."

Xannon: "Oh, dear. It seems we have a dispute as to what is fair. For myself, I want to divide the pie the same way as Yancy. But let us resolve this dispute over the meaning of fairness, fairly: that is, giving equal weight to each of our desires. Zaire desires the pie to be divided {1/4, 1/4, 1/2}, and Yancy and I desire the pie to be divided {1/3, 1/3, 1/3}. So the fair compromise is {11/36, 11/36, 14/36}."

Zaire: "*What*? That's crazy. There's two different opinions as to how fairness works—why should the opinion that happens to be yours, get twice as much weight as the opinion that happens to be mine? Do you think your theory is twice as good? *I* think my theory is a *hundred* times as good as yours! So there!"

Yancy: "Craziness indeed. Xannon, I already took Zaire's desires into account in saying that he should get 1/3 of the pie. You can't count the same factor twice. Even if we count fairness as an inherent desire, why should Zaire be rewarded for being selfish? Think about which agents thrive under your system!"

Xannon: "Alas! I was hoping that, even if we could not agree on how to distribute the pie, we could agree on a fair resolution procedure for our dispute, such as averaging our desires together. But even that hope was dashed. Now what are we to do?"

Yancy: "Xannon, you are overcomplicating things. 1/3 apiece. It's not that complicated. A fair distribution is an even split, not a distribution arrived at by a 'fair resolution procedure' that everyone agrees on. What if we'd all been raised in a society that believed that men should get twice as much pie as women? Then we would split the pie unevenly, and even though no one of us disputed the split, it would *still* be unfair."

Xannon: "*What*? Where is this 'fairness' stored if not in human minds? Who says that something is unfair if no intelligent agent does so? Not upon the stars or the mountains is 'fairness' written."

Yancy: "So what you're saying is that if you've got a whole society where women are chattel and men sell them like farm animals and it hasn't occurred to anyone that things could be other than they are, that this society is fair, and at the exact moment where someone first realizes it shouldn't have to be that way, the whole society suddenly becomes unfair."

Xannon: "How can a society be unfair without some specific party who claims injury and receives no reparation? If it hasn't occurred to anyone that things could work differently, and no one's *asked* for things to work differently, then—"

Yancy: "Then the women are still being treated like farm animals and *that is unfair.* Where's your common sense? Fairness is not agreement, fairness is symmetry."

Zaire: "Is this all working out to my getting half the pie?"

Yancy: "No."

Xannon: "I don't know... maybe as the limit of an infinite sequence of meta-meta-fairnesses..."

Zaire: "I fear I must accord with Yancy on one point, Xannon; your desire for perfect accord among us is misguided. I want half the pie. Yancy wants me to have a third of the pie. This is all there is to the world, and all there ever was. If two monkeys want the same banana, in the end one will have it, and the other will cry morality. Who gets to form the committee to decide the rules that will be used to determine what is 'fair'? Whoever it is, got the banana."

Yancy: "I wanted to give you a third of the pie, and you equate this to seizing the whole thing for myself? Small wonder that you don't want to acknowledge the existence of morality—you don't want to acknowledge that anyone can be so much less of a jerk."

Xannon: "You oversimplify the world, Zaire. Banana-fights occur across thousands and perhaps millions of species, in the animal kingdom. But if this were all there was, *Homo sapiens* would never have evolved moral intuitions. Why would the human animal evolve to cry morality, if the cry had no effect?"

Zaire: "To make themselves feel better."

Yancy: "Ha! You fail at evolutionary biology."

Xannon: "A murderer accosts a victim, in a dark alley; the murderer desires the victim to die, and the victim desires to live. Is there nothing more to the universe than their conflict? No, because if I happen along, I will side with the victim, and not with the murderer. The victim's plea crosses the gap of persons, to me; it is not locked up inside the victim's own mind. But the murderer cannot obtain my sympathy, nor incite me to help murder. Morality crosses the gap between persons; you might not see it in a conflict between two people, but you would see it in a society."

Yancy: "So you define morality as that which crosses the gap of persons?"

Xannon: "It seems to me that *social* arguments over disputed goals are how human moral intuitions arose, beyond the simple clash over bananas. So that is how I define the term."

Yancy: "Then I disagree. If someone wants to murder me, and the two of us are alone, then I am still in the right and they are still in the wrong, even if no one else is present."

Zaire: "And the murderer says, 'I am in the right, you are in the wrong'. So what?"

Xannon: "How does your statement that you are in the right, and the murderer is in the wrong, impinge upon the universe—if there is no one else present to be persuaded?"

Yancy: "It licenses *me* to resist being murdered; which I might not do, if I thought that my desire to avoid being murdered was wrong, and the murderer's desire to kill me was right. I can distinguish between things I merely want, and things that are right—though alas, I do not always live up to my own standards. The murderer is blind to the morality, perhaps, but that doesn't change the morality. And if we were *both* blind, the morality *still* would not change."

Xannon: "Blind? What is being seen, what sees it?"

Yancy: "You're trying to treat fairness as... I don't know, something like an array-mapped 2-place function that goes out and eats a list of human minds, and returns a list of what each person thinks is 'fair', and then averages it together. The problem with this isn't just that different people could have different ideas about fairness. It's not just that they could have different ideas about how to combine the results. It's that it leads to infinite recursion outright—passing the recursive buck. You want there to be some level on which everyone agrees, but at least some possible minds will disagree with any statement you make."

Xannon: "Isn't the whole point of fairness to let people agree on a division, instead of fighting over it?"

Yancy: "What is *fair* is one question, and whether someone else *accepts* that this is fair is another question. What is fair? That's easy: an equal division of the pie is fair. Anything else won't be fair no matter what kind of pretty arguments you put around it. Even if I *gave* Zaire a sixth of my pie, that might be a *voluntary* division but it wouldn't be a *fair* division. Let *fairness* be a simple and object-level procedure, instead of this infinite meta-recursion, and the buck will stop immediately."

Zaire: "If the word 'fair' simply means 'equal division' then why not just say 'equal division' instead of this strange additional word, 'fair'? You want the pie divided equally, I want half the pie for myself. That's the whole fact of the matter; this word 'fair' is merely an attempt to get more of the pie for yourself."

Xannon: "If that's the whole fact of the matter, why would anyone talk about 'fairness' in the first place, I wonder?"

Zaire: "Because they all share the same delusion."

Yancy: "A delusion of *what?* What is it that you are saying people *think incorrectly* the universe is like?"

Zaire: "I am under no obligation to describe other people's confusions."

Yancy: "If you can't dissolve their confusion, how can you be sure they're confused? But it seems clear enough to me that if the word *fair* is going to have any meaning at all, it has to finally add up to each of us getting one-third of the pie."

Xannon: "How odd it is to have a procedure of which we are more sure of the result than the procedure itself."

Zaire: "Speak for yourself."

Part of *The Metaethics Sequence*

Next post: "Moral Complexities"

Previous post: "Created Already In Motion"

Moral Complexities

Followup to: The Bedrock of Fairness

Discussions of morality seem to me to often end up turning around two different intuitions, which I might label morality-as-preference and morality-as-given. The former crowd tends to equate morality with what people want; the latter to regard morality as something you can't change by changing people.

As for me, I have my own notions, which I am working up to presenting. But above all, I try to avoid avoiding difficult questions. Here are what I see as (some of) the difficult questions for the two intuitions:

- For morality-as-preference:
  - Why do people seem to mean different things by "I want the pie" and "It is right that I should get the pie"? Why are the two propositions argued in different ways?
  - When and why do people change their terminal values? Do the concepts of "moral error" and "moral progress" have referents? Why would anyone want to change what they want?
  - Why and how does anyone ever "do something they know they shouldn't", or "want something they know is wrong"? Does the notion of morality-as-preference really add up to moral normality?
- For morality-as-given:
  - Would it be possible for everyone in the world to be wrong about morality, *and* wrong about how to update their beliefs about morality, *and* wrong about how to choose between metamoralities, etcetera? So that there would be a morality, but it would be entirely outside our frame of reference? What distinguishes this state of affairs, from finding a random stone tablet showing the words "You should commit suicide"?
  - How does a world in which a moral proposition is true, differ from a world in which that moral proposition is false? If the answer is "no", how does anyone perceive moral givens?
  - Is it better for people to be happy than sad? If so, why does morality look amazingly like godshatter of natural selection?
  - Am I not allowed to construct an alien mind that evaluates morality differently? What will stop me from doing so?

Part of *The Metaethics Sequence*

Next post: "Is Morality Preference?"

Previous post: "The Bedrock of Fairness"

Is Morality Preference?

Followup to: Moral Complexities

In the dialogue "The Bedrock of Fairness", I intended Yancy to represent morality-as-raw-fact, Zaire to represent morality-as-raw-whim, and Xannon to be a particular kind of attempt at compromising between them. Neither Xannon, Yancy, or Zaire represent my own views—rather they are, in their disagreement, showing the *problem* that I am trying to solve. It is futile to present answers to which questions are lacking.

But characters have independent life in the minds of all readers; when I create a *dialogue*, I don't view my authorial intent as primary. Any good interpretation can be discussed. I meant Zaire to be asking for half the pie out of pure selfishness; many readers interpreted this as a genuine need... which is as interesting a discussion to have as any, though it's a different discussion.

With this in mind, I turn to Subhan and Obert, who shall try to answer yesterday's questions on behalf of their respective viewpoints.

Subhan makes the opening statement:

Subhan: "I defend this proposition: that there is no reason to talk about a 'morality' distinct from what people want."

Obert: "I challenge. Suppose someone comes to me and says, 'I want a slice of that pie you're holding.' It seems to me that they have just made a very different statement from 'It is *right* that I should get a slice of that pie'. I have no reason at all to doubt the former statement—to suppose that they are lying to me about their desires. But when it comes to the latter proposition, I have reason indeed to be skeptical. Do you say that these two statements *mean* the same thing?"

Subhan: "I suggest that when the pie-requester says to you, 'It is right for me to get some pie', this asserts that *you* want the pie-requester to get a slice."

Obert: "Why should *I* need to be told what *I* want?"

Subhan: "You take a needlessly restrictive view of wanting, Obert; I am not setting out to reduce humans to creatures of animal instinct. Your wants include those desires you label 'moral values', such as wanting the hungry to be fed—"

Obert: "And you see no distinction between my desire to feed the hungry, and my desire to eat all the delicious pie myself?"

Subhan: "No! They are both desires—backed by *different* emotions, perhaps, but both desires. To continue, the pie-requester hopes that you have a desire to feed the hungry, and so says, 'It is right that I should get a slice of this pie', to remind you of your own desire. We do not automatically know all the consequences of our own wants; we are not logically omniscient."

Obert: "This seems psychologically unrealistic—I don't think that's what goes through the mind of the person who says, 'I have a right to some pie'. In this latter case, if I deny them pie, they will feel *indignant.* If they are only trying to remind me of my own desires, why should they feel indignant?"

Subhan: "Because they didn't get any pie, so they're frustrated."

Obert: "Unrealistic! Indignation at moral transgressions has a psychological dimension that goes beyond struggling with a struck door."

Subhan: "Then consider the evolutionary psychology. The pie-requester's emotion of indignation would evolve as a display, first to remind you of the potential consequences of offending fellow tribe-members, and second, to remind any observing tribe-members of goals *they* may have to feed the hungry. By refusing to share, you would offend against a social norm—which is to say, a widely shared want."

Obert: "So you take refuge in social wants as the essence of morality? But people seem to see a difference between desire and morality, *even* in the quiet of their own minds. They say things like: 'I want X, but the right thing to do is Y... what shall I do?'"

Subhan: "So they experience a conflict between their want to eat pie, and their want to feed the hungry—which they know is also a want of society. It's not predetermined that the prosocial impulse will be victorious, but they are both impulses."

Obert: "And when, during WWII, a German hides Jews in their basement—*against* the wants of surrounding society—how then?"

Subhan: "People do not always define their in-group by looking at their next-door neighbors; they may conceive of *their group* as 'good Christians' or 'humanitarians'."

Obert: "I should sooner say that people choose their in-groups by looking for others who share their beliefs about morality—not that they construct their morality from their in-group."

Subhan: "Oh, *really?* I should not be surprised if that were experimentally testable—if so, how much do you want to bet?"

Obert: "That the Germans who hid Jews in their basements, chose who to call *their people* by looking at their beliefs about morality? Sure. I'd bet on that."

Subhan: "But in any case, even if a German resister has a desire to preserve life which is so strong as to go against their own perceived 'society', it is still *their desire*."

Obert: "Yet they would attribute to that desire, the same distinction they make between 'right' and 'want'—even when going *against* society. They might think to themselves, 'How dearly I wish I could stay out of this, and keep my family safe. But it is my duty to hide these Jews from the Nazis, and I must fulfill that duty.' There is an interesting moral question, as to whether it reveals greater heroism, to fulfill a duty eagerly, or to fulfill your duties when you are not eager. For myself I should just total up the lives saved, and call that their score. But I digress... The distinction between 'right' and 'want' is not explained by your distinction of socially shared and individual wants. The distinction between desire and duty seems to me a basic thing, which someone could experience floating alone in a spacesuit a thousand light-years from company."

Subhan: "Even if I were to grant this *psychological* distinction, perhaps that is simply a matter of emotional flavoring. Why should I not describe perceived duties as a differently flavored want?"

Obert: "Duties, and should-ness, seem to have a dimension that goes beyond our whims. If we want different pizza toppings today, we can order a different pizza without guilt; but we cannot choose to make murder a good thing."

Subhan: "Schopenhauer: 'A man can do as he wills, but not will as he wills.' You cannot decide to make salad taste better to you than cheeseburgers, and you cannot decide *not* to dislike murder. Furthermore, people do change, albeit rarely, those wants that you name 'values'; indeed they are easier to change than our food tastes."

Obert: "Ah! That is something I meant to ask *you* about. People sometimes change their morals; *I* would call this updating their beliefs about morality, but *you* would call it changing their wants. Why would anyone want to change their wants?"

Subhan: "Perhaps they simply find that their wants have changed; brains do change over time. Perhaps they have formed a *verbal belief* about what they want, which they have discovered to be mistaken. Perhaps society has changed, or their perception of society has changed. But really, in most cases you don't have to go that far, to explain apparent changes of morality."

Obert: "Oh?"

Subhan: "Let's say that someone begins by thinking that Communism is a good social system, has some arguments, and ends by believing that Communism is a bad social system. This does not mean that their *ends* have changed—they may simply have gotten a good look at the history of Russia, and decided that Communism is a poor *means* to the end of raising standards of living. I challenge you to find me a case of changing morality in which people change their terminal values, and not just their beliefs about which acts have which consequences."

Obert: "Someone begins by believing that God ordains against premarital sex; they find out there is no God; subsequently they approve of premarital sex. This, let us specify, is *not* because of fear of Hell; but because previously they believed that God had the power to ordain, or knowledge to tell them, what is *right;* in ceasing to believe in God, they updated their belief about what is right."

Subhan: "I am not responsible for straightening others' confusions; this one is merely in a general state of disarray around the 'God' concept."

Obert: "All right; suppose I get into a moral argument with a man from a society that practices female circumcision. I do not think our argument is about the *consequences* to the woman; the argument is about the morality of these consequences."

Subhan: "Perhaps the one falsely believes that women have no feelings—"

Obert: "Unrealistic, unrealistic! It is far more likely that the one hasn't really considered whether the woman has feelings, because he doesn't see any obligation to care. The happiness of women is not a terminal value to him. Thousands of years ago, most societies devalued consequences to women. They also had false beliefs about women, true—and false beliefs about men as well, for that matter—but nothing like the Victorian era's complex rationalizations for how paternalistic rules really benefited women. The Old Testament doesn't explain *why* it levies the death penalty for a woman wearing men's clothing. It certainly doesn't explain how this rule really benefits women after all. It's not the sort of argument it would have occurred to the authors to rationalize! They didn't *care* about the consequences to women."

Subhan: "So they wanted different things than you; what of it?"

Obert: "See, now that is exactly why I cannot accept your viewpoint. *Somehow,* societies went from Old Testament attitudes, to democracies with female suffrage. And this transition—however it occurred—was caused by people saying, 'What this society does to women is a great wrong!', not, 'I would personally prefer to treat women better.' That's not just a change in semantics—it's the difference between being obligated to stand and deliver a justification, versus being able to just say, 'Well, I prefer differently, end of discussion.' And who says that humankind has finished with its moral progress? You're yanking the ladder out from underneath a very important climb."

Subhan: "Let us suppose that the change of human societies over the last ten thousand years, has been accompanied by a change in terminal values—"

Obert: "You call this a *supposition?* Modern political debates turn around vastly different valuations of consequences than in ancient Greece!"

Subhan: "I am not so sure; human cognitive psychology has not had time to change evolutionarily over that period. Modern democracies tend to appeal to our empathy for those suffering; that empathy existed in ancient Greece as well, but it was invoked less often. In each single moment of argument, I doubt you would find modern politicians appealing to *emotions that didn't exist* in ancient Greece."

Obert: "I'm not saying that emotions have changed; I'm saying that beliefs about morality have changed. Empathy merely provides emotional depth to an argument that can be made on a purely logical level: 'If it's wrong to enslave you, if it's wrong to enslave your family and your friends, then how can it be right to enslave people who happen to be a different color? What difference does the color make?' If morality is just preference, then there's a very simple answer: 'There is no right or wrong, I just like my own family better.' You see the problem here?"

Subhan: "Logical fallacy: Appeal to consequences."

Obert: "I'm not appealing to consequences. I'm showing that when I reason about 'right' or 'wrong', I am reasoning about something that does *not* behave like 'want' and 'don't want'."

Subhan: "Oh? But I think that in reality, your rejection of morality-as-preference has a great deal to do with your fear of where the truth leads."

Obert: "Logical fallacy: Ad hominem."

Subhan: "Fair enough. Where were we?"

Obert: "If morality is preference, why would you want to change your wants to be more inclusive? Why would you want to change your wants at all?"

Subhan: "The answer to your first question probably has to do with a fairness instinct, I would suppose—a notion that the tribe should have the same rules for everyone."

Obert: "I don't think that's an instinct. I think that's a triumph of three thousand years of moral philosophy."

Subhan: "That could be tested."

Obert: "And my second question?"

Subhan: "Even if terminal values change, it doesn't mean that terminal values are stored on a great stone tablet outside humanity. Indeed, it would seem to argue against it! It just means that some of the events that go on in our brains, can change what we want."

Obert: "*That's* your concept of moral progress? *That's* your view of the last three thousand years? *That's* why we have free speech, democracy, mass street protests against wars, nonlethal weapons, no more slavery—"

Subhan: "If you wander on a random path, and you compare all past states to your present state, you will see continuous 'advancement' toward your present condition—"

Obert: "*Wander on a random path?*"

Subhan: "I'm just pointing out that saying, 'Look how much better things are now', when your criterion for 'better' is comparing past moral values to yours, does not establish any directional trend in human progress."

Obert: "Your strange beliefs about the nature of morality have destroyed your soul. I don't even believe in souls, and I'm saying that."

Subhan: "Look, depending on which arguments do, in fact, move us, you might be able to regard the process of changing terminal values as a directional progress. You might be able to show that the change had a consistent trend as we thought of more and more arguments. But that doesn't show that morality is something *outside* us. We could even—though this is psychologically unrealistic—choose to *regard you* as computing a converging approximation to your 'ideal wants', so that you would have meta-values that defined both your present value and the rules for updating them. But these would be *your* meta-values and *your* ideals and *your* computation, just as much as pepperoni is *your own* taste in pizza toppings. You may not know your *real* favorite ever pizza topping, until you've tasted many possible flavors."

Obert: "Leaving out *what* it is that you just compared to pizza toppings, I begin to be suspicious of the all-embracingness of your viewpoint. No matter *what* my mind does, you can simply call it a still-more-modified 'want'. I think that *you* are the one suffering from meta-level confusion, not I. Appealing to right is not the same as appealing to desire. Just because the appeal is judged *inside my brain*, doesn't mean that the appeal is not *to* something more than my desires. Why can't my brain compute duties as well as desires?"

Subhan: "What is the difference between duty and desire?"

Obert: "A duty is something you must do whether you want to or not."

Subhan: "Now you're just being incoherent. Your brain computes something it wants to do whether it wants to or not?"

Obert: "No, *you* are the one whose theory makes this incoherent. Which is why your theory ultimately fails to add up to morality."

Subhan: "I say again that you underestimate the power of mere wanting. And more: *You* accuse *me* of incoherence? You say that *I* suffer from meta-level confusion?"

Obert: "Er... yes?"

To be continued...

Part of *The Metaethics Sequence*

Next post: "Is Morality Given?"

Previous post: "Moral Complexities"

Is Morality Given?

Continuation of: Is Morality Preference?

(Disclaimer: Neither Subhan nor Obert represent my own position on morality; rather they represent different sides of the *questions* I hope to answer.)

Subhan: "What is this 'morality' stuff, if it is *not* a preference within you?"

Obert: "I know that my mere wants, don't change what is *right;* but I don't claim to have absolute knowledge of what is right—"

Subhan: "You're not escaping that easily! How does a universe in which murder is wrong, differ from a universe in which murder is right? How can you detect the difference experimentally? If the answer to that is 'No', then how does any human being come to *know* that murder is wrong?"

Obert: "Am I allowed to say 'I don't know'?"

Subhan: "No. You believe *now* that murder is wrong. You must believe you *already* have evidence and you should be able to present it *now.*"

Obert: "That's too strict! It's like saying to a hunter-gatherer, 'Why is the sky blue?' and expecting an immediate answer."

Subhan: "No, it's like saying to a hunter-gatherer: Why do you *believe* the sky is blue?"

Obert: "Because it seems blue, just as murder seems wrong. Just don't ask me what the sky is, or how I can see it."

Subhan: "But—aren't we discussing the nature of morality?"

Obert: "That, I confess, is not one of my strong points. I specialize in plain old morality. And as a matter of morality, I know that I can't make murder *right* just by wanting to kill someone."

Subhan: "But if you *wanted* to kill someone, you would say, 'I know murdering this guy is right, and I couldn't make it wrong just by not wanting to do it.'"

Obert: "Then, if I said that, I would be wrong. That's common moral sense, right?"

Subhan: "Argh! It's difficult to even argue with you, since you won't tell me exactly what you think morality is made of, or where you're getting all these amazing moral truths—"

Obert: "Well, I do regret having to frustrate you. But it's more important that I *act morally,* than that I come up with amazing new theories of the *nature* of morality. I don't claim that my strong point is in explaining the fundamental nature of morality. Rather, my strong point is coming up with theories of morality that give normal moral answers to questions like, 'If you feel like killing someone, does that make it right to do so?' The common-sense answer is 'No' and I really see no reason to adopt a theory that makes the answer 'Yes'. Adding up to moral normality—*that* is my theory's strong point."

Subhan: "Okay... look. You say that, if you believed it was right to murder someone, you would be *wrong.*"

Obert: "Yes, of course! And just to cut off any quibbles, we'll specify that we're not talking about going back in time and shooting Stalin, but rather, stalking some innocent bystander through a dark alley and slitting their throat for no other reason but my own enjoyment. That's *wrong*."

Subhan: "And *anyone* who says murder is right, is mistaken."

Obert: "Yes."

Subhan: "Suppose there's an alien species somewhere in the vastness of the multiverse, who evolved from carnivores. In fact, through most of their evolutionary history, they were cannibals. They've evolved different emotions from us, and they have no concept that murder is wrong—"

Obert: "Why doesn't their society fall apart in an orgy of mutual killing?"

Subhan: "That doesn't matter for our purposes of theoretical metaethical investigation. But since you ask, we'll suppose that the Space Cannibals have a strong sense of *honor*—they won't kill someone they promise not to kill; they have a very strong idea that violating an oath is wrong. Their society holds together on that basis, and on the basis of vengeance contracts with private assassination companies. But so far as the actual killing is concerned, the aliens just think it's fun. When someone gets executed for, say, driving through a traffic light, there's a bidding war for the rights to personally tear out the offender's throat."

Obert: "Okay... where is this going?"

Subhan: "I'm proposing that the Space Cannibals not only have no sense that murder is wrong—indeed, they have a positive sense that killing is an important part of life—but moreover, there's no path of arguments you could use to *persuade* a Space Cannibal of your view that murder is wrong. There's no fact the aliens can learn, and no chain of reasoning they can discover, which will *ever* cause them to conclude that murder is a moral wrong. Nor is there any way to persuade them that they *should* modify themselves to perceive things differently."

Obert: "I'm not sure I believe *that's* possible—"

Subhan: "Then you believe in universally compelling arguments processed by a ghost in the machine. For every possible mind whose utility function assigns terminal value +1, mind design space contains an equal and opposite mind whose utility function assigns terminal value—1. A mind is a physical device and you can't have a little blue woman pop out of nowhere and make it say 1 when the physics calls for it to say 0."

Obert: "Suppose I were to concede this. Then?"

Subhan: "Then it's possible to have an alien species that believes murder is not wrong, and moreover, will continue to believe this given knowledge of every possible fact and every possible argument. Can you say these aliens are *mistaken?*"

Obert: "Maybe it's the right thing to do in *their* very different, alien world—"

Subhan: "And then they land on Earth and start slitting human throats, laughing all the while, because they don't believe it's wrong. Are they *mistaken?*"

Obert: "Yes."

Subhan: "Where exactly is the mistake? In which step of reasoning?"

Obert: "I don't know exactly. My guess is that they've got a bad axiom."

Subhan: "Dammit! Okay, look. Is it possible that—by analogy with the Space Cannibals—there are true moral facts of which the human species is not only *presently* unaware, but incapable of perceiving *in principle?* Could we have been born defective—incapable even of being *compelled* by the arguments that would lead us to the light? Moreover, born without any desire to modify ourselves to be capable of understanding such arguments? Could we be *irrevocably mistaken* about morality—just like you say the Space Cannibals are?"

Obert: "I... guess so..."

Subhan: "You guess so? Surely this is an inevitable consequence of believing that morality is a given, independent of anyone's preferences! Now, is it possible that *we,* not the Space Cannibals, are the ones who are irrevocably mistaken in believing that murder is wrong?"

Obert: "*That* doesn't seem likely."

Subhan: "I'm not asking you if it's likely, I'm asking you if it's *logically possible!* If it's *not* possible, then you have just confessed that human morality is ultimately determined by our human constitutions. And if it *is* possible, then what distinguishes this scenario of 'humanity is irrevocably mistaken about morality', from finding a stone tablet on which is written the phrase 'Thou Shalt Murder' without any known justification attached? How is a given morality any different from an unjustified stone tablet?"

Obert: "Slow down. Why does this argument show that morality is determined by our own constitutions?"

Subhan: "Once upon a time, theologians tried to say that God was the foundation of morality. And even since the time of the ancient Greeks, philosophers were sophisticated enough to go on and ask the next question—'*Why* follow God's commands?' Does God have *knowledge* of morality, so that we should follow Its orders as good advice? But then what is this morality, outside God, of which God has knowledge? Do God's commands *determine* morality? But then why, *morally,* should one follow God's orders?"

Obert: "Yes, this demolishes attempts to answer questions about the nature of morality just by saying 'God!', unless you answer the obvious further questions. But so what?"

Subhan: "And furthermore, let us castigate those who made the argument originally, for the sin of trying to *cast off responsibility*—trying to wave a scripture and say, 'I'm just following God's orders!' Even if God *had* told them to do a thing, it would still have been *their own decision* to follow God's orders."

Obert: "I agree—as a matter of morality, there is no evading of moral responsibility. Even if your parents, or your government, or some kind of hypothetical superintelligence, tells you to do something, you are responsible for your decision in doing it."

Subhan: "But you see, this also demolishes the idea of any morality that is outside, beyond, or above human preference. Just substitute 'morality' for 'God' in the argument!"

Obert: "*What?*"

Subhan: "John McCarthy said: 'You say you couldn't live if you thought the world had no purpose. You're saying that you can't form purposes of your own-that you need someone to tell you what to do. The average child has more gumption than that.' For every kind of stone tablet that you might imagine anywhere, in the trends of the universe or in the structure of logic, you are still left with the question: 'And *why* obey this morality?' It would be *your decision* to follow this trend of the universe, or obey this structure of logic. Your decision—and *your preference.*"

Obert: "That doesn't follow! Just because it is *my decision* to be moral—and even because there are drives in me that lead me to make that decision—it doesn't follow that the morality I follow consists *merely* of my preferences. If someone gives me a pill that makes me prefer to *not* be moral, to commit murder, then this just alters my preference—but *not* the morality; murder is still wrong. That's common moral sense—"

Subhan: "I beat my head against my keyboard! What about *scientific* common sense? If morality is this mysterious *given* thing, from beyond space and time—and I don't even see why we *should* follow it, in that case—but in any case, if morality exists independently of human nature, then isn't it a *remarkable coincidence* that, say, *love* is good?"

Obert: "Coincidence? How so?"

Subhan: "Just where on Earth do you think the emotion of *love* comes from? If the ancient Greeks had ever thought of the theory of natural selection, they could have looked at the human institution of sexual romance, or parental love for that matter, and deduced in one flash that human beings had evolved—or at least derived tremendous Bayesian evidence for human evolution. Parental bonds and sexual romance clearly display the signature of evolutionary psychology—they're archetypal cases, in fact, so obvious we usually don't even see it."

Obert: "But love isn't just about reproduction—"

Subhan: "Of course not; individual organisms are adaptation-executers, not fitness-maximizers. But for something independent of humans, morality looks remarkably like godshatter of natural selection. Indeed, it is far too much coincidence for me to credit. Is happiness morally preferable to pain? What a coincidence! And if you claim that there is any emotion, any instinctive preference, any complex brain circuitry in humanity which was created by some external morality thingy and not natural selection, then you are infringing upon science and you will surely be torn to shreds—science has never needed to postulate anything but evolution to explain any feature of human psychology—"

Obert: "I'm *not* saying that humans got here by anything except evolution."

Subhan: "Then why does morality look so amazingly like a product of an evolved psychology?"

Obert: "I don't claim perfect access to moral truth; maybe, being human, I've made certain mistakes about morality—"

Subhan: "Say *that*—forsake love and life and happiness, and follow some useless damn trend of the universe or whatever—and you will lose every scrap of the moral normality that you once touted as your strong point. And I will be right here, asking, 'Why even bother?' It would be a pitiful mind indeed that demanded authoritative answers so strongly, that it would forsake all good things to have some authority beyond itself to follow."

Obert: "All right... then maybe the reason morality seems to bear certain similarities to our human constitutions, is that we could only perceive morality at all, if we happened, by luck, to evolve in consonance with it."

Subhan: "Horsemanure."

Obert: "Fine... you're right, that wasn't very plausible. Look, I admit you've driven me into quite a corner here. But even if there *were* nothing more to morality than preference, I would still prefer to act as morality were real. I mean, if it's all just preference, that way is as good as anything else—"

Subhan: "Now you're just trying to avoid facing reality! Like someone who says, 'If there is no Heaven or Hell, then I may as well still act as if God's going to punish me for sinning.'"

Obert: "That may be a good metaphor, in fact. Consider two theists, in the process of becoming atheists. One says, 'There is no Heaven or Hell, so I may as well cheat and steal, if I can get away without being caught, since there's no God to watch me.' And the other says, 'Even though there's no God, I intend to *pretend* that God is watching me, so that I can go on being a moral person.' Now they are both mistaken, but the first is straying much further from the path."

Subhan: "And what is the second one's flaw? *Failure to accept personal responsibility!*"

Obert: "Well, and I admit I find that a more compelling argument than anything else you have said. Probably because it is a moral argument, and it has always been morality, not metaethics, with which I claimed to be concerned. But even so, after our whole conversation, I still maintain that wanting to murder someone does not make murder *right.* Everything that you have said about preference is interesting, but it is ultimately *about* preference—about minds and what they are designed to desire—and not about this other thing that humans sometimes talk about, 'morality'. I can just ask Moore's Open Question: Why should I *care* about human preferences? What makes following human preferences *right*? By changing a mind, you can change what it prefers; you can even change what it *believes* to be right; but you cannot change what *is* right. Anything you talk about, that can be changed in this way, is not 'right-ness'."

Subhan: "So you take refuge in arguing from definitions?"

Obert: "You know, when I reflect on this whole argument, it seems to me that your position has the definite advantage when it comes to arguments about ontology and

reality and all that stuff—"

Subhan: "'*All that stuff*'? What else *is* there, besides reality?"

Obert: "Okay, the morality-as-preference viewpoint is a lot easier to shoehorn into a universe of quarks. But I still think the morality-as-given viewpoint has the advantage when it comes to, you know, the actual *morality* part of it—giving answers that are good in the sense of being *morally* good, not in the sense of being a good reductionist. Because, you know, there *are* such things as moral errors, there *is* moral progress, and you really *shouldn't* go around thinking that murder would be right if you wanted it to be right."

Subhan: "That sounds to me like the logical fallacy of appealing to consequences."

Obert: "Oh? Well, it sounds to *me* like an incomplete reduction—one that doesn't quite add up to normality."

Part of *The Metaethics Sequence*

Next post: "Where Recursive Justification Hits Bottom"

Previous post: "Is Morality Preference?"

Where Recursive Justification Hits Bottom

Why do I believe that the Sun will rise tomorrow?

Because I've seen the Sun rise on thousands of previous days.

Ah... but why do I believe the future will be like the past?

Even if I go past the mere surface observation of the Sun rising, to the apparently universal and exceptionless laws of gravitation and nuclear physics, then I am still left with the question: "Why do I believe this will also be true tomorrow?"

I could appeal to Occam's Razor, the principle of using the simplest theory that fits the facts... but why believe in Occam's Razor? Because it's been successful on past problems? But who says that this means Occam's Razor will work tomorrow?

And lo, the one said:

> "Science also depends on unjustified assumptions. Thus science is ultimately based on faith, *so don't you criticize me* for believing in [silly-belief-#238721]."

As I've previously observed:

> It's a most peculiar psychology—this business of "Science is based on faith too, so there!" Typically this is said by people who claim that faith is a *good* thing. Then why do they say "Science is based on faith too!" in that angry-triumphal tone, rather than as a compliment?

Arguing that you should be immune to criticism is rarely a good sign.

But this doesn't answer the legitimate philosophical dilemma: If every belief must be justified, and those justifications in turn must be justified, then how is the infinite recursion terminated?

And if you're allowed to end in something assumed-without-justification, then why aren't you allowed to assume *anything* without justification?

A similar critique is sometimes leveled against Bayesianism—that it requires assuming some prior—by people who apparently think that the problem of induction is a *particular* problem of Bayesianism, which you can avoid by using classical statistics. I will speak of this later, perhaps.

But first, let it be clearly admitted that the rules of Bayesian updating, do *not* of themselves solve the problem of induction.

Suppose you're drawing red and white balls from an urn. You observe that, of the first 9 balls, 3 are red and 6 are white. What is the probability that the next ball drawn will be red?

That depends on your prior beliefs about the urn. If you think the urn-maker generated a uniform random number between 0 and 1, and used that number as the fixed probability of each ball being red, then the answer is 4/11 (by Laplace's Law of

Succession). If you think the urn originally contained 10 red balls and 10 white balls, then the answer is 7/11.

Which goes to say that, with the right prior—or rather the wrong prior—the chance of the Sun rising tomorrow, would seem to go *down* with each succeeding day... if you were absolutely certain, a priori, that there was a great barrel out there from which, on each day, there was drawn a little slip of paper that determined whether the Sun rose or not; and that the barrel contained only a limited number of slips saying "Yes", and the slips were drawn without replacement.

There are possible minds in mind design space who have anti-Occamian and anti-Laplacian priors; they believe that simpler theories are less likely to be correct, and that the more often something happens, the less likely it is to happen again.

And when you ask these strange beings why they keep using priors that never seem to work in real life... they reply, "Because it's never worked for us before!"

Now, one lesson you might derive from this, is "Don't be born with a stupid prior." This is an amazingly helpful principle on many real-world problems, but I doubt it will satisfy philosophers.

Here's how I treat this problem myself: I try to approach questions like "Should I trust my brain?" or "Should I trust Occam's Razor?" as though they were *nothing special*—or at least, nothing special as deep questions go.

Should I trust Occam's Razor? Well, how well does (any particular version of) Occam's Razor seem to work in practice? What kind of probability-theoretic justifications can I find for it? When I look at the universe, does it seem like the kind of universe in which Occam's Razor would work well?

Should I trust my brain? Obviously not; it doesn't always work. But nonetheless, the human brain seems much more powerful than the most sophisticated computer programs I could consider trusting otherwise. How well does my brain work in practice, on which sorts of problems?

When I examine the causal history of my brain—its origins in natural selection—I find, on the one hand, all sorts of specific reasons for doubt; my brain was optimized to run on the ancestral savanna, not to do math. But on the other hand, it's also clear why, loosely speaking, it's possible that the brain really could work. Natural selection would have quickly eliminated brains so *completely* unsuited to reasoning, so *anti*-helpful, as anti-Occamian or anti-Laplacian priors.

So what I did in practice, does *not* amount to declaring a sudden halt to questioning and justification. I'm not halting the chain of examination at the point that I encounter Occam's Razor, or my brain, or some other unquestionable. The chain of examination continues—but it continues, unavoidably, using my current brain and my current grasp on reasoning techniques. *What else could I possibly use?*

Indeed, no matter *what* I did with this dilemma, it would be me doing it. Even if I trusted something else, like some computer program, it would be my own decision to trust it.

The technique of rejecting beliefs that have absolutely no justification, is in general an extremely important one. I sometimes say that the fundamental question of rationality is "Why do you believe what you believe?" I don't even want to say

something that *sounds* like it might allow a single exception to the rule that everything needs justification.

Which is, itself, a dangerous sort of motivation; you can't always avoid everything that might be risky, and when someone annoys you by saying something silly, you can't reverse that stupidity to arrive at intelligence.

But I would nonetheless emphasize the difference between saying:

> "Here is this assumption I cannot justify, which must be simply taken, and not further examined."

Versus saying:

> "Here the inquiry continues to examine this assumption, with the full force of my *present intelligence*—as opposed to the full force of something else, like a random number generator or a magic 8-ball—even though my present intelligence happens to be founded on this assumption."

Still... wouldn't it be nice if we could examine the problem of how much to trust our brains *without* using our current intelligence? Wouldn't it be nice if we could examine the problem of how to think, *without* using our current grasp of rationality?

When you phrase it *that* way, it starts looking like the answer might be "No".

E. T. Jaynes used to say that you must always use all the information available to you —he was a Bayesian probability theorist, and had to clean up the paradoxes other people generated when they used different information at different points in their calculations. The principle of "*Always put forth your true best effort*" has at least as much appeal as "*Never do anything that might look circular.*" After all, the alternative to putting forth your best effort is presumably doing less than your best.

But still... wouldn't it be nice if there were some way to justify using Occam's Razor, or justify predicting that the future will resemble the past, *without* assuming that those methods of reasoning which have worked on previous occasions are better than those which have continually failed?

Wouldn't it be nice if there were some chain of justifications that neither ended in an unexaminable assumption, nor was forced to examine itself under its own rules, but, instead, could be explained starting from absolute scratch to an ideal philosophy student of perfect emptiness?

Well, I'd certainly be interested, but I don't expect to see it done any time soon. I've argued elsewhere in several places against the idea that you can have a perfectly empty ghost-in-the-machine; there is no argument that you can explain to a rock.

Even if someone cracks the First Cause problem and comes up with *the actual reason the universe is simple, which does not itself presume a simple universe...* then I would still expect that the explanation could only be understood by a mindful listener, and not by, say, a rock. A listener that didn't start out already implementing modus ponens might be out of luck.

So, at the end of the day, what happens when someone keeps asking me "Why do you believe what you believe?"

At present, I start going around in a loop at the point where I explain, "I predict the future as though it will resemble the past on the simplest and most stable level of organization I can identify, because previously, this rule has usually worked to generate good results; and using the simple assumption of a simple universe, I can see *why* it generates good results; and I can even see how my brain might have evolved to be able to observe the universe with some degree of accuracy, if my observations are correct."

But then... haven't I just licensed *circular logic?*

Actually, I've just licensed *reflecting on your mind's degree of trustworthiness, using your current mind as opposed to something else.*

Reflection of this sort is, indeed, the reason we reject most circular logic in the first place. We want to have a coherent causal story about how our mind comes to know something, a story that explains how the process we used to arrive at our beliefs, is itself trustworthy. This is the essential demand behind the rationalist's fundamental question, "Why do you believe what you believe?"

Now suppose you write on a sheet of paper: "(1) Everything on this sheet of paper is true, (2) The mass of a helium atom is 20 grams." If that trick actually *worked in real life*, you would be able to know the true mass of a helium atom just by believing some circular logic which asserted it. Which would enable you to arrive at a true map of the universe sitting in your living room with the blinds drawn. Which would violate the second law of thermodynamics by generating information from nowhere. Which would not be a plausible story about how your mind could end up believing something true.

Even if you started out believing the sheet of paper, it would not seem that you had any reason for why the paper corresponded to reality. It would just be a miraculous coincidence that (a) the mass of a helium atom was 20 grams, and (b) the paper happened to say so.

Believing, in general, self-validating statement sets, does not seem like it should work to map external reality—when we *reflect on it as a causal story about minds*—using, of course, our *current* minds to do so.

But what about evolving to give more credence to simpler beliefs, and to believe that algorithms which have worked in the past are more likely to work in the future? *Even when* we reflect on this as a causal story of the origin of minds, it still seems like this could plausibly work to map reality.

And what about trusting reflective coherence in general? Wouldn't most possible minds, randomly generated and allowed to settle into a state of reflective coherence, be incorrect? Ah, but *we* evolved by natural selection; we were not generated randomly.

If trusting this argument seems worrisome to you, then forget about the problem of philosophical justifications, and ask yourself whether it's really truly true.

(You will, of course, use your own mind to do so.)

Is this the same as the one who says, "I believe that the Bible is the word of God, because the Bible says so"?

Couldn't they argue that their blind faith must also have been placed in them by God, and is therefore trustworthy?

In point of fact, when religious people finally come to reject the Bible, they do *not* do so by magically jumping to a non-religious state of pure emptiness, and then evaluating their religious beliefs in that non-religious state of mind, and then jumping back to a new state with their religious beliefs removed.

People go from being religious, to being non-religious, because even in a religious state of mind, doubt seeps in. They notice their prayers (and worse, the prayers of seemingly much worthier people) are not being answered. They notice that God, who speaks to them in their heart in order to provide seemingly consoling answers about the universe, is not able to tell them the hundredth digit of pi (which would be a lot more reassuring, if God's purpose were reassurance). They examine the story of God's creation of the world and damnation of unbelievers, and it doesn't seem to make sense even under their own religious premises.

Being religious doesn't make you less than human. Your brain still has the abilities of a human brain. The dangerous part is that being religious might stop you from *applying* those native abilities to your religion—stop you from *reflecting fully* on yourself. People don't heal their errors by resetting themselves to an ideal philosopher of pure emptiness and reconsidering all their sensory experiences from scratch. They heal themselves by becoming more willing to question their current beliefs, using more of the power of their current mind.

This is why it's important to distinguish between *reflecting on your mind using your mind* (it's not like you can use anything else) and *having an unquestionable assumption that you can't reflect on.*

"I believe that the Bible is the word of God, because the Bible says so." Well, if the Bible *were* an astoundingly reliable source of information about all other matters, if it had not said that grasshoppers had four legs or that the universe was created in six days, but had instead contained the Periodic Table of Elements centuries before chemistry—if the Bible had served us only well and told us only truth—then we might, in fact, be inclined to take seriously the additional statement in the Bible, that the Bible had been generated by God. We might not trust it entirely, because it could also be aliens or the Dark Lords of the Matrix, but it would at least be worth taking seriously.

Likewise, if everything *else* that priests had told us, turned out to be true, we might take more seriously their statement that faith had been placed in us by God and was a systematically trustworthy source—especially if people could divine the hundredth digit of pi by faith as well.

So the important part of appreciating the circularity of "I believe that the Bible is the word of God, because the Bible says so," is not so much that you are going to reject the idea of reflecting on your mind using your current mind. But, rather, that you realize that anything which calls into question the Bible's trustworthiness, also calls into question the Bible's assurance of its trustworthiness.

This applies to rationality too: if the future should cease to resemble the past—even on its lowest and simplest and most stable observed levels of organization—well, mostly, I'd be dead, because my brain's processes require a lawful universe where chemistry goes on working. But if somehow I survived, then I would have to start questioning the principle that the future should be predicted to be like the past.

But for now... what's the *alternative* to saying, "I'm going to believe that the future will be like the past on the most stable level of organization I can identify, because that's previously worked better for me than any other algorithm I've tried"?

Is it saying, "I'm going to believe that the future will *not* be like the past, because that algorithm has always failed before"?

At this point I feel obliged to drag up the point that rationalists are not out to win arguments with ideal philosophers of perfect emptiness; we are simply out to win. For which purpose we want to get as close to the truth as we can possibly manage. So at the end of the day, I embrace the principle: "Question your brain, question your intuitions, question your principles of rationality, *using the full current force of your mind, and doing the best you can do at every point.*"

If one of your current principles does come up wanting—according to your own mind's examination, since you can't step outside yourself—then change it! And then go back and look at things again, using your new improved principles.

The point is not to be reflectively consistent. The point is to win. But *if* you look at yourself and play to win, you are making yourself more reflectively consistent—that's what it means to "play to win" while "looking at yourself".

Everything, without exception, needs justification. Sometimes—unavoidably, as far as I can tell—those justifications will go around in reflective loops. I do think that reflective loops have a meta-character which should enable one to distinguish them, by common sense, from circular logics. But anyone seriously considering a circular logic in the first place, is probably out to lunch in matters of rationality; and will simply insist that their circular logic is a "reflective loop" even if it consists of a single scrap of paper saying "Trust me". Well, you can't always optimize your rationality techniques according to the sole consideration of preventing those bent on self-destruction from abusing them.

The important thing is to *hold nothing back* in your criticisms of how to criticize; nor should you regard the unavoidability of loopy justifications as a warrant of *immunity from questioning*.

Always apply full force, whether it loops or not—do the best you can possibly do, whether it loops or not—and play, ultimately, to win.

My Kind of Reflection

In "Where Recursive Justification Hits Bottom", I concluded that it's okay to use induction to reason about the probability that induction will work in the future, given that it's worked in the past; or to use Occam's Razor to conclude that the simplest explanation for why Occam's Razor works is that the universe itself is fundamentally simple.

Now I am far from the first person to consider reflective application of reasoning principles. Chris Hibbert compared my view to Bartley's Pan-Critical Rationalism (I was wondering whether that would happen). So it seems worthwhile to state what I see as the distinguishing features of my view of reflection, which may or may not happen to be shared by any other philosopher's view of reflection.

• All of my philosophy here *actually* comes from trying to figure out how to build a self-modifying AI that applies its own reasoning principles to itself in the process of rewriting its own source code. So whenever I talk about using induction to license induction, I'm *really* thinking about an inductive AI considering a rewrite of the part of itself that performs induction. If you wouldn't want the AI to rewrite its source code to not use induction, your philosophy had better not label induction as unjustifiable.

• One of the most powerful general principles I know for AI in general, is that the true Way generally turns out to be *naturalistic*—which for reflective reasoning, means treating transistors inside the AI, just as if they were transistors found in the environment; *not* an ad-hoc special case. This is the real source of my insistence in "Recursive Justification" that questions like "How well does my version of Occam's Razor work?" should be considered just like an ordinary question—or at least an ordinary very deep question. I strongly suspect that a correctly built AI, in pondering modifications to the part of its source code that implements Occamian reasoning, will not have to do anything special as it ponders—in particular, it shouldn't have to make a special effort to avoid using Occamian reasoning.

• I don't think that "reflective coherence" or "reflective consistency" should be considered as a desideratum in itself. As I said in the *Twelve Virtues* and the *Simple Truth,* if you make five accurate maps of the same city, then the maps will necessarily be consistent with each other; but if you draw one map by fantasy and then make four copies, the five will be consistent but not accurate. In the same way, no one is deliberately pursuing reflective consistency, and reflective consistency is not a special warrant of trustworthiness; the goal is to win. But anyone who pursues the goal of winning, using their current notion of winning, and modifying their own source code, will end up reflectively consistent as a side effect—just like someone continually striving to improve their map of the world should find the parts becoming more consistent among themselves, as a side effect. If you put on your AI goggles, then the AI, rewriting its own source code, is not trying to make itself "reflectively consistent"—it is trying to optimize the expected utility of its source code, and it happens to be doing this using its current mind's anticipation of the consequences.

• One of the ways I license using induction and Occam's Razor to consider "induction" and "Occam's Razor", is by appealing to E. T. Jaynes's principle that we should always use all the information available to us (computing power permitting) in a calculation. If you think induction works, then you should use it in order to use your maximum power, including when you're thinking about induction.

• In general, I think it's valuable to distinguish a defensive posture where you're imagining how to justify your philosophy to a philosopher that questions you, from an aggressive posture where you're trying to get as close to the truth as possible. So it's not that being suspicious of Occam's Razor, but using your current mind and intelligence to inspect it, shows that you're being *fair* and *defensible* by questioning your foundational beliefs. Rather, the reason why you would inspect Occam's Razor is to see if you could improve your application of it, or if you're worried it might really be wrong. I tend to deprecate mere dutiful doubts.

• If you run around inspecting your foundations, I expect you to actually improve them, not just dutifully investigate. Our brains are built to assess "simplicity" in a certain intuitive way that makes Thor sound simpler than Maxwell's Equations as an explanation for lightning. But, having gotten a better look at the way the universe really works, we've concluded that differential equations (which few humans master) are actually *simpler* (in an information-theoretic sense) than heroic mythology (which is how most tribes explain the universe). This being the case, we've tried to import our notions of Occam's Razor into math as well.

• On the other hand, the improved foundations should still add up to normality; 2 + 2 should still end up equalling 4, not something new and amazing and exciting like "fish".

• I think it's very important to distinguish between the questions "Why does induction work?" and "Does induction work?" The reason *why the universe itself is regular* is still a mysterious question unto us, for now. Strange speculations here may be temporarily needful. But on the other hand, if you start claiming that the universe *isn't actually regular,* that the answer to "Does induction work?" is "No!", then you're wandering into 2 + 2 = 3 territory. You're trying too hard to make your philosophy interesting, instead of correct. An inductive AI asking what probability assignment to make on the next round is asking "*Does* induction work?", and this is the question that it may answer by inductive reasoning. If you ask "*Why* does induction work?" then answering "Because induction works" is circular logic, and answering "Because I believe induction works" is magical thinking.

• I don't think that going around in a loop of justifications through the meta-level is the same thing as circular logic. I think the notion of "circular logic" applies within the object level, and is something that is definitely bad and forbidden, on the object level. Forbidding *reflective coherence* doesn't sound like a good idea. But I haven't yet sat down and formalized the exact difference—my reflective theory is something I'm trying to work out, not something I have in hand.

The Genetic Fallacy

In lists of logical fallacies, you will find included “the genetic fallacy”—the fallacy of attacking a belief based on someone’s causes for believing it.

This is, at first sight, a very strange idea—if the causes of a belief do not determine its systematic reliability, what does? If Deep Blue advises us of a chess move, we trust it based on our understanding of the *code* that searches the game tree, being unable to evaluate the actual game tree ourselves. What could license any probability assignment as “rational,” except that it was produced by some systematically reliable process?

Articles on the genetic fallacy will tell you that genetic reasoning is not always a fallacy—that the origin of evidence *can* be relevant to its evaluation, as in the case of a trusted expert. But other times, say the articles, it *is* a fallacy; the chemist Kekulé first saw the ring structure of benzene in a dream, but this doesn’t mean we can never trust this belief.

So sometimes the genetic fallacy is a fallacy, and sometimes it’s not?

The genetic fallacy is formally a fallacy, because the *original cause* of a belief is not the same as its *current justificational status*, the sum of all the support and antisupport *currently* known.

Yet we change our minds less often than we think. Genetic accusations have a force among humans that they would not have among ideal Bayesians.

Clearing your mind is a *powerful heuristic* when you’re faced with new suspicion that many of your ideas may have come from a flawed source.

Once an idea gets into our heads, it’s not always easy for evidence to root it out. Consider all the people out there who grew up believing in the Bible; later came to reject (on a deliberate level) the idea that the Bible was written by the hand of God; and who nonetheless think that the Bible is full of indispensable ethical wisdom. They have failed to clear their minds; they could do significantly better by doubting anything the Bible said *because the Bible said it*.

At the same time, they would have to bear firmly in mind the principle that reversed stupidity is not intelligence; the goal is to genuinely shake your mind loose and do independent thinking, not to negate the Bible and let that be your algorithm.

Once an idea gets into your head, you tend to find support for it everywhere you look—and so when the original source is suddenly cast into suspicion, you would be very wise indeed to suspect all the leaves that originally grew on that branch . . .

If you can! It’s not easy to clear your mind. It takes a convulsive effort to *actually reconsider* , instead of letting your mind fall into the pattern of rehearsing cached arguments. “It ain’t a true crisis of faith unless things could just as easily go either way,” said Thor Shenkel.

You should be *extremely suspicious* if you have many ideas suggested by a source that you now know to be untrustworthy, but by golly, it seems that all the ideas still ended up being right—the Bible being the obvious archetypal example.

On the other hand . . . there's such a thing as sufficiently clear-cut evidence, that it no longer significantly matters where the idea originally came from. Accumulating that kind of clear-cut evidence is what Science is all about. It doesn't matter any more that Kekulé first saw the ring structure of benzene in a dream—it wouldn't matter if we'd found the hypothesis to test by generating random computer images, or from a spiritualist revealed as a fraud, or even from the Bible. The ring structure of benzene is pinned down by enough experimental evidence to make the source of the suggestion irrelevant.

In the absence of such clear-cut evidence, then you do need to pay attention to the original sources of ideas—to give experts more credence than layfolk, if their field has earned respect—to suspect ideas you originally got from suspicious sources—to distrust those whose motives are untrustworthy, *if* they cannot present arguments independent of their own authority.

The genetic fallacy is a *fallacy* when there exist justifications *beyond* the genetic fact asserted, but the genetic accusation is presented as if it settled the issue. Hal Finney suggests that we call correctly appealing to a claim's origins "the genetic heuristic."[1]

Some good rules of thumb (for humans):

- Be suspicious of genetic accusations against beliefs that you dislike, especially if the proponent claims justifications beyond the simple authority of a speaker. "Flight is a religious idea, so the Wright Brothers must be liars" is one of the classically given examples.
- By the same token, don't think you can get good information about a technical issue just by sagely psychoanalyzing the personalities involved and their flawed motives. If technical arguments exist, they get priority.
- When new suspicion is cast on one of your fundamental sources, you really *should* doubt all the branches and leaves that grew from that root. You are not licensed to reject them outright as conclusions, because reversed stupidity is not intelligence, but . . .
- Be extremely suspicious if you find that you still believe the early suggestions of a source you later rejected.

---

[1]Source: http://lesswrong.com/lw/s3/the_genetic_fallacy/lls.

Fundamental Doubts

Followup to: The Genetic Fallacy, Where Recursive Justification Hits Bottom

Yesterday I said that—because humans are not perfect Bayesians—the genetic fallacy is not entirely a fallacy; when new suspicion is cast on one of your fundamental sources, you really *should* doubt all the branches and leaves of that root, even if they *seem* to have accumulated new evidence in the meanwhile.

This is one of the most difficult techniques of rationality (on which I will separately post, one of these days). Descartes, setting out to "doubt, insofar as possible, all things", ended up trying to prove the existence of God—which, if he wasn't a secret atheist trying to avoid getting burned at the stake, is pretty pathetic. It is *hard* to doubt an idea to which we are deeply attached; our mind naturally reaches for cached thoughts and rehearsed arguments.

But today's post concerns a different kind of difficulty—the case where the doubt is so deep, of a source so fundamental, that you *can't* make a true fresh beginning.

Case in point: Remember when, in the *The Matrix*, Morpheus told Neo that the machines were harvesting the body heat of humans for energy, and liquefying the dead to feed to babies? I suppose you thought something like, "Hey! That violates the second law of thermodynamics."

Well, it *does* violate the second law of thermodynamics. But if the *Matrix*'s makers had cared about the flaw once it was pointed out to them, they could have fixed the plot hole in any of the sequels, in fifteen seconds, this easily:

> Neo: "Doesn't harvesting human body heat for energy, violate the laws of thermodynamics?"
>
> Morpheus: "Where'd you learn about thermodynamics, Neo?"
>
> Neo: "In school."
>
> Morpheus: "Where'd you go to school, Neo?"
>
> Neo: "Oh."
>
> Morpheus: "The machines tell elegant lies."

Now, mind you, I am not saying that this excuses the original mistake in the script. When my mind generated this excuse, it came clearly labeled with that warning sign of which I have spoken, "Tada! Your mind can generate an excuse for *anything!*" You do not need to tell me that my plot-hole-patch is a nitwit idea, I am well aware of that...

...but, in point of fact, if you woke up out of a virtual reality pod one day, you *would* have to suspect all the physics you knew. Even if you looked down and saw that you had hands, you couldn't rely on there being blood and bone inside them. Even if you looked up and saw stars, you couldn't rely on their being trillions of miles away. And even if you found yourself thinking, you couldn't rely on your head containing a brain.

You could still try to doubt, even so. You could do your best to unwind your thoughts past every lesson in school, every science paper read, every sensory experience, every math proof whose seeming approval by other mathematicians might have been choreographed to conceal a subtle flaw...

But suppose you discovered that you were a computer program and that the Dark Lords of the Matrix were actively tampering with your thoughts.

Well... in that scenario, you're pretty much screwed, I'd have to say.

Descartes vastly underestimated the powers of an infinitely powerful deceiving demon when he supposed he could trust "I think therefore I am." Maybe that's just what *they* want you to think. Maybe *they* just inserted that conclusion into your mind with a memory of it seeming to have an irrefutable chain of logical support, along with some peer pressure to label it "unquestionable" just like all your friends.

(Personally, I don't trust "I think therefore I am" even in real life, since it contains a term "am" whose meaning I find confusing, and I've learned to spread my confidence intervals very widely in the presence of basic confusion. As for absolute certainty, don't be silly.)

Every memory of justification could be faked. Every feeling of support could be artificially induced. Modus ponens could be a lie. Your concept of "rational justification"—not just your specific concept, but your notion that any such thing exists at all—could have been manufactured to mislead you. Your trust in Reason itself could have been inculcated to throw you off the trail.

So you might as well not think about the possibility that you're a brain with choreographed thoughts, because there's nothing you can do about it...

Unless, of course, that's what *they* want you to think.

Past a certain level of doubt, it's not possible to start over fresh. There's nothing you can *un*assume to find some firm rock on which to stand. You cannot unwind yourself into a perfectly empty and perfectly reliable ghost in the machine.

This level of meta-suspicion should be a rare occasion. For example, suspecting that all academic science is an organized conspiracy, should not run into anything like these meta-difficulties. Certainly, someone does not get to plead that unwinding past the Bible is impossible because it is too foundational; atheists walk the Earth without falling into comas. Remember, when Descartes tried to outwit an infinitely powerful deceiving demon, he first tried to make himself absolutely certain of a highly confusing statement, and then proved the existence of God. Consider that a caution about what you try to claim is "too basic for a fresh beginning". And even basic things can still be doubted, it is only that we use our untrustworthy brains to doubt them.

Or consider the case of our existence as evolved brains. Natural selection isn't trustworthy, and we have specific reason to suspect it. We know that evolution is stupid. We know many specific ways in which our human brains fail, taken beyond the savanna. But you *can't* clear your mind of evolutionary influences and start over. It would be like deciding that you don't trust neurons, so you're going to clear your mind of brains.

And evolution certainly gets a chance to influence every single thought that runs through your mind! It is the very reason why you exist as a thinker, rather than a

lump of carbon—and that doesn't mean evolution summoned a ghost-in-the-machine into you; it *designed* the ghost. If you learn culture, it is because you were built to learn culture.

But in fact, we *don't* run into unmanageable meta-trouble in trying to come up with specific patches for specific known evolved biases. And evolution is stupid, so even though it has set up self-deceptive circuits in us, these circuits are not infinitely difficult to comprehend and outwit.

Or so it seems! But it really *does* seem that way, on reflection.

There is no button you can press to rewind past your noisy brain, and become a perfectly reliable ghost of perfect emptiness. That's not just because your brain *is you*. It's also because you can't unassume things like modus ponens or belief updating. You can unassume them as explicit premises for deliberate reasoning—a hunter-gatherer has no *explicit* concept of modus ponens—but you can't delete the actual dynamics (and all their products!)

So, in the end, I think we must allow the use of brains to think about thinking; and the use of evolved brains to think about evolution; and the use of inductive brains to think about induction; and the use of brains with an Occam prior to think about whether the universe appears to be simple; for these things we really *cannot* unwind entirely, even when we have reason to distrust them. Strange loops through the meta level, I think, are not the same as circular logic.

Part of *The Metaethics Sequence*

Next post: "Rebelling Within Nature"

Previous post: "My Kind of Reflection"

Rebelling Within Nature

Followup to: Fundamental Doubts, Where Recursive Justification Hits Bottom, No Universally Compelling Arguments, Joy in the Merely Real, Evolutionary Psychology

> "Let us understand, once and for all, that the ethical progress of society depends, not on imitating the cosmic process, still less in running away from it, but in combating it."
>
> —T. H. Huxley ("Darwin's bulldog", early advocate of evolutionary theory)

There is a quote from some Zen Master or other, who said something along the lines of:

> "Western man believes that he is rebelling against nature, but he does not realize that, in doing so, he is acting according to nature."

The Reductionist Masters of the West, strong in their own Art, are not so foolish; they *do* realize that they always act within Nature.

You can narrow your focus and rebel against a *facet* of existing Nature—polio, say—but in so doing, you act within the *whole* of Nature. The syringe that carries the polio vaccine is forged of atoms; our minds, that understood the method, embodied in neurons. If Jonas Salk had to fight laziness, he fought something that evolution instilled in him—a reluctance to work that conserves energy. And he fought it *with* other emotions that natural selection also inscribed in him: feelings of friendship that he extended to humanity, heroism to protect his tribe, maybe an explicit desire for fame that he never acknowledged to himself—who knows? (I haven't actually read a biography of Salk.)

The point is, you can't fight Nature from beyond Nature, only from within it. There is no acausal fulcrum on which to stand outside reality and move it. There is no ghost of perfect emptiness by which you can judge your brain from outside your brain. You can fight the cosmic process, but only by recruiting other abilities that evolution originally gave to you.

And if you fight one emotion within yourself—looking upon your own nature, and judging yourself less than you think should be—saying perhaps, "I should not *want* to kill my enemies"—then you make *that* judgment, by...

How exactly *does* one go about rebelling against one's own goal system?

From within it, naturally.

This is perhaps *the* primary thing that I didn't quite understand as a teenager.

At the age of fifteen (fourteen?), I picked up a copy of TIME magazine and read an article on evolutionary psychology. It seemed like one of the most massively obvious-in-retrospect ideas I'd ever heard. I went on to read *The Moral Animal* by Robert Wright. And later *The Adapted Mind*—but from the perspective of personal epiphanies, *The Moral Animal* pretty much did the job.

I'm reasonably sure that if I had not known the basics of evolutionary psychology from my teenage years, I would not currently exist as the Eliezer Yudkowsky you know.

Indeed, let me drop back a bit further:

At the age of... I think it was nine... I discovered the truth about sex by looking it up in my parents' home copy of the Encyclopedia Britannica (stop that laughing). Shortly after, I learned a good deal more by discovering where my parents had hidden the secret *15th volume* of my long-beloved Childcraft series. I'd been avidly reading the first 14 volumes—some of them, anyway—since the age of five. But the 15th volume wasn't meant for me—it was the "Guide for Parents".

The 15th volume of Childcraft described the life cycle of children. It described the horrible confusion of the teenage years—teenagers experimenting with alcohol, with drugs, with unsafe sex, with reckless driving, the hormones taking over their minds, the overwhelming importance of peer pressure, the tearful accusations of "You don't love me!" and "I hate you!"

I took one look at that description, at the tender age of nine, and said to myself in quiet revulsion, *I'm not going to do that.*

And I didn't.

My teenage years were not untroubled. But I didn't do any of the things that the *Guide to Parents* warned me against. I didn't drink, drive, drug, lose control to hormones, pay any attention to peer pressure, or ever once think that my parents didn't love me.

In a safer world, I would have wished for my parents to have hidden that book better.

But in this world, which needs me as I am, I don't regret finding it.

I still rebelled, of course. I rebelled against the rebellious nature the *Guide to Parents* described to me. That was part of how I defined my identity in my teenage years —"I'm not doing the standard stupid stuff." Some of the time, this just meant that I invented amazing new stupidity, but in fact that *was* a major improvement.

Years later, *The Moral Animal* made suddenly obvious the *why* of all that disastrous behavior I'd been warned against. Not that Robert Wright pointed any of this out explicitly, but it was obvious given the elementary concept of evolutionary psychology:

Physiologically adult humans are not meant to spend an additional 10 years in a school system; their brains map that onto "I have been assigned low tribal status". And so, of course, they plot rebellion—accuse the existing tribal overlords of corruption—plot perhaps to split off their own little tribe in the savanna, not realizing that this is impossible in the Modern World. The teenage males map their own fathers onto the role of "tribal chief"...

Echoes in time, thousands of repeated generations in the savanna carving the pattern, ancient repetitions of form, reproduced in the present in strange twisted mappings, across genes that didn't know anything had changed...

The world grew older, of a sudden.

And I'm not going to go into the evolutionary psychology of "teenagers" in detail, not now, because that would deserve its own post.

But when I read *The Moral Animal*, the world suddenly acquired *causal depth.* Human emotions existed for *reasons,* they weren't just unexamined givens. I might previously have questioned whether an emotion was appropriate to its circumstance—whether it made sense to hate your parents, if they did really love you—but I wouldn't have thought, before then, to judge *the existence of hatred as an evolved emotion.*

And then, having come so far, and having avoided with instinctive ease all the classic errors that evolutionary psychologists are traditionally warned against—I was never once tempted to confuse evolutionary causation with psychological causation—I went wrong at the last turn.

The echo in time that was teenage psychology was obviously wrong and stupid—a *distortion* in the way things should be—so clearly you were supposed to unwind past it, compensate in the opposite direction or disable the feeling, to arrive at the correct answer.

It's hard for me to remember exactly what I was thinking in this era, but I think I tended to focus on one facet of human psychology at any given moment, trying to unwind myself a piece at a time. IIRC I did think, in full generality, "Evolution is bad; the effect it has on psychology is bad." (Like it had some kind of "effect" that could be isolated!) But somehow, I managed not to get to "Evolutionary psychology is the cause of altruism; altruism is bad."

It was easy for me to see all sorts of *warped* altruism as having been *warped by* evolution.

People who wanted to trust themselves with power, for the good of their tribe—*that* had an obvious evolutionary explanation; it was, therefore, a distortion to be corrected.

People who wanted to be altruistic in ways their friends would approve of—obvious evolutionary explanation; therefore a distortion to be corrected.

People who wanted to be altruistic in a way that would optimize their fame and repute—obvious evolutionary distortion to be corrected.

People who wanted to help only their family, or only their nation—acting out ancient selection pressures on the savanna; move past it.

But the fundamental will to help people?

Well, the notion of *that* being merely evolved, was something that, somehow, I managed to *never quite accept*. Even though, in retrospect, the causality is just as obvious as teen revolutionism.

IIRC, I did think something along the lines of: "Once you unwind past evolution, then the true morality isn't likely to contain a clause saying, 'This person matters but this person doesn't', so everyone should matter equally, so you should be as eager to help others as help yourself." And so I thought that even if the emotion of altruism had merely evolved, it was a right emotion, and I should keep it.

But why think that people mattered at all, if you were trying to unwind past all evolutionary psychology? Why think that it was better for people to be happy than sad, rather than the converse?

If I recall correctly, I *did* ask myself that, and sort of waved my hands mentally and said, "It just seems like one of the best guesses—I mean, I don't know that people are valuable, but I can't think of what else could be."

This is the Avoiding Your Belief's Real Weak Points / Not Spontaneously Thinking About Your Belief's Most Painful Weaknesses antipattern in full glory: Get just far enough to place yourself on the first fringes of real distress, and then stop thinking.

And also the antipattern of trying to unwind past everything that is causally responsible for your existence as a mind, to arrive at a perfectly reliable ghost of perfect emptiness.

Later, having also seen others making similar mistakes, it seems to me that the general problem is an illusion of mind-independence that comes from picking something that appeals to you, while still seeming philosophically simple.

As if the appeal to you, of the moral argument, weren't still a feature of your particular point in mind design space.

As if there weren't still an ordinary and explicable causal history behind the appeal, and your selection of that particular principle.

As if, by making things philosophically simpler-seeming, you could enhance their appeal to a ghost-in-the-machine who would hear your justifications starting from scratch, as fairness demands.

As if your very sense of simplicity were not an aesthetic sense inscribed in you by evolution.

As if your very intuitions of "moral argument" and "justification", were not an architecture-of-reasoning inscribed in you by natural selection, and just as causally explicable as any other feature of human psychology...

You can't throw away evolution, and end up with a perfectly moral creature that humans would have been, if only we had never evolved; that's really not how it works.

Why accept intuitively appealing arguments about the nature of morality, rather than intuitively unappealing ones, if you're going to distrust everything in you that ever evolved?

Then what *is* right? What *should* we do, having been inscribed by a blind mad idiot god whose incarnation-into-reality takes the form of millions of years of ancestral murder and war?

But even this question—every fragment of it—the notion that a blind mad idiocy is an ugly property for a god to have, or that murder is a poisoned well of order, even the words "right" and "should"—all a phenomenon within nature. All traceable back to debates built around arguments appealing to intuitions that evolved in me.

You can't jump out of the system. You really can't. Even *wanting* to jump out of the system—the sense that something isn't justified "just because it evolved"—is something that you feel from *within* the system. Anything you might try to use to jump—any sense of what morality *should* be like, if you could unwind past evolution—is *also* there as a causal result of evolution.

Not everything we think about morality is *directly* inscribed by evolution, of course. We have values that we got from our parents teaching them to us as we grew up; after it won out in a civilizational debate conducted with reference to other moral principles; that were themselves argued into existence by appealing to built-in emotions; using an architecture-of-interpersonal-moral-argument that evolution burped into existence.

It all goes back to evolution. This doesn't just include things like instinctive concepts of fairness, or empathy, it includes the whole notion of arguing morals as if they were propositional beliefs. Evolution created within you that *frame of reference* within which you can *formulate the concept* of moral questioning. Including questioning evolution's fitness to create our moral frame of reference. If you *really* try to unwind outside the system, you'll unwind your unwinders.

That's what I didn't quite get, those years ago.

I do plan to dissolve the cognitive confusion that makes words like "right" and "should" seem difficult to grasp. I've been working up to that for a while now.

But I'm not there yet, and so, for now, I'm going to jump ahead and peek at an answer I'll only later be able to justify as moral philosophy:

Embrace reflection. You can't unwind to emptiness, but you can bootstrap from a starting point.

Go on morally questioning the existence (and not just appropriateness) of emotions. But don't treat the mere fact of their *having evolved* as a reason to reject them. Yes, I know that "X evolved" doesn't seem like a good justification for having an emotion; but don't let that be a reason to reject X, any more than it's a reason to accept it. Hence the post on the Genetic Fallacy: causation is conceptually distinct from justification. If you try to apply the Genetic Accusation to automatically convict and expel your *genes,* you're going to run into foundational trouble—so don't!

Just ask if the emotion is justified—don't treat its evolutionary cause as proof of mere distortion. Use your current mind to examine the emotion's pluses and minuses, without being ashamed; *use your full strength of morality.*

Judge emotions *as emotions*, not as evolutionary relics. When you say, "motherly love outcompeted its alternative alleles because it protected children that could carry the allele for motherly love", this is only a *cause,* not a sum of all moral arguments. The evolutionary psychology may grant you helpful insight into the pattern and process of motherly love, but it neither justifies the emotion as natural, nor convicts it as coming from an unworthy source. You don't make the Genetic Accusation either way. You just, y'know, think about motherly love, and ask yourself if it seems like a good thing or not; considering its effects, not its source.

You tot up the balance of moral justifications, using your current mind—without worrying about the fact that the entire debate takes place within an evolved framework.

That's the moral normality to which my yet-to-be-revealed moral philosophy will add up.

And if, in the meanwhile, it seems to you like I've just proved that there is no morality... well, I haven't proved any such thing. But, meanwhile, just ask yourself if

you might want to help people even if there were no morality. If you find that the answer is yes, then you will later discover that you discovered morality.

Part of *The Metaethics Sequence*

Next post: "Probability is Subjectively Objective"

Previous post: "Fundamental Doubts"

Probability is Subjectively Objective

Followup to: Probability is in the Mind

> "Reality is that which, when you stop believing in it, doesn't go away."
> —Philip K. Dick

There are two kinds of Bayesians, allegedly. Subjective Bayesians believe that "probabilities" are degrees of uncertainty existing in our minds; if you are uncertain about a phenomenon, that is a fact about your state of mind, not a property of the phenomenon itself; probability theory constrains the logical coherence of uncertain beliefs. Then there are objective Bayesians, who... I'm not quite sure what it means to be an "objective Bayesian"; there are multiple definitions out there. As best I can tell, an "objective Bayesian" is anyone who uses Bayesian methods and isn't a subjective Bayesian.

If I recall correctly, E. T. Jaynes, master of the art, once described himself as a subjective-objective Bayesian. Jaynes certainly believed very firmly that probability was in the mind; Jaynes was the one who coined the term Mind Projection Fallacy. But Jaynes also didn't think that this implied a license to make up whatever priors you liked. There was only one *correct* prior distribution to use, given your state of partial information at the start of the problem.

How can something be in the mind, yet still be objective?

It appears to me that a good deal of philosophical maturity consists in being able to keep separate track of nearby concepts, without mixing them up.

For example, to understand evolutionary psychology, you have to keep separate track of the psychological purpose of an act, and the evolutionary pseudo-purposes of the adaptations that execute as the psychology; this is a common failure of newcomers to evolutionary psychology, who read, misunderstand, and thereafter say, "You think you love your children, but you're just trying to maximize your fitness!"

What is it, exactly, that the terms "subjective" and "objective", mean? Let's say that I hand you a sock. Is it a subjective or an objective sock? You believe that 2 + 3 = 5. Is *your belief* subjective or objective? What about two plus three *actually* equaling five—is that subjective or objective? What about a specific act of adding two apples and three apples and getting five apples?

I don't intend to confuse you in shrouds of words; but I do mean to point out that, while you may feel that you know very well what is "subjective" or "objective", you might find that you have a bit of trouble saying out loud what those words mean.

Suppose there's a calculator that computes "2 + 3 = 5". We punch in "2", then "+", then "3", and lo and behold, we see "5" flash on the screen. We accept this as *evidence* that 2 + 3 = 5, but we wouldn't say that the calculator's physical output *defines* the answer to the question 2 + 3 = ?. A cosmic ray could strike a transistor, which might give us misleading evidence and cause us to believe that 2 + 3 = 6, but it wouldn't affect the *actual* sum of 2 + 3.

Which proposition is common-sensically true, but philosophically interesting: while we can easily point to the physical location of a symbol on a calculator screen, or observe

the result of putting two apples on a table followed by another three apples, it is rather harder to track down the whereabouts of 2 + 3 = 5. (Did you look in the garage?)

But let us leave aside the question of *where* the fact 2 + 3 = 5 is located—in the universe, or somewhere else—and consider the assertion that the proposition is "objective". If a cosmic ray strikes a calculator and makes it output "6" in response to the query "2 + 3 = ?", and you add two apples to a table followed by three apples, then you'll still see five apples on the table. If you do the calculation in your own head, expending the necessary computing power—we assume that 2 + 3 is a very difficult sum to compute, so that the answer is not immediately obvious to you—then you'll get the answer "5". So the cosmic ray strike didn't change anything.

And similarly—exactly similarly—what if a cosmic ray strikes a neuron inside your brain, causing you to compute "2 + 3 = 7"? Then, adding two apples to three apples, you will expect to see seven apples, but instead you will be surprised to see five apples.

If instead we found that no one was ever mistaken about addition problems, and that, moreover, you could change the answer by an act of will, then we might be tempted to call addition "subjective" rather than "objective". I am not saying that this is *everything* people mean by "subjective" and "objective", just pointing to one aspect of the concept. One might summarize this aspect thus: "If you can change something by thinking differently, it's subjective; if you can't change it by anything you do strictly inside your head, it's objective."

Mind is not magic. Every act of reasoning that we human beings carry out, is *computed within* some particular human brain. But not every computation is *about* the state of a human brain. Not every thought that you think is *about* something that can be changed by thinking. Herein lies the opportunity for confusion-of-levels. The quotation is not the referent. If you are going to consider thoughts as referential at all—if not, I'd like you to explain the mysterious correlation between my thought "2 + 3 = 5" and the observed behavior of apples on tables—then, while the quoted thoughts will always change with thoughts, the referents *may or may not* be entities that change with changing human thoughts.

The calculator computes "What is 2 + 3?", not "What does this calculator compute as the result of 2 + 3?" The answer to the former question is 5, but if the calculator were to ask the latter question instead, the result could self-consistently be anything at all! If the calculator returned 42, then indeed, "What does this calculator compute as the result of 2 + 3?" would in fact be 42.

So just because a computation takes place inside your brain, does not mean that the computation *explicitly mentions* your brain, that it has your brain as a *referent,* any more than the calculator mentions the calculator. The calculator does not attempt to contain a representation of itself, only of numbers.

Indeed, in the most straightforward implementation, the calculator that asks "What does this calculator compute as the answer to the query 2 + 3 = ?" will *never* return a result, just simulate itself simulating itself until it runs out of memory.

But if you punch the keys "2", "+", and "3", and the calculator proceeds to compute "What do I output when someone punches '2 + 3'?", the resulting computation does have one interesting characteristic: the *referent* of the computation is highly

subjective, since it depends on the computation, and can be made to be anything just by changing the computation.

Is probability, then, subjective or objective?

Well, probability is computed within human brains or other calculators. A probability is a state of partial information that is possessed by you; if you flip a coin and press it to your arm, the coin is showing heads or tails, but you assign the probability 1/2 until you reveal it. A friend, who got a tiny but not fully informative peek, might assign a probability of 0.6.

So can you make the probability of winning the lottery be anything you like?

Forget about many-worlds for the moment—you should almost always be able to forget about many-worlds—and pretend that you're living in a single Small World where the lottery has only a single outcome. You will nonetheless have a need to call upon probability. Or if you prefer, we can discuss the ten trillionth decimal digit of pi, which I believe is not yet known. (If you are foolish enough to refuse to assign a probability distribution to this entity, you might pass up an excellent bet, like betting $1 to win $1000 that the digit is not 4.) Your uncertainty is a state of your mind, of partial information that you possess. Someone else might have different information, complete or partial. And the entity itself will only ever take on a single value.

So can you make the probability of winning the lottery, or the probability of the ten trillionth decimal digit of pi equaling 4, be anything you like?

You might be tempted to reply: "Well, since I *currently* think the probability of winning the lottery is one in a hundred million, then obviously, I will *currently* expect that assigning any other probability than this to the lottery, will decrease my expected log-score—or if you prefer a decision-theoretic formulation, I will expect this modification to myself to decrease expected utility. So, obviously, I will not choose to modify my probability distribution. It wouldn't be reflectively coherent."

So reflective coherency is the goal, is it? Too bad you weren't born with a prior that assigned probability 0.9 to winning the lottery! Then, by exactly the same line of argument, you wouldn't want to assign any probability except 0.9 to winning the lottery. And you would still be reflectively coherent. And you would have a 90% probability of winning millions of dollars! Hooray!

"No, then I would *think* I had a 90% probability of winning the lottery, but *actually,* the probability would only be one in a hundred million."

Well, of course *you* would be expected to say that. And if you'd been born with a prior that assigned 90% probability to your winning the lottery, you'd consider an alleged probability of 10^-8, and say, "No, then I would *think* I had almost no probability of winning the lottery, but *actually,* the probability would be 0.9."

"Yeah? Then just modify your probability distribution, and buy a lottery ticket, and then wait and see what happens."

What happens? Either the ticket will win, or it won't. That's what will happen. We won't get to see that some particular probability was, in fact, the exactly right probability to assign.

"Perform the experiment a hundred times, and—"

Okay, let's talk about the ten trillionth digit of pi, then. Single-shot problem, no "long run" you can measure.

Probability is subjectively objective: Probability exists in your mind: if you're ignorant of a phenomenon, that's an attribute of you, not an attribute of the phenomenon. Yet it will seem to you that you can't change probabilities by wishing.

You could make yourself compute something *else,* perhaps, *rather than* probability. You could compute "What do I say is the probability?" (answer: anything you say) or "What do I wish were the probability?" (answer: whatever you wish) but these things are not the *probability,* which is subjectively objective.

The thing about subjectively objective quantities is that they *really do* seem objective to you. You don't look them over and say, "Oh, well, of course I don't want to modify my own probability estimate, because no one can just modify their probability estimate; but if I'd been born with a different prior I'd be saying something different, and I wouldn't want to modify that either; and so none of us is superior to anyone else." That's the way a subjectively *subjective* quantity would seem.

No, it will seem to you that, if the lottery sells a hundred million tickets, and you don't get a peek at the results, then the probability of a ticket winning, *is* one in a hundred million. And that you could be born with different priors but that wouldn't give you any better odds. And if there's someone next to you saying the same thing about *their* 90% probability estimate, you'll just shrug and say, "Good luck with that." You won't expect them to *win.*

Probability is subjectively *really* objective, not just subjectively *sort of* objective.

Jaynes used to recommend that no one ever write out an unconditional probability: That you never, ever write simply P(A), but always write P(A|I), where I is your prior information. I'll use Q instead of I, for ease of reading, but Jaynes used I. Similarly, one would not write P(A|B) for the posterior probability of A given that we learn B, but rather P(A|B,Q), the probability of A given that we learn B and had background information Q.

This is good advice in a purely pragmatic sense, when you see how many false "paradoxes" are generated by accidentally using different prior information in different places.

But it also makes a deep philosophical point as well, which I never saw Jaynes spell out explicitly, but I think he would have approved: *there is no such thing as a probability that isn't in any mind.* Any mind that takes in evidence and outputs probability estimates of the next event, remember, can be viewed as a prior—so there is no probability without priors/minds.

You can't unwind the Q. You can't ask "What is the *unconditional* probability of our background information being true, P(Q)?" To make that estimate, you would still need *some* kind of prior. No way to unwind back to an ideal ghost of perfect emptiness...

You might argue that you and the lottery-ticket buyer do not really have a disagreement about *probability.* You say that the probability of the ticket winning the lottery is one in a hundred million given your prior, P(W|Q1) = 10^-8. The other fellow says the probability of the ticket winning given his prior is P(W|Q2) = 0.9. Every time

you say "The probability of X is Y", you really mean, "P(X|Q1) = Y". And when *he* says, "No, the probability of X is Z", he really *means,* "P(X|Q2) = Z".

Now you might, if you traced out his mathematical calculations, agree that, indeed, the conditional probability of the ticket winning, given his weird prior is 0.9. But you wouldn't agree that "the probability of the ticket winning" is 0.9. Just as he wouldn't agree that "the probability of the ticket winning" is 10^-8.

Even if the two of you refer to different mathematical calculations when you say the word "probability", *you* don't think that puts you on equal ground, neither of you being better than the other. And neither does he, of course.

So you see that, subjectively, probability really *does* feel objective—even after you have subjectively taken all apparent subjectivity into account.

And this is not mistaken, because, by golly, the probability of winning the lottery really *is* 10^-8, not 0.9. It's not as if you're doing your probability calculation *wrong,* after all. If you weren't worried about being fair or about justifying yourself to philosophers, if you only wanted to get the correct answer, your betting odds would be 10^-8.

Somewhere out in mind design space, there's a mind with any possible prior; but that doesn't mean that you'll say, "All priors are created equal."

When you judge those alternate minds, you'll do so using your own mind—your own beliefs about the universe—your own posterior that came out of your own prior, your own posterior probability assignments P(X|A,B,C,...,Q1). But there's nothing wrong with that. It's not like you could judge using something other than yourself. It's not like you could have a probability assignment without any prior, a degree of uncertainty that isn't in any mind.

And so, when all that is said and done, it still seems like the probability of winning the lottery really *is* 10^-8, not 0.9. No matter what other minds in design space say differently.

Which shouldn't be surprising. When you compute probabilities, you're thinking about lottery balls, not thinking about brains or mind designs or other people with different priors. Your probability computation makes no mention of that, any more than it explicitly represents itself. Your goal, after all, is to win, not to be fair. So of course probability will *seem* to be independent of what other minds might think of it.

Okay, but... you *still* can't win the lottery by assigning a higher probability to winning.

If you like, we could regard probability as an idealized computation, just like 2 + 2 = 4 seems to be independent of any particular error-prone calculator that computes it; and you could regard your mind as trying to approximate this ideal computation. In which case, it is good that your mind does not mention people's opinions, and only thinks of the lottery balls; the ideal computation makes no mention of people's opinions, and we are trying to reflect this ideal as accurately as possible...

But what you will calculate is the "ideal calculation" to plug into your betting odds, will depend on your prior, even though the calculation won't have an explicit dependency on "your prior". Someone who thought the universe was anti-Occamian, would advocate an anti-Occamian calculation, regardless of whether or not anyone thought the universe was anti-Occamian.

Your calculations get checked against reality, in a probabilistic way; you either win the lottery or not. But interpreting these results, is done with your prior; once again there is no probability that isn't in any mind.

I am not trying to argue that you can win the lottery by wishing, of course. Rather, I am trying to inculcate the ability to *distinguish between levels*.

When you think about the ontological nature of probability, and perform reductionism on it—when you try to explain how "probability" fits into a universe in which states of mind do not exist *fundamentally*—then you find that probability is computed within a brain; and you find that other possible minds could perform mostly-analogous operations with different priors and arrive at different answers.

But, when you consider probability *as probability,* think about the *referent* instead of the thought process—which thinking you will do in your own thoughts, which are physical processes—then you will conclude that the vast majority of possible priors are *probably wrong*. (You will also be able to conceive of priors which are, in fact, better than yours, because they assign more probability to the actual outcome; you just won't know in advance which alternative prior is the truly better one.)

If you again swap your goggles to think about how probability is implemented in the brain, the seeming objectivity of probability is the way the probability algorithm feels from inside; so it's no *mystery* that, considering probability as probability, you feel that it's not subject to your whims. That's just what the probability-computation would be expected to say, since the computation doesn't represent any dependency on your whims.

But when you swap out those goggles and go back to thinking about probabilities, then, by golly, your algorithm seems to be *right* in computing that probability is not subject to your whims. You *can't* win the lottery just by changing your beliefs about it. And if that is the way you would be expected to feel, then so what? The feeling has been explained, not explained away; it is not a *mere* feeling. Just because a calculation is implemented in your brain, doesn't mean it's *wrong,* after all.

Your "probability that the ten trillionth decimal digit of pi is 4", is an attribute of yourself, and exists in your mind; the real digit is either 4 or not. And if you could change your belief about the probability by editing your brain, you wouldn't expect that to change the probability.

Therefore I say of probability that it is "subjectively objective".

Part of *The Metaethics Sequence*

Next post: "Whither Moral Progress?"

Previous post: "Rebelling Within Nature"

Whither Moral Progress?

Followup to: Is Morality Preference?

In the dialogue "Is Morality Preference?", Obert argues for the existence of moral progress by pointing to free speech, democracy, mass street protests against wars, the end of slavery... and we could also cite female suffrage, or the fact that burning a cat alive was once a popular entertainment... and many other things that our ancestors believed were right, but which we have come to see as wrong, or vice versa.

But Subhan points out that if your only measure of progress is to take a difference against your current state, then you can follow a random walk, and still see the appearance of inevitable progress.

One way of refuting the simplest version of this argument, would be to say that we don't automatically think ourselves the very apex of possible morality; that we can imagine our descendants being more moral than us.

But can you *concretely* imagine a being morally wiser than yourself—one who knows that some particular thing is wrong, when you believe it to be right?

Certainly: I am not sure of the moral status of chimpanzees, and hence I find it easy to imagine that a future civilization will label them definitely people, and castigate us for failing to cryopreserve the chimpanzees who died in human custody.

Yet this still doesn't prove the existence of moral progress. Maybe I am simply mistaken about the nature of changes in morality that have previously occurred—like looking at a time chart of "differences between past and present", noting that the difference has been steadily decreasing, and saying, without being able to visualize it, "Extrapolating this chart into the future, we find that the future will be even less different from the present than the present."

So let me throw the question open to my readers: Whither moral progress?

You might say, perhaps, "Over time, people have become more willing to help one another—that is the very substance and definition of moral progress."

But as John McCarthy put it:

> "If everyone were to live for others all the time, life would be like a procession of ants following each other around in a circle."

Once you make "People helping each other more" the *definition* of moral progress, then people helping each other all the time, is *by definition* the *apex* of moral progress.

At the very least we have Moore's Open Question: It is not clear that helping others all the time is *automatically* moral progress, whether or not you argue that it is; and so we apparently have some notion of what constitutes "moral progress" that goes beyond the direct identification with "helping others more often".

Or if you identify moral progress with "Democracy!", then at some point there was a first democratic civilization—at some point, people went from having no notion of democracy as a good thing, to inventing the idea of democracy as a good thing. If increasing democracy is the very substance of moral progress, then how did this moral progress come about to exist in the world? How did people invent, without knowing it, this very substance of moral progress?

It's easy to come up with *concrete* examples of moral progress. Just point to a moral disagreement between past and present civilizations; or point to a disagreement between yourself and present civilization, and claim that future civilizations might agree with you.

It's harder to answer Subhan's challenge—to show *directionality,* rather than a random walk, on the meta-level. And explain how this directionality is implemented, on the meta-level: how people go from not having a moral ideal, to having it.

(I have my own ideas about this, as some of you know. And I'll thank you *not* to link to them in the comments, or quote them and attribute them to me, until at least 24 hours have passed from this post.)

Part of *The Metaethics Sequence*

Next post: "The Gift We Give To Tomorrow"

Previous post: "Probability is Subjectively Objective"

The Gift We Give To Tomorrow

How, oh how, did an unloving and mindless universe, cough up minds who were capable of love?

"No mystery in that," you say, "it's just a matter of natural selection."

But natural selection is cruel, bloody, and bloody stupid. Even when, on the surface of things, biological organisms aren't *directly* fighting each other—aren't *directly* tearing at each other with claws—there's still a deeper competition going on between the genes. Genetic information is created when genes increase their *relative* frequency in the next generation—what matters for "genetic fitness" is not how many children you have, but that you have *more* children than others. It is quite possible for a species to evolve to extinction, if the winning genes are playing negative-sum games.

How, oh how, could such a process create beings capable of love?

"No mystery," you say, "there is never any mystery-in-the-world; mystery is a property of questions, not answers. A mother's children share her genes, so the mother loves her children."

But sometimes mothers adopt children, and still love them. And mothers love their children for themselves, not for their genes.

"No mystery," you say, "Individual organisms are adaptation-executers, not fitness-maximizers. Evolutionary psychology is not about deliberately maximizing fitness—through most of human history, we didn't know genes existed. We don't calculate our acts' effect on genetic fitness consciously, or even subconsciously."

But human beings form friendships even with non-relatives: how, oh how, can it be?

"No mystery, for hunter-gatherers often play Iterated Prisoner's Dilemmas, the solution to which is reciprocal altruism. Sometimes the most dangerous human in the tribe is not the strongest, the prettiest, or even the smartest, but the one who has the most allies."

Yet not all friends are fair-weather friends; we have a concept of true friendship—and some people have sacrificed their life for their friends. Would not such a devotion tend to remove itself from the gene pool?

"You said it yourself: we have a concept of true friendship and fair-weather friendship. We can tell, or try to tell, the difference between someone who considers us a valuable ally, and someone executing the friendship adaptation. We wouldn't be true friends with someone who we didn't think was a true friend to us—and someone with many *true* friends is far more formidable than someone with many fair-weather allies."

And Mohandas Gandhi, who really did turn the other cheek? Those who try to serve all humanity, whether or not all humanity serves them in turn?

"That perhaps is a more complicated story. Human beings are not just social animals. We are political animals who argue linguistically about policy in adaptive tribal contexts. Sometimes the formidable human is not the strongest, but the one who can most skillfully argue that their preferred policies match the preferences of others."

Um... that doesn't explain Gandhi, or am I missing something?

"The point is that we have the ability to *argue* about 'What should be done?' as a *proposition*—we can make those arguments and respond to those arguments, without which politics could not take place."

Okay, but Gandhi?

"Believed certain complicated propositions about 'What should be done?' and did them."

That sounds like it could explain any possible human behavior.

"If we traced back the chain of causality through all the arguments, it would involve: a moral architecture that had the ability to argue *general abstract* moral propositions like 'What should be done to people?'; appeal to hardwired intuitions like fairness, a concept of duty, pain aversion + empathy; something like a preference for simple moral propositions, probably reused from our previous Occam prior; and the end result of all this, plus perhaps memetic selection effects, was 'You should not hurt people' in full generality—"

And that gets you Gandhi.

"Unless you think it was magic, it has to fit into the lawful causal development of the universe somehow."

Well... I certainly won't postulate magic, under any name.

"Good."

But come on... doesn't it seem a little... *amazing*... that hundreds of millions of years worth of evolution's death tournament could cough up mothers and fathers, sisters and brothers, husbands and wives, steadfast friends and honorable enemies, true altruists and guardians of causes, police officers and loyal defenders, even artists sacrificing themselves for their art, all practicing so many kinds of love? For so many things other than genes? Doing their part to make their world less ugly, something besides a sea of blood and violence and mindless replication?

"Are you claiming to be surprised by this? If so, question your underlying model, for it has led you to be surprised by the true state of affairs. Since the beginning, not one unusual thing has ever happened."

But how is it *not* surprising?

"What are you suggesting, that some sort of shadowy figure stood behind the scenes and directed evolution?"

Hell no. But—

"Because if you *were* suggesting that, I would have to ask how that shadowy figure *originally* decided that love was a *desirable* outcome of evolution. I would have to ask where that figure got preferences that included things like love, friendship, loyalty, fairness, honor, romance, and so on. On evolutionary psychology, we can see how *that specific outcome* came about—how *those particular goals rather than others* were *generated in the first place.* You can call it 'surprising' all you like. But when you

really do understand evolutionary psychology, you can see how parental love and romance and honor, and even true altruism and moral arguments, *bear the specific design signature of natural selection* in particular adaptive contexts of the hunter-gatherer savanna. So if there was a shadowy figure, it must itself have evolved—and that obviates the whole point of postulating it."

I'm not postulating a shadowy figure! I'm just asking how human beings ended up so *nice.*

"*Nice!* Have you *looked* at this planet lately? We also bear all those other emotions that evolved, too—which would tell you very well that we evolved, should you begin to doubt it. Humans aren't always nice."

We're one hell of a lot nicer than the process that produced us, which lets elephants starve to death when they run out of teeth, and doesn't anesthetize a gazelle even as it lays dying and is of no further importance to evolution one way or the other. It doesn't take much to be nicer than evolution. To have the *theoretical capacity* to make one single gesture of mercy, to feel a single twinge of empathy, is to be nicer than evolution. How did evolution, which is itself so uncaring, create minds on that qualitatively higher moral level than itself? How did evolution, which is so ugly, end up doing anything so *beautiful?*

"Beautiful, you say? Bach's *Little Fugue in G Minor* may be beautiful, but the sound waves, as they travel through the air, are not stamped with tiny tags to specify their beauty. If you wish to find *explicitly encoded* a measure of the fugue's beauty, you will have to look at a human brain—nowhere else in the universe will you find it. Not upon the seas or the mountains will you find such judgments written: they are not minds, they cannot think."

Perhaps that is so, but still I ask: How did evolution end up doing anything so beautiful, as giving us the ability to admire the beauty of a flower?

"Can you not see the circularity in your question? If beauty were like some great light in the sky that shined from outside humans, then your question might make sense—though there would still be the question of how humans came to perceive that light. You evolved with a psychology unlike evolution: Evolution has nothing like the intelligence or the precision required to exactly quine its goal system. In coughing up the first true minds, evolution's simple fitness criterion shattered into a thousand values. You evolved with a psychology that attaches utility to things which evolution does not care about, like human life and happiness. And then you look back and say, 'How marvelous, that uncaring evolution produced minds that care about sentient life!' So your great marvel and wonder, that seems like far too much coincidence, is really no coincidence at all."

But then it is still amazing that this particular circular loop, happened to loop around such important things as beauty and altruism.

"I don't think you're following me here. To you, it seems natural to privilege the beauty and altruism as special, as preferred, because you value them highly; and you don't see this as a unusual fact about yourself, because many of your friends do likewise. So you expect that a ghost of perfect emptiness would also value life and happiness—and then, from this standpoint outside reality, a great coincidence would indeed have occurred."

But you can make arguments for the importance of beauty and altruism from first principles—that our aesthetic senses lead us to create new complexity, instead of repeating the same things over and over; and that altruism is important because it takes us outside ourselves, gives our life a higher meaning than sheer brute selfishness.

"Oh, and *that* argument is going to move even a ghost of perfect emptiness—now that you've appealed to slightly different values? Those aren't first principles, they're just *different* principles. Even if you've adopted a high-falutin' philosophical tone, still there are no *universally* compelling arguments. All you've done is pass the recursive buck."

You don't think that, somehow, we evolved to *tap into* something beyond—

"What good does it do to suppose something beyond? Why should we pay more attention to that beyond thing, than we pay to our existence as humans? How does it alter your personal responsibility, to say that you were only following the orders of the beyond thing? And you would still have evolved to let the beyond thing, rather than something else, direct your actions. You are only passing the recursive buck. Above all, it would be *too much coincidence.*"

Too much coincidence?

"A flower is beautiful, you say. Do you think there is no story behind that beauty, or that science does not know the story? Flower pollen is transmitted by bees, so by sexual selection, flowers evolved to attract bees—by imitating certain mating signs of bees, as it happened; the flowers' patterns would look more intricate, if you could see in the ultraviolet. Now healthy flowers are a sign of fertile land, likely to bear fruits and other treasures, and probably prey animals as well; so is it any wonder that humans evolved to be attracted to flowers? But for there to be some great light written upon the very stars—those huge unsentient balls of burning hydrogen—which *also* said that flowers were beautiful, now *that* would be far too much coincidence."

So you explain away the beauty of a flower?

"No, I explain it. Of course there's a story behind the beauty of flowers and the fact that we find them beautiful. Behind ordered events, one finds ordered stories; and what has no story is the product of random noise, which is hardly any better. If you cannot take joy in things that have stories behind them, your life will be empty indeed. I don't think I take any less joy in a flower than you do; more so, perhaps, because I take joy in its story as well."

Perhaps as you say, there is no surprise from a causal viewpoint—no disruption of the physical order of the universe. But it still seems to me that, in this creation of humans by evolution, something happened that is precious and marvelous and wonderful. If we cannot call it a physical miracle, then call it a moral miracle.

"Because it's only a miracle from the perspective of the morality that was produced, thus explaining away all of the apparent coincidence from a merely causal and physical perspective?"

Well... I suppose you could interpret the term that way, yes. I just meant something that was immensely surprising and wonderful on a moral level, even if it is not surprising on a physical level.

"I think that's what I said."

But it still seems to me that you, from your own view, drain something of that wonder away.

"Then you have problems taking joy in the merely real. Love has to begin *somehow,* it has to enter the universe *somewhere.* It is like asking how life itself begins—and though you were born of your father and mother, and they arose from their living parents in turn, if you go far and far and far away back, you will finally come to a replicator that arose by pure accident—the border between life and unlife. So too with love.

"A complex pattern must be explained by a cause which is not already that complex pattern. Not just the event must be explained, but the very shape and form. For love to first enter Time, it must come of something that is not love; if this were not possible, then love could not be.

"Even as life itself required that first replicator to come about by accident, parentless but still caused: far, far back in the causal chain that led to you: 3.85 billion years ago, in some little tidal pool.

"Perhaps your children's children will ask how it is that they are capable of love.

"And their parents will say: Because we, who also love, created you to love.

"And your children's children will ask: But how is it that *you* love?

"And their parents will reply: Because our own parents, who also loved, created us to love in turn.

"Then your children's children will ask: But where did it all begin? Where does the recursion end?

"And their parents will say: Once upon a time, long ago and far away, ever so long ago, there were intelligent beings who were not themselves intelligently designed. Once upon a time, there were lovers created by something that did not love.

"Once upon a time, when all of civilization was a single galaxy and a single star: and a single planet, a place called Earth.

"Long ago, and far away, ever so long ago."

Could Anything Be Right?

Years ago, Eliezer_{1999} was convinced that he knew *nothing* about morality.

For all he knew, morality could require the extermination of the human species; and if so he saw no virtue in taking a stand against morality, because he thought that, by definition, if he postulated that moral fact, that meant human extinction was what "should" be done.

I thought I could *figure out* what was right, perhaps, given enough reasoning time and enough facts, but that I currently had no information about it. I could not trust evolution which had built me. What foundation did that leave on which to stand?

Well, indeed Eliezer_{1999} was massively mistaken about the nature of morality, so far as his explicitly represented philosophy went.

But as Davidson once observed, if you believe that "beavers" live in deserts, are pure white in color, and weigh 300 pounds when adult, then you do not have any beliefs *about* beavers, true or false. You must get at least some of your beliefs right, before the remaining ones can be wrong *about* anything.

My belief that I had *no* information *about* morality was not internally consistent.

Saying that I knew nothing felt virtuous, for I had once been taught that it was virtuous to confess my ignorance. "The only thing I know is that I know nothing," and all that. But in this case I would have been better off considering the admittedly exaggerated saying, "The greatest fool is the one who is not aware they are wise." (This is nowhere near the *greatest* kind of foolishness, but it is a kind of foolishness.)

Was it wrong to kill people? Well, I thought so, but I wasn't sure; maybe it was right to kill people, though that seemed less likely.

What kind of *procedure* would answer whether it was right to kill people? I didn't know that either, but I thought that if you built a generic superintelligence (what I would later label a "ghost of perfect emptiness") then it could, you know, reason about what was likely to be right and wrong; and since it was *superintelligent,* it was bound to come up with the right answer.

The problem that I somehow managed not to think too hard about, was where the superintelligence would get the procedure that discovered the procedure that discovered the procedure that discovered morality—if I couldn't write it into the start state that wrote the successor AI that wrote the successor AI.

As Marcello Herreshoff later put it, "We never bother running a computer program unless we don't know the output and we know an important fact about the output." If I knew nothing about morality, and did not even claim to know the nature of morality, then how could I construct any computer program whatsoever—even a "superintelligent" one or a "self-improving" one—and claim that it would output something called "morality"?

There are no-free-lunch theorems in computer science—in a maxentropy universe, no plan is better on average than any other. If you have no knowledge at all about "morality", there's also no computational procedure that will seem more likely than

others to compute "morality", and no meta-procedure that's more likely than others to produce a procedure that computes "morality".

I thought that surely even a ghost of perfect emptiness, finding that it knew nothing of morality, would see a moral imperative to *think about morality*.

But the difficulty lies in the word *think.* Thinking is not an activity that a ghost of perfect emptiness is automatically able to carry out. Thinking requires running some *specific* computation that is the thought. For a reflective AI to decide to think, requires that it know some computation which it believes is *more* likely to tell it what it wants to know, than consulting an Ouija board; the AI must also have a notion of how to interpret the output.

If one knows nothing about morality, what does the word "should" mean, at all? If you don't know whether death is right or wrong—and don't know how you can discover whether death is right or wrong—and don't know whether any given procedure might *output* the procedure for saying whether death is right or wrong—then what do these words, "right" and "wrong", even *mean*?

If the words "right" and "wrong" have *nothing* baked into them—no starting point—if *everything* about morality is up for grabs, not just the content but the structure and the starting point and the determination procedure—then what is their meaning? What distinguishes, "I don't know what is right" from "I don't know what is wakalixes"?

A scientist may say that everything is up for grabs in science, since any theory may be disproven; but then they have some idea of what would count as *evidence* that could disprove the theory. Could there be something that would change what a scientist regarded as evidence?

Well, yes, in fact; a scientist who read some Karl Popper and thought they knew what "evidence" meant, could be presented with the coherence and uniqueness proofs underlying Bayesian probability, and that might change their definition of evidence. They might not have had any *explicit notion,* in advance, that such a proof could exist. But they would have had an implicit notion. It would have been baked into their brains, if not explicitly represented therein, that such-and-such an argument would in fact persuade them that Bayesian probability gave a better definition of "evidence" than the one they had been using.

In the same way, you could say, "I don't know what morality is, but I'll know it when I see it," and make sense.

But then you are not rebelling completely against your own evolved nature. You are supposing that whatever has been baked into you to recognize "morality", is, if not absolutely trustworthy, then at least your initial condition with which you start debating. Can you trust your moral intuitions to give you any information about morality *at all*, when they are the product of mere evolution?

But if you discard every procedure that evolution gave you *and all its products*, then you discard your whole brain. You discard everything that could potentially recognize morality when it sees it. You discard everything that could potentially respond to moral arguments by updating your morality. You even unwind past the unwinder: you discard the intuitions underlying your conclusion that *you can't trust evolution* to be moral. It is your *existing* moral intuitions that tell you that evolution doesn't seem like a very *good* source of morality. What, then, will the words "right" and "should" and "better" even *mean*?

Humans do not perfectly recognize truth when they see it, and hunter-gatherers do not have an explicit concept of the Bayesian criterion of evidence. But all our science and all our probability theory was built on top of a chain of appeals to our instinctive notion of "truth". Had this core been flawed, there would have been nothing we could do *in principle* to arrive at the present notion of science; the notion of science would have just sounded completely unappealing and pointless.

One of the arguments that might have shaken my teenage self out of his mistake, if I could have gone back in time to argue with him, was the question:

Could there be some morality, some given rightness or wrongness, that human beings do not perceive, do not want to perceive, will not see any appealing moral argument for adopting, nor any moral argument for adopting a procedure that adopts it, etcetera? Could there be a morality, and ourselves *utterly* outside its frame of reference? But then what makes this thing *morality*—rather than a stone tablet somewhere with the words 'Thou shalt murder' written on them, with absolutely no *justification* offered?

So all this suggests that you should be willing to accept that you might know a *little* about morality. Nothing unquestionable, perhaps, but an initial state with which to start questioning yourself. Baked into your brain but not explicitly known to you, perhaps; but still, that which your brain *would* recognize as *right* is what you are talking *about*. You will accept at least enough of the way you *respond to moral arguments* as a *starting point*, to identify "morality" as something to think about.

But that's a rather large step.

It implies accepting your own mind as identifying a moral frame of reference, rather than all morality being a great light shining from beyond (that in principle you might not be able to perceive at all). It implies accepting that even if there were a light and your brain decided to recognize it as "morality", it would still be your own brain that recognized it, and you would not have evaded causal responsibility—or evaded moral responsibility either, on my view.

It implies dropping the notion that a ghost of perfect emptiness will necessarily agree with you, because the ghost might occupy a different moral frame of reference, respond to different arguments, be *asking a different question* when it computes what-to-do-next.

And if you're willing to bake at least a few things into the very meaning of this topic of "morality", this quality of *rightness* that you are talking about when you talk about "rightness"—if you're willing to accept even that morality is what you argue about when you argue about "morality"—then why not accept other intuitions, other pieces of yourself, into the starting point as well?

Why not accept that, *ceteris paribus*, joy is preferable to sorrow?

You might later find some ground within yourself or built upon yourself with which to criticize this—but why not accept it for now? Not just as a personal preference, mind you; but as something baked into the *question* you ask when you ask "What is truly right"?

But then you might find that you know rather a lot about morality! Nothing certain—nothing unquestionable—nothing unarguable—but still, quite a bit of information. Are you willing to relinquish your Socratean ignorance?

I don't argue by definitions, of course. But if you claim to know nothing at all about morality, then you will have problems with the *meaning* of your words, not just their plausibility.

Existential Angst Factory

Followup to: The Moral Void

A widespread excuse for avoiding rationality is the widespread belief that it is "rational" to believe life is meaningless, and thus suffer existential angst. This is one of the secondary reasons why it is worth discussing the nature of morality. But it's also worth attacking existential angst directly.

I suspect that most existential angst is not really existential. I think that most of what is labeled "existential angst" comes from trying to solve the wrong problem.

Let's say you're trapped in an unsatisfying relationship, so you're unhappy. You consider going on a skiing trip, or you actually go on a skiing trip, and you're still unhappy. You eat some chocolate, but you're still unhappy. You do some volunteer work at a charity (or better yet, work the same hours professionally and donate the money, thus applying the Law of Comparative Advantage) and you're still unhappy because you're in an unsatisfying relationship.

So you say something like: "Skiing is meaningless, chocolate is meaningless, charity is meaningless, life is doomed to be an endless stream of woe." And you blame this on the universe being a mere dance of atoms, empty of meaning. Not necessarily because of some kind of subconsciously deliberate Freudian substitution to avoid acknowledging your real problem, but because you've stopped hoping that your real problem is solvable. And so, as a sheer unexplained background fact, you observe that you're always unhappy.

Maybe you're poor, and so always unhappy. Nothing you do solves your poverty, so it starts to seem like a universal background fact, along with your unhappiness. So when you *observe* that you're always unhappy, you blame this on the universe being a mere dance of atoms. Not as some kind of Freudian substitution, but because it has *ceased to occur to you* that there *does* exist some possible state of affairs in which life is not painful.

What about rich heiresses with everything in the world available to buy, who still feel unhappy? Perhaps they can't get themselves into satisfying romantic relationships. One way or another, they don't *know* how to use their money to create happiness—they lack the expertise in hedonic psychology and/or self-awareness and/or simple competence.

So they're constantly unhappy—and they blame it on existential angst, because they've already solved the only problem they know how to solve. They already have enough money and they've already bought all the toys. Clearly, if there's still a problem, it's because life is meaningless.

If someone who weighs 560 pounds suffers from "existential angst", *allegedly* because the universe is a mere dance of particles, then stomach reduction surgery might drastically change their views of the metaphysics of morality.

I'm not a fan of Timothy Ferris, but *The Four-Hour Workweek* does make an interesting fun-theoretic observation:

> Let's assume we have 10 goals and we achieve them—what is the desired outcome that makes all the effort worthwhile? The most common response is what I also would have suggested five years ago: happiness. I no longer believe this is a good answer. Happiness can be bought with a bottle of wine and has become ambiguous through overuse. There is a more precise alternative that reflects what I believe the actual objective is.
>
> Bear with me. What is the opposite of happiness? Sadness? No. Just as love and hate are two sides of the same coin, so are happiness and sadness. Crying out of happiness is a perfect illustration of this. The opposite of love is indifference, and the opposite of happiness is—here's the clincher—boredom.
>
> *Excitement is the more practical synonym for happiness, and it is precisely what you should strive to chase. It is the cure-all.* When people suggest you follow your "passion" or your "bliss," I propose that they are, in fact, referring to the same singular concept: excitement.
>
> This brings us full circle. The question you should be asking isn't "What do I want?" or "What are my goals?" but "What would excite me?"
>
> Remember—boredom is the enemy, not some abstract "failure."
>
> *Living* like a millionaire requires *doing* interesting things and not just owning enviable things.

I don't endorse all of the above, of course. But note the SolvingTheWrongProblem anti-pattern Ferris describes. It was on reading the above that I first generalized ExistentialAngstFactory.

Now, *if* someone is in a unproblematic, loving relationship; and they have enough money; and no major health problems; and they're signed up for cryonics so death is not approaching inexorably; and they're doing exciting work that they enjoy; and they believe they're having a positive effect on the world...

...and they're *still* unhappy because it seems to them that the universe is a mere dance of atoms empty of meaning, *then* we may have a legitimate problem here. One that, perhaps, can *only* be resolved by a very long discussion of the nature of morality and how it fits into a reductionist universe.

But, mostly, I suspect that when people complain about the empty meaningless void, it is because they have at least one problem that they aren't thinking about solving—perhaps because they never identified it. Being able to identify your own problems is a feat of rationality that schools don't explicitly train you to perform. And they haven't even been told that an un-focused-on problem might be the source of their "existential angst"—they've just been told to blame it on existential angst.

That's the other reason it might be helpful to understand the nature of morality—even if it just adds up to moral normality—because it tells you that if you're constantly unhappy, it's *not* because the universe is empty of meaning.

Or maybe believing the universe is a "mere dance of particles" is one more factor contributing to human unhappiness; in which case, again, people can benefit from eliminating that factor.

If it seems to you like nothing you do makes you happy, and you can't even imagine what would make you happy, it's not because the universe is made of particle fields. It's because you're *still* solving the wrong problem. Keep searching, until you find the visualizable state of affairs in which the existential angst seems like it should go away —that might (or might not) tell you the real problem; but at least, don't blame it on reductionism.

Added: Several commenters pointed out that random acts of brain chemistry may also be responsible for depression, even if your life is otherwise fine. As far as I know, this is true. But, once again, it won't help to mistake that random act of brain chemistry as being *about* existential issues; that might prevent you from trying neuropharmaceutical interventions.

Part of *The Metaethics Sequence*

Next post: "Can Counterfactuals Be True?"

Previous post: "Could Anything Be Right?"

Can Counterfactuals Be True?

Followup to: Probability is Subjectively Objective

The classic explanation of counterfactuals begins with this distinction:

1. If Lee Harvey Oswald didn't shoot John F. Kennedy, then someone else did.
2. If Lee Harvey Oswald hadn't shot John F. Kennedy, someone else would have.

In ordinary usage we would agree with the first statement, but not the second (I hope).

If, somehow, we learn the definite fact that Oswald did not shoot Kennedy, then someone else must have done so, since Kennedy was in fact shot.

But if we went back in time and removed Oswald, while leaving everything else the same, then—unless you believe there was a conspiracy—there's no particular reason to believe Kennedy would be shot:

We start by imagining the same historical situation that existed in 1963—by a further act of imagination, we remove Oswald from our vision—we run forward the laws that we think govern the world—visualize Kennedy parading through in his limousine—and find that, in our imagination, no one shoots Kennedy.

It's an interesting question whether counterfactuals can be *true* or *false.* We never get to experience them directly.

If we disagree on what *would have* happened if Oswald hadn't been there, what experiment could we perform to find out which of us is right?

And if the counterfactual is something unphysical—like, "If gravity had stopped working three days ago, the Sun would have exploded"—then there aren't even any alternate histories out there to provide a truth-value.

It's not as simple as saying that if the bucket contains three pebbles, and the pasture contains three sheep, the bucket is true.

Since the counterfactual event *only* exists in your imagination, how can it be true or false?

So... is it just as fair to say that "If Oswald hadn't shot Kennedy, the Sun would have exploded"?

After all, the event only exists in our imaginations—surely that means it's subjective, so we can say anything we like?

But so long as we have a lawful specification of how counterfactuals are constructed—a lawful computational procedure—then the counterfactual result of removing Oswald, depends entirely on the empirical state of the world.

If there was no conspiracy, then any reasonable computational procedure that simulates removing Oswald's bullet from the course of history, ought to return an answer of Kennedy not getting shot.

"Reasonable!" you say. "Ought!" you say.

But that's not the point; the point is that if you *do* pick some fixed computational procedure, whether it is reasonable or not, then either it *will* say that Kennedy gets shot, or not, and what it says will depend on the empirical state of the world. So that, if you tell me, "I believe that *this-and-such* counterfactual construal, run over Oswald's removal, preserves Kennedy's life", then I can deduce that you don't believe in the conspiracy.

Indeed, so long as we take this computational procedure as fixed, then the actual state of the world (which either does include a conspiracy, or does not) presents a ready truth-value for the output of the counterfactual.

In general, if you give me a fixed computational procedure, like "multiply by 7 and add 5", and then you point to a 6-sided die underneath a cup, and say, "The result-of-procedure is 26!" then it's not hard at all to assign a truth value to this statement. Even if the actual die under the cup only ever takes on the values between 1 and 6, so that "26" is not found anywhere under the cup. The statement is still true if and only if the die is showing 3; that is its empirical truth-condition.

And what about the statement ((3 * 7) + 5) = 26? Where is the truth-condition for *that* statement located? This I don't know; but I am nonetheless quite confident that it is true. Even though I am not confident that this 'true' means exactly the same thing as the 'true' in "the bucket is 'true' when it contains the same number of pebbles as sheep in the pasture".

So if someone I trust—presumably someone I *really* trust—tells me, "If Oswald hadn't shot Kennedy, someone else would have", and I believe this statement, then I believe the empirical reality is such as to make the counterfactual computation come out this way. Which would seem to imply the conspiracy. And I will anticipate accordingly.

Or if I find out that there *was* a conspiracy, then this will *confirm the truth-condition of the counterfactual*—which might make a bit more sense than saying, "Confirm that the counterfactual is true."

But how do you *actually* compute a counterfactual? For this you must consult Judea Pearl. Roughly speaking, you perform surgery on graphical models of causal processes; you sever some variables from their ordinary parents and surgically set them to new values, and then recalculate the probability distribution.

There are other ways of defining counterfactuals, but I confess they all strike me as entirely odd. Even worse, you have philosophers arguing over what the value of a counterfactual *really is* or *really means*, as if there were some counterfactual world actually floating out there in the philosophical void. If you think I'm attacking a strawperson here, I invite you to consult the philosophical literature on Newcomb's Problem.

A lot of philosophy seems to me to suffer from "naive philosophical realism"—the belief that philosophical debates are about things that automatically and directly exist as propertied objects floating out there in the void.

You can talk about an ideal computation, or an ideal process, that would ideally be applied to the empirical world. You can talk about your uncertain beliefs about the output of this ideal computation, or the result of the ideal process.

So long as the computation is fixed, and so long as the computational itself is only over actually existent things. Or the results of other computations previously defined—you should not have your computation be over "nearby possible worlds" unless you can tell me how to compute those, as well.

A chief sign of naive philosophical realism is that it does not tell you how to write a computer program that computes the objects of its discussion.

I have yet to see a camera that peers into "nearby possible worlds"—so even after you've analyzed counterfactuals in terms of "nearby possible worlds", I still can't write an AI that computes counterfactuals.

But Judea Pearl tells me just how to compute a counterfactual, given only my beliefs about the *actual* world.

I strongly privilege the *real world that actually exists,* and to a slightly lesser degree, logical truths about mathematical objects (preferably finite ones). Anything *else* you want to talk about, I need to figure out how to describe in terms of the first two—for example, as the output of an ideal computation run over the empirical state of the real universe.

The absence of this requirement as a condition, or at least a goal, of modern philosophy, is one of the primary reasons why modern philosophy is often surprisingly useless in my AI work. I've read whole books about decision theory that take counterfactual distributions as givens, and never tell you how to compute the counterfactuals.

Oh, and to talk about "the probability that John F. Kennedy was shot, given that Lee Harvey Oswald didn't shoot him", we write:

```
P(Kennedy_shot|Oswald_not)
```

And to talk about "the probability that John F. Kennedy would have been shot, if Lee Harvey Oswald hadn't shot him", we write:

```
P(Oswald_not []-> Kennedy_shot)
```

That little symbol there is supposed to be a box with an arrow coming out of it, but I don't think Unicode has it.

Part of *The Metaethics Sequence*

Next post: "Math is Subjunctively Objective"

Previous post: "Existential Angst Factory"

Math is Subjunctively Objective

Followup to: Probability is Subjectively Objective, Can Counterfactuals Be True?

I am quite confident that the statement 2 + 3 = 5 is *true*; I am far less confident of what it *means* for a mathematical statement to be true.

In "The Simple Truth" I defined a pebble-and-bucket system for tracking sheep, and defined a condition for whether a bucket's pebble level is "true" in terms of the sheep. The bucket is the belief, the sheep are the reality. I believe 2 + 3 = 5. Not just that two sheep plus three sheep equal five sheep, but that 2 + 3 = 5. That is my belief, but where is the reality?

So now the one comes to me and says: "Yes, two sheep plus three sheep equals five sheep, and two stars plus three stars equals five stars. I won't deny that. But this notion that 2 + 3 = 5, *exists only in your imagination, and is purely subjective.*"

So I say: Excuse me, *what?*

And the one says: "Well, I know what it means to observe two sheep and three sheep leave the fold, and five sheep come back. I know what it means to press '2' and '+' and '3' on a calculator, and see the screen flash '5'. I even know what it means to ask someone 'What is two plus three?' and hear them say 'Five.' But you insist that there is some fact *beyond* this. You insist that 2 + 3 = 5."

Well, it kinda *is.*

"Perhaps you just mean that when you *mentally visualize* adding two dots and three dots, you end up visualizing five dots. Perhaps this is the content of what you mean by saying, 2 + 3 = 5. I have no trouble with that, for brains are as real as sheep."

No, for it seems to me that 2 + 3 equaled 5 *before* there were any humans around to do addition. When humans showed up on the scene, they did not *make* 2 + 3 equal 5 by virtue of thinking it. Rather, they thought that '2 + 3 = 5' *because* 2 + 3 did in fact equal 5.

"Prove it."

I'd love to, but I'm busy; I've got to, um, eat a salad.

"The *reason you believe* that 2 + 3 = 5, is your mental visualization of two dots plus three dots yielding five dots. Does this not imply that this physical event in your physical brain is the *meaning* of the statement '2 + 3 = 5'?"

But I honestly don't think that *is* what I mean. Suppose that by an amazing cosmic coincidence, a flurry of neutrinos struck my neurons, causing me to imagine two dots colliding with three dots and visualize six dots. I would then say, '2 + 3 = 6'. But this wouldn't mean that 2 + 3 actually *had* become equal to 6. Now, if what I mean by '2 + 3' consists entirely of what my mere physical brain merely *happens to output*, then a neutrino *could* make 2 + 3 = 6. But you can't change arithmetic by tampering with a calculator.

"Aha! I have you now!"

Is that so?

"Yes, you've given your whole game away!"

Do tell.

"You visualize a subjunctive world, a counterfactual, where your brain is struck by neutrinos, and says, '2 + 3 = 6'. So you know that in this case, your future self will *say* that '2 + 3 = 6'. But then you add up dots in your *own, current brain,* and your *current* self gets five dots. So you say: 'Even if I believed "2 + 3 = 6", then 2 + 3 would still equal 5.' You say: '2 + 3 = 5 regardless of what anyone thinks of it.' So your *current* brain, computing the same question while it *imagines* being different but is not *actually* different, finds that the answer *seems to be the same*. Thus your brain creates the *illusion* of an additional reality that exists outside it, independent of any brain."

Now hold on! You've *explained* my belief that 2 + 3 = 5 regardless of what anyone thinks, but that's not the same as explaining away my belief. Since 2 + 3 = 5 does not, *in fact*, depend on what any human being thinks of it, therefore it is *right and proper* that when I imagine counterfactual worlds in which people (including myself) *think* '2 + 3 = 6', and I ask what 2 + 3 *actually* equals in this counterfactual world, it still comes out as 5.

"Don't you see, that's just like trying to visualize motion stopping everywhere in the universe, by imagining yourself as an observer outside the universe who experiences time passing while nothing moves. But really there is no time without motion."

I see the analogy, but I'm not sure it's a deep analogy. Not everything you can imagine seeing, doesn't exist. It seems to me that a brain can *easily* compute quantities that don't depend on the brain.

"*What?* Of *course* everything that the brain computes depends on the brain! Everything that the brain computes, is computed inside the brain!"

That's not what I mean! I just mean that the brain can perform computations that *refer to* quantities outside the brain. You can set up a question, like 'How many sheep are in the field?', that isn't *about* any particular person's brain, and whose *actual* answer doesn't *depend on* any particular person's brain. And then a brain can faithfully compute that answer.

If I count two sheep and three sheep returning from the field, and Autrey's brain gets hit by neutrinos so that Autrey thinks there are six sheep in the fold, then that's not going to *cause* there to be six sheep in the fold—right? The whole question here is just *not about* what Autrey thinks, it's *about* how many sheep are in the fold.

Why should I care what *my* subjunctive future self thinks is the sum of 2 + 3, any more than I care what *Autrey* thinks is the sum of 2 + 3, when it comes to asking what is *really* the sum of 2 + 3?

"Okay... I'll take another tack. Suppose you're a psychiatrist, right? And you're an expert witness in court cases—basically a hired gun, but you try to deceive yourself about it. Now wouldn't it be a bit suspicious, to find yourself saying: 'Well, the only reason *that I in fact believe* that the defendant is insane, is because I was paid to be an expert psychiatric witness for the defense. And if I had been paid to witness for the prosecution, I undoubtedly would have come to the conclusion that the defendant

is sane. But my belief that the defendant is insane, is *perfectly justified;* it is justified by my observation that the defendant used his own blood to paint an Elder Sign on the wall of his jail cell.'"

Yes, that *does* sound suspicious, but I don't see the point.

"My point is that the *physical cause* of your belief that 2 + 3 = 5, is the physical event of your brain visualizing two dots and three dots and coming up with five dots. If your brain came up six dots, due to a neutrino storm or whatever, you'd think '2 + 3 = 6'. How can you possibly say that your belief *means* anything other than the number of dots your brain came up with?"

Now hold on just a second. Let's say that the psychiatrist is paid by the judge, and when he's paid by the judge, he renders an honest and neutral evaluation, and his evaluation is that the defendant is sane, just played a bit too much Mythos. So it is true to say that if the psychiatrist had been paid by the defense, then the psychiatrist would have found the defendant to be insane. But that doesn't mean that when the psychiatrist is paid by the *judge,* you should dismiss his evaluation as telling you *nothing more than* 'the psychiatrist was paid by the judge'. On those occasions where the psychiatrist *is* paid by the judge, his opinion varies with the defendant, and conveys real evidence about the defendant.

"Okay, so now what's y*our* point?"

That when my brain is *not* being hit by a neutrino storm, it yields honest and informative evidence that 2 + 3 = 5.

"And if your brain *was* hit by a neutrino storm, you'd be saying, '2 + 3 = 6 regardless of what anyone thinks of it'. Which shows how reliable *that* line of reasoning is."

I'm not claiming that my saying '2 + 3 = 5 no matter what anyone thinks' represents stronger *numerical* evidence than my saying '2 + 3 = 5'. My saying the former just tells you something extra about my epistemology, not numbers.

"And you don't think your epistemology is, oh, a little... *incoherent?*"

No! I think it is perfectly coherent to simultaneously hold all of the following:

- 2 + 3 = 5.
- If neutrinos make me believe "2 + 3 = 6", then 2 + 3 = 5.
- If neutrinos make me believe "2 + 3 = 6", then I will say "2 + 3 = 6".
- If neutrinos make me believe that "2 + 3 = 6", then I will thereafter assert that "If neutrinos make me believe '2 + 3 = 5', then 2 + 3 = 6".
- The cause of my thinking that "2 + 3 = 5 independently of what anyone thinks" is that my *current* mind, when it subjunctively recomputes the value of 2 + 3 under the assumption that my *imagined* self is hit by neutrinos, does not see the *imagined* self's beliefs as changing the dots, and my *current* brain just visualizes two dots plus three dots, as before, so that the imagination of my *current* brain shows the same result.
- If I were *actually* hit by neutrinos, my brain would compute a different result, and I would assert "2 + 3 = 6 independently of what anyone thinks."
- 2 + 3 = 5 independently of what anyone thinks.
- Since 2 + 3 will *in fact* go on equaling 5 *regardless* of what I imagine about it or how my brain visualizes cases where my future self has different beliefs, it's a

> *good thing* that my imagination doesn't visualize the result as depending on my beliefs.

"Now that's just crazy talk!"

No, *you're* the crazy one! You're *collapsing your levels*; you think that just because my brain asks a question, it should start mixing up queries about the state of my brain *into* the question. Not every question my brain asks is *about* my brain!

Just because something is computed *in* my brain, doesn't mean that my computation has to depend on my brain's *representation of* my brain. It certainly doesn't mean that the *actual quantity* depends on my brain! It's my brain that computes my beliefs about gravity, and if neutrinos hit me I will come to a different conclusion; but that doesn't mean that I can think different and fly. And I don't *think* I can think different and fly, either!

I am not a calculator who, when someone presses my "2" and "+" and "3" buttons, computes, "What do I output when someone presses 2 + 3?" I am a calculator who computes "What is 2 + 3?" The former is a circular question that can consistently return any answer—which makes it not very *helpful.*

Shouldn't we expect non-circular questions to be the *normal* case? The brain evolved to guess at the state of the environment, not guess at 'what the brain will think is the state of the environment'. Even when the brain models itself, it is trying to *know itself*, not trying to know *what it will think about itself*.

Judgments that depend on our representations of *anyone's* state of mind, like "It's okay to kiss someone only if they want to be kissed", are the exception rather than the rule.

Most quantities we bother to think about at all, will appear to be 'the same regardless of what anyone thinks of them'. When we imagine thinking differently about the quantity, we will imagine the quantity coming out the same; it will feel "subjunctively objective".

And there's nothing wrong with that! If something *appears* to be the same regardless of what anyone thinks, then maybe that's because it *actually is* the same regardless of what anyone thinks.

Even if you explain that the quantity *appears* to stay the same in my imagination, *merely* because my current brain computes it the same way—well, how *else* would I imagine something, *except* with my current brain? Should I imagine it using a rock?

"Okay, so it's possible for something that appears thought-independent, to actually be thought-independent. But why do you think that 2 + 3 = 5, in particular, has some kind of existence independently of the dots you imagine?"

Because two sheep plus three sheep equals five sheep, and this appears to be true in every mountain and every island, every swamp and every plain and every forest.

And moreover, it is also true of two rocks plus three rocks.

And further, when I press buttons upon a calculator and activate a network of transistors, it *successfully predicts* how many sheep or rocks I will find.

Since all these quantities, correlate with each other and successfully predict each other, surely they must have something *like* a common cause, a similarity that factors out? Something that is true beyond and before the concrete observations? Something that the concrete observations hold in common? And this commonality is then also the sponsor of my answer, 'five', that I find in my own brain.

"But my dear sir, if the fact of 2 + 3 = 5 exists somewhere outside your brain... *then where is it?*"

Damned if I know.

Part of *The Metaethics Sequence*

Next post: "Does Your Morality Care What You Think?"

Previous post: "Can Counterfactuals Be True?"

Does Your Morality Care What You Think?

Followup to: Math is Subjunctively Objective, The Moral Void, Is Morality Given?

Thus I recall the study, though I cannot recall the citation:

Children, at some relatively young age, were found to distinguish between:

- The teacher, by saying that we're allowed to stand on our desks, can make it right to do so.
- The teacher, by saying that I'm allowed to take something from another child's backpack, *cannot* make it right to do so.

Obert: "Well, I don't know the citation, but it sounds like a fascinating study. So even children, then, realize that moral facts are givens, beyond the ability of teachers or parents to alter."

Subhan: "You say that like it's a good thing. Children may also think that people in Australia have to wear heavy boots from falling off the other side of the Earth."

Obert: "Call me Peter Pan, then, because I never grew up on this one. Of course it doesn't matter what the teacher says. It doesn't matter what I say. It doesn't even matter what I *think.* Stealing is wrong. Do you *disagree?*"

Subhan: "You don't see me picking your pockets, do you? Isn't it enough that I *choose* not to steal from you—do I have to pretend it's the law of the universe?"

Obert: "Yes, or I can't trust your commitment."

Subhan: "A... revealing remark. But really, I don't think that this experimental result seems at all confusing, in light of the recent discussion of subjunctive objectivity—a discussion in which Eliezer strongly supported my position, by the way."

Obert: "Really? *I* thought Eliezer was finally coming out in favor of *my* position."

Subhan: "Huh? How do you get *that*?"

Obert: "The whole subtext of 'Math is Subjunctively Objective' is that morality is just like math! Sure, we compute morality inside our own brains—where else would we compute it? But just because we compute a quantity inside our own brains, doesn't mean that *what is computed* has a dependency on our own state of mind."

Subhan: "I think we must have been reading different *Overcoming Bias* posts! The whole subtext of 'Math is Subjunctively Objective' is to explain *away* why morality *seems* objective—to show that the *feeling* of a fixed given can arise without any external referent. When you *imagine* yourself thinking that killing is right, your brain-that-imagines hasn't *yet* been altered, so you carry out that moral imagination with your *current* brain, and conclude: 'Even if I thought killing were right, killing would still be wrong.' But this *doesn't* show that killing-is-wrong is a fixed fact from outside you."

Obert: "Like, say, 2 + 3 = 5 is a fixed fact. Eliezer wrote: 'If something *appears* to be the same regardless of what anyone thinks, then maybe that's because it *actually is* the same regardless of what anyone thinks.' I'd say that subtext is pretty clear!"

Subhan: "On the contrary. Naively, you might imagine your future self thinking differently of a thing, and visualize that the thing wouldn't thereby change, and conclude that the thing existed outside you. Eliezer shows how this is not *necessarily* the case. So you shouldn't *trust your intuition* that the thing is objective—it might be that the thing exists outside you, or it might *not*. It has to be *argued separately* from the feeling of subjunctive objectivity. In the case of 2 + 3 = 5, it's at least reasonable to wonder if math existed before humans. Physics itself seems to be made of math, and if we don't tell a story where physics was around before humans could observe it, it's hard to give a coherent account of how we got here. But there's not the slightest evidence that *morality* was at work in the universe before humans got here. *We* created it."

Obert: "I know some very wise children who would disagree with you."

Subhan: "Then they're wrong! If children learned in school that it was okay to steal, they would grow up believing it was okay to steal."

Obert: "Not if they saw that stealing hurt the other person, and felt empathy for their pain. Empathy is a human universal."

Subhan: "So we take a step back and say that evolution created the emotions that gave rise to morality, it doesn't put morality anywhere outside us. But what you say might not even be true—if theft weren't considered a crime, the other child might not feel so hurt by it. And regardless, it is rare to find any child capable of fully reconsidering the moral teachings of its society."

Obert: "I hear that, in a *remarkable similarity* to Eliezer, your parents were Orthodox Jewish and you broke with religion as a very young child."

Subhan: "I doubt that I was internally generating *de novo* moral philosophy. I was probably just wielding, against Judaism, the morality of the science fiction that actually socialized me."

Obert: "Perhaps you underestimate yourself. How much science fiction had you read at the age of five, when you realized it was dumb to recite Hebrew prayers you couldn't understand? Children may see errors that adults are too adept at fooling themselves to realize."

Subhan: "Hah! In all probability, if the teacher *had in fact* said that it was okay to take things from other children's backpacks, the children *would in fact* have thought it was right to steal."

Obert: "Even if true, that doesn't prove anything. It is quite coherent to simultaneously hold that:"

- "Stealing is wrong."
- "If a neutrino storm makes me believe 'stealing is right', then stealing is wrong."
- "If a neutrino storm makes me believe 'stealing is right', then I will say, 'If a neutrino storm makes me believe ''stealing is wrong'', then stealing is right.'"

Subhan: "Fine, it's *coherent*, but that doesn't mean it's *true*. The morality that the child has *in fact* learned from the teacher—or their parents, or the other children, or the television, or their parents' science fiction collection—doesn't say, 'Don't steal *because the teacher says so.*' The learned morality just says, 'Don't steal.' The cognitive procedure by which the children were taught to judge, does not have an internal dependency on what the children believe the teacher believes. That's why, in their moral imagination, it feels objective. But where did they acquire that morality in the first place? From the teacher!"

Obert: "So? I don't understand—you're saying that because they learned about morality from the teacher, they should think that morality has to be *about* the teacher? That they should think the teacher has the power to make it right to steal? How does that follow? It is quite coherent to simultaneously hold that—"

Subhan: "I'm saying that they got the morality *from the teacher!* Not from some mysterious light in the sky!"

Obert: "Look, I too read science fiction and fantasy as a child, and I think I may have been to some degree socialized by it—"

Subhan: "What a *remarkable coincidence*."

Obert: "The stories taught me that it was right to care about people who were different from me—aliens with strange shapes, aliens made of something other than carbon atoms, AIs who had been created rather than evolved, even things that didn't think like a human. But none of the stories ever said, 'You should care about people of different shapes and substrates *because science fiction told you to do it, and what science fiction says, goes.*' I wouldn't have bought that."

Subhan: "Are you sure you wouldn't have? That's how religion works."

Obert: "Didn't work on you. Anyway, the novels said to care about the aliens *because* they had inner lives and joys—or *because* I wouldn't want aliens to mistreat humans—or *because* shape and substrate never had anything to do with what makes a person a person. And you know, that still seems to me like a good justification."

Subhan: "Of course; you were *told* it was a good justification—maybe not directly, but the author showed other characters responding to the argument."

Obert: "It's not like the science fiction writers were making up their morality from scratch. They were working at the end of a chain of moral arguments and debates that stretches back to the Greeks, probably to before writing, maybe to before the dawn of modern humanity. You can *learn* morality, not just get pressed into it like a Jello mold. If you learn 2 + 3 = 5 from a teacher, it doesn't mean the teacher has the power to add two sheep to three sheep and get six sheep. If you would have spouted back '2 + 3 = 6' if the teacher said so, that doesn't change the sheep, it just means that you don't really understand the subject. So too with morality."

Subhan: "Okay, let me try a different tack. You, I take it, agree with both of these statements:"

- "If I preferred to kill people, it would not become right to kill people."
- "If I preferred to eat anchovy pizzas, it would become right to eat anchovy pizzas."

Obert: "Well, there are various caveats I'd attach to both of those. Like, in any circumstance where I really did prefer to kill someone, there'd be a high probability he was about to shoot me, or something. And there's all kinds of ways that eating an anchovy pizza could be wrong, like if I was already overweight. And I don't claim to be certain of anything when it comes to morality. But on the whole, and omitting all objections and knock-on effects, I agree."

Subhan: "It's that second statement I'm really interested in. How does your wanting to eat an anchovy pizza *make* it right?"

Obert: "Because ceteris paribus, in the course of ordinary life as we know it, and barring unspecified side effects, it is good for sentient beings to get what they want."

Subhan: "And why doesn't that apply to the bit about killing, then?"

Obert: "Because the other person doesn't want to die. Look, the whole reason why it's right *in the first place* for me to eat pepperoni pizza—the *original justification*—is that I enjoy doing so. Eating pepperoni pizza makes me happy, which is ceteris paribus a good thing. And eating anchovy pizza—blegh! Ceteris paribus, it's not good for sentient beings to experience disgusting tastes. But if my taste in pizza changes, that changes the consequneces of eating, which changes the moral justification, and so the moral judgment changes as well. But the reasons for not *killing* are in terms of the other person having an inner life that gets snuffed out—a fact that doesn't change depending on my own state of mind."

Subhan: "Oh? I was guessing that the difference had something to do with the social disapproval that would be leveled at murder, but not at eating anchovy pizza."

Obert: "As usual, your awkward attempts at rationalism have put you out of touch with self-evident moral truths. That's just not how I, or other real people, actually think! If I want to *bleep bleep bleep* a consenting adult, it doesn't matter whether society approves. Society can go *bleep bleep bleep bleep bleep* -"

Subhan: "Or so science fiction taught you."

Obert: "Spider Robinson's science fiction, to be precise. 'Whatever turns you on' shall be the whole of the law. So long as the 'you' is plural."

Subhan: "So that's where you got that particular self-evident moral truth. Was it also Spider Robinson who told you that it was self-evident?"

Obert: "No, I thought about that for a while, and then decided myself."

Subhan: "You seem to be paying remarkably close attention to what people *want.* Yet you insist that what validates this attention, is some external standard that makes the satisfaction of desires, *good.* Can't you just admit that, by empathy and vicarious experience and evolved fellow-feeling, you want others to get what they want? When does this external standard ever say that it's good for something to happen that someone *doesn't* want?"

Obert: "Every time you've got to tell your child to lay off the ice cream, he'll grow more fat cells that will make it impossible for him to lose weight as an adult."

Subhan: "And could something good happen that *no one* wanted?"

Obert: "I rather expect so. I don't think we're all *entirely* past our childhoods. In some ways the human species itself strikes me as being a sort of toddler in the 'No!' stage."

Subhan: "Look, there's a perfectly normal and non-mysterious chain of causality that describes where morality comes from, and it's not from outside humans. If you'd been told that killing was right, or if you'd evolved to enjoy killing—much more than we already do, I mean—or if you *really did* have a mini-stroke that damaged your frontal lobe, then you'd be going around saying, 'Killing is right regardless of what anyone thinks of it'. No great light in the sky would correct you. There is nothing else to the story."

Obert: "Really, I think that in this whole debate between us, there is surprisingly litle information to be gained by such observations as 'You only say that because your brain makes you say it.' If a neutrino storm hit me, I might say '2 + 3 = 6', but that wouldn't change arithmetic. It would just make my brain compute something other than arithmetic. And these various misfortunes that you've described, wouldn't change the crime of murder. They would just make my brain compute something other than morality."

Part of *The Metaethics Sequence*

Next post: "Changing Your Metaethics"

Previous post: "Math is Subjunctively Objective"

Changing Your Metaethics

If you say, "Killing people is wrong," that's morality. If you say, "You shouldn't kill people because God prohibited it," or "You shouldn't kill people because it goes against the trend of the universe", that's metaethics.

Just as there's far more agreement on Special Relativity than there is on the question "What is science?", people find it much easier to agree "Murder is bad" than to agree *what* makes it bad, or what it *means* for something to be bad.

People do get attached to their metaethics. Indeed they frequently insist that if their metaethic is wrong, all morality necessarily falls apart. It might be interesting to set up a panel of metaethicists—theists, Objectivists, Platonists, etc.—all of whom agree that killing is wrong; all of whom disagree on what it means for a thing to be "wrong"; and all of whom insist that if their metaethic is untrue, then morality falls apart.

Clearly a good number of people, if they are to make philosophical progress, will need to shift metathics at some point in their lives. *You* may have to do it.

At that point, it might be useful to have an open line of retreat—not a retreat from morality, but a retreat from Your-Current-Metaethic. (You know, the one that, if it is not true, leaves no possible basis for not killing people.)

And so I've been setting up these lines of retreat, in many and various posts, summarized below. For I have learned that to change metaethical beliefs is nigh-impossible in the presence of an unanswered attachment.

If, for example, someone believes the authority of "Thou Shalt Not Kill" derives from God, then there are several and well-known things to say that can help set up a line of retreat—as opposed to immediately attacking the plausibility of God. You can say, "Take personal responsibility! Even if you got orders from God, it would be your own decision to obey those orders. Even if God didn't order you to be moral, you could just be moral anyway."

The above argument actually generalizes to quite a number of metaethics—you just substitute Their-Favorite-Source-Of-Morality, or even the word "morality", for "God". Even if your particular source of moral authority failed, couldn't you just drag the child off the train tracks *anyway?* And indeed, who is it but you, that ever decided to follow this source of moral authority in the first place? What responsibility are you really passing on?

So the most important line of retreat is the one given in The Moral Void: If your metaethic stops telling you to save lives, you can just drag the kid off the train tracks anyway. To paraphrase Piers Anthony, only those who have moralities worry over whether or not they have them. If your metaethic tells you to kill people, why *should* you even listen? Maybe that which you would do even if there were no morality, *is* your morality.

The point being, of course, not that no morality exists; but that you can hold your will in place, and not fear losing sight of what's important to you, while your notions of the *nature* of morality change.

Other posts are there to set up lines of retreat specifically for more *naturalistic* metaethics. It may make more sense where I'm coming from on these, once I *actually* present my metaethic; but I thought it wiser to set them up in advance, to leave lines of retreat.

Joy in the Merely Real and Explaining vs. Explaining Away argue that you shouldn't be disappointed in any facet of life, just because it turns out to be *explicable* instead of inherently mysterious: for if we cannot take joy in the merely real, our lives shall be empty indeed.

No Universally Compelling Arguments sets up a line of retreat from the desire to have *everyone* agree with our moral arguments. There's a strong moral intuition which says that if our moral arguments are right, by golly, we ought to be able to *explain* them to people. This may be valid among humans, but you can't explain moral arguments to a rock. There is no ideal philosophy student of perfect emptiness who can be persuaded to implement modus ponens, starting without modus ponens. If a mind doesn't contain that which is moved by your moral arguments, it won't respond to them.

But then isn't all morality circular logic, in which case it falls apart? Where Recursive Justification Hits Bottom and My Kind of Reflection explain the difference between a self-consistent loop through the meta-level, and actual circular logic. You shouldn't find yourself saying "The universe is simple because it is simple", or "Murder is wrong because it is wrong"; but neither should you try to abandon Occam's Razor while evaluating the probability that Occam's Razor works, nor should you try to evaluate "Is murder wrong?" from somewhere outside your brain. There is no ideal philosophy student of perfect emptiness to which you can unwind yourself—try to find the perfect rock to stand upon, and you'll end up as a rock. So instead use the full force of your intelligence, your full rationality and your full morality, when you investigate the foundations of yourself.

The Gift We Give To Tomorrow sets up a line of retreat for those afraid to allow a *causal* role for evolution, in their account of how morality came to be. (Note that this is extremely distinct from granting evolution a *justificational* status in moral theories.) Love has to come into existence somehow—for if we cannot take joy in things that can come into existence, our lives will be empty indeed. Evolution may not be a particularly *pleasant* way for love to evolve, but judge the end product—not the source. Otherwise you would be committing what is known (appropriately) as The Genetic Fallacy: causation is not the same concept as justification. It's not like you can step outside the brain evolution gave you: Rebelling against nature is only possible from within nature.

The earlier series on Evolutionary Psychology should dispense with the metaethical confusion of believing that any normal human being thinks about their reproductive fitness, even unconsciously, in the course of making decisions. Only evolutionary biologists even know how to *define* genetic fitness, and they know better than to think it defines morality.

Alarming indeed is the thought that morality might be computed inside our own minds—doesn't this imply that morality is a mere thought? Doesn't it imply that whatever you think is right, must be right? Posts such as Does Your Morality Care What You Think? and its predecessors, Math is Subjunctively Objective and Probability is Subjectively Objective, set up the needed line of retreat: Just because a quantity is computed inside your head, doesn't mean that the quantity computed is *about* your

thoughts. There's a difference between a calculator that calculates "What is 2 + 3?" and "What do I output when someone presses '2', '+', and '3'?"

And finally Existential Angst Factory offers the notion that if life seems painful, reductionism may not be the real source of your problem—if living in a world of mere particles seems too unbearable, maybe your life isn't exciting enough on its own?

If all goes well, my next post will set up the metaethical question and its methodology, and I'll present my actual answer on Monday.

And if you're wondering why I deem this business of metaethics important, when it is all going to end up adding up to moral normality... telling you to pull the child off the train tracks, rather than the converse...

Well, there *is* opposition to rationality from people who think it drains meaning from the universe.

And this is a special case of a general phenomenon, in which many many people get messed up by misunderstanding where their morality comes from. Poor metaethics forms part of the teachings of many a cult, including the big ones. My target audience is not just people who are afraid that life is meaningless, but also those who've concluded that love is a delusion because real morality has to involve maximizing your inclusive fitness, or those who've concluded that unreturned kindness is evil because real morality arises only from selfishness, etc.

But the *real* reason, of course...

Setting Up Metaethics

Followup to: Is Morality Given?, Is Morality Preference?, Moral Complexities, Could Anything Be Right?, The Bedrock of Fairness, ...

Intuitions about morality seem to split up into two broad camps: morality-as-given and morality-as-preference.

Some perceive morality as a *fixed given*, independent of our whims, about which we form changeable *beliefs*. This view's great advantage is that it seems more normal up at the level of everyday moral conversations: it is the intuition underlying our everyday notions of "moral error", "moral progress", "moral argument", or "just because you want to murder someone doesn't make it *right*".

Others choose to describe morality as a *preference*—as a desire in some particular person; nowhere else is it written. This view's great advantage is that it has an easier time living with reductionism—fitting the notion of "morality" into a universe of mere physics. It has an easier time at the *meta* level, answering questions like "What is morality?" and "Where does morality come from?"

Both intuitions must contend with seemingly impossible questions. For example, Moore's Open Question: Even if you come up with some simple answer that fits on T-Shirt, like "Happiness is the sum total of goodness!", you would need to *argue* the identity. It isn't instantly obvious to everyone that goodness is happiness, which seems to indicate that happiness and rightness were different concepts to start with. What was that second concept, then, originally?

Or if "Morality is mere preference!" then *why care* about human preferences? How is it possible to establish any "ought" at all, in a universe seemingly of mere "is"?

So what we should want, ideally, is a metaethic that:

1. Adds up to moral normality, including moral errors, moral progress, and things you should do whether you want to or not;
2. Fits naturally into a non-mysterious universe, postulating no exception to reductionism;
3. Does not oversimplify humanity's complicated moral arguments and many terminal values;
4. Answers all the impossible questions.

I'll present that view tomorrow.

Today's post is devoted to setting up the question.

Consider "free will", already dealt with in these posts. On one level of organization, we have mere physics, particles that make no choices. On another level of organization, we have human minds that extrapolate possible futures and choose between them. How can we control anything, even our own choices, when the universe is deterministic?

To dissolve the puzzle of free will, you have to simultaneously imagine two levels of organization while keeping them conceptually distinct. To get it on a gut level, you have to *see* the level transition—the way in which free will *is* how the human decision

algorithm feels from inside. (Being told flatly "one level emerges from the other" just relates them by a magical transition rule, "emergence".)

For free will, the key is to understand how your brain computes whether you "could" do something—the algorithm that labels reachable states. Once you understand this label, it does not appear particularly meaningless—"could" makes sense—and the label does not conflict with physics following a deterministic course. If you can *see* that, you can *see* that there is no conflict between your feeling of freedom, and deterministic physics. Indeed, I am perfectly willing to say that the feeling of freedom is correct, when the feeling is interpreted correctly.

In the case of morality, once again there are two levels of organization, seemingly quite difficult to fit together:

On one level, there are just particles without a shred of *should*-ness built into them—just like an electron has no notion of what it "could" do—or just like a flipping coin is not uncertain of its own result.

On another level is the ordinary morality of everyday life: moral errors, moral progress, and things you ought to do whether you want to do them or not.

And in between, the level transition question: What is this *should*-ness stuff?

Award yourself a point if you thought, "But wait, that problem isn't quite analogous to the one of free will. With free will it was just a question of factual investigation—look at human psychology, figure out how it *does in fact* generate the feeling of freedom. But here, it won't be enough to figure out how the mind generates its feelings of should-ness. Even after we know, we'll be left with a remaining question—is that how we *should* calculate should-ness? So it's not just a matter of sheer factual reductionism, it's a moral question."

Award yourself *two* points if you thought, "...oh, wait, I recognize *that* pattern: It's one of those strange loops through the meta-level we were talking about earlier."

And if you've been reading along this whole time, you know the answer isn't going to be, "Look at this *fundamentally* moral stuff!"

Nor even, "Sorry, morality is *mere* preference, and right-ness is just what serves you or your genes; all your moral intuitions otherwise are wrong, but I won't explain where they come from."

Of the art of answering impossible questions, I have already said much: Indeed, vast segments of my *Overcoming Bias* posts were created with that specific hidden agenda.

The sequence on anticipation fed into Mysterious Answers to Mysterious Questions, to prevent the Primary Catastrophic Failure of stopping on a poor answer.

The Fake Utility Functions sequence was directed at the problem of oversimplified moral answers particularly.

The sequence on words provided the first and basic illustration of the Mind Projection Fallacy, the understanding of which is one of the Great Keys.

The sequence on words also showed us how to play Rationalist's Taboo, and Replace the Symbol with the Substance. What is "right", if you can't say "good" or "desirable" or "better" or "preferable" or "moral" or "should"? What happens if you try to carry out the operation of replacing the symbol with what it stands for?

And the sequence on quantum physics, among other purposes, was there to teach the fine art of not *running away* from Scary and Confusing Problems, even if others have failed to solve them, even if great minds failed to solve them for generations. Heroes screw up, time moves on, and each succeeding era gets an entirely new chance.

If you're just joining us here (Belldandy help you) then you might want to think about reading all those posts before, oh, say, tomorrow.

If you've been reading this whole time, then you should think about trying to dissolve the question on your own, before tomorrow. It doesn't require more than 96 insights beyond those already provided.

Next: *The Meaning of Right.*

Part of *The Metaethics Sequence*

Next post: "The Meaning of Right"

Previous post: "Changing Your Metaethics"

The Meaning of Right

Continuation of: Changing Your Metaethics, Setting Up Metaethics
Followup to: Does Your Morality Care What You Think?, The Moral Void, Probability is Subjectively Objective, Could Anything Be Right?, The Gift We Give To Tomorrow, Rebelling Within Nature, Where Recursive Justification Hits Bottom, ...

(The culmination of a *long* series of *Overcoming Bias* posts; if you start here, I accept no responsibility for any resulting confusion, misunderstanding, or unnecessary angst.)

What *is* morality? What does the word "should", *mean*? The many pieces are in place: This question I shall now dissolve.

The key—as it has always been, in my experience so far—is to understand how a certain cognitive algorithm feels from inside. Standard procedure for righting a wrong question: If you don't know what right-ness is, then take a step beneath and ask how your brain labels things "right".

It is not the *same* question—it has no moral aspects to it, being strictly a matter of fact and cognitive science. But it is an *illuminating* question. Once we know how our brain labels things "right", perhaps we shall find it easier, afterward, to ask what is really and truly *right.*

But with that said—the easiest way to begin investigating *that* question, will be to jump back up to the level of morality and ask what *seems* right. And if that seems like too much recursion, get used to it—the other 90% of the work lies in handling recursion properly.

(Should you find your grasp on meaningfulness wavering, at any time following, check Changing Your Metaethics for the appropriate prophylactic.)

So! In order to investigate how the brain labels things "right", we are going to *start out* by talking about what is right. That is, we'll start out wearing our *morality-goggles,* in which we consider morality-as-morality and talk about moral questions directly. As opposed to wearing our *reduction-goggles,* in which we talk about cognitive algorithms and mere physics. Rigorously distinguishing between these two views is the first step toward mating them together.

As a first step, I offer this observation, on the level of morality-as-morality: Rightness is contagious backward in time.

Suppose there is a switch, currently set to OFF, and it is *morally desirable* for this switch to be flipped to ON. Perhaps the switch controls the emergency halt on a train bearing down on a child strapped to the railroad tracks, this being my canonical example. If this is the case, then, *ceteris paribus* and presuming the absence of exceptional conditions or further consequences that were not explicitly specified, we may consider it *right* that this switch should be flipped.

If it is right to flip the switch, then it is right to pull a string that flips the switch. If it is good to pull a string that flips the switch, it is right and proper to press a button that pulls the string: Pushing the button seems to have more *should*-ness than not pushing it.

It seems that—all else being equal, and assuming no other consequences or exceptional conditions which were not specified—value flows backward along arrows of causality.

Even in deontological moralities, if you're *obligated* to save the child on the tracks, then you're *obligated* to press the button. Only *very* primitive AI systems have motor outputs controlled by strictly local rules that don't model the future *at all*. Duty-based or virtue-based ethics are only *slightly* less consequentialist than consequentialism. It's hard to say whether moving your arm left or right is more virtuous without talking about what happens next.

> Among my readers, there may be some who presently assert—though I hope to persuade them otherwise—that the life of a child is of no value to them. If so, they may substitute anything else that they prefer, at the end of the switch, and ask if they should press the button.
>
> But I also suspect that, among my readers, there are some who wonder if the *true* morality might be something quite different from what is presently believed among the human kind. They may find it imaginable—plausible?—that human life is of no value, or negative value. They may wonder if the goodness of human happiness, is as much a self-serving delusion as the justice of slavery.
>
> I myself was once numbered among these skeptics, because I was always very suspicious of anything that looked self-serving.
>
> Now here's a little question I never thought to ask, during those years when I thought I knew nothing about morality:
>
> Could make sense to have a morality in which, if we *should* save the child from the train tracks, then we *should not* flip the switch, *should* pull the string, and *should not* push the button, so that, finally, we do not push the button?
>
> Or perhaps someone says that it is better to save the child, than to not save them; but doesn't see why anyone would think this implies it is better to press the button than not press it. (Note the resemblance to the Tortoise who denies *modus ponens*.)
>
> It seems imaginable, to at least some people, that entirely different things could be *should*. It didn't seem nearly so imaginable, at least to me, that *should*-ness could fail to flow backward in time. When I was trying to question everything else, that thought simply did not occur to me.
>
> Can you question it? Should you?

Every now and then, in the course of human existence, we question what *should* be done and what is *right* to do, what is *better* or *worse*; others come to us with assertions along these lines, and we question them, asking "Why is it right?" Even when we believe a thing is right (because someone told us that it is, or because we wordlessly feel that it is) we may still question why it is right.

Should-ness, it seems, flows backward in time. This gives us one way to question why or whether a particular event has the *should*-ness property. We can look for some *consequence* that has the *should*-ness property. If so, the *should*-ness of the original event seems to have been plausibly proven or explained.

Ah, but what about the consequence—why is *it* should? Someone comes to you and says, "You should give me your wallet, because then I'll have your money, and I should have your money." If, at this point, you stop asking questions about *should*-ness, you're vulnerable to a moral mugging.

So we keep asking the next question. Why should we press the button? To pull the string. Why should we pull the string? To flip the switch. Why should we flip the switch? To pull the child from the railroad tracks. Why pull the child from the railroad tracks? So that they live. Why should the child live?

Now there are people who, caught up in the enthusiasm, go ahead and answer that question in the same style: for example, "Because the child might eventually grow up and become a trade partner with you," or "Because you will gain honor in the eyes of others," or "Because the child may become a great scientist and help achieve the Singularity," or some such. But even if we were to answer in this style, it would only beg the next question.

Even if you try to have a chain of *should* stretching into the infinite future—a trick I've yet to see anyone try to pull, by the way, though I may be only ignorant of the breadths of human folly—then you would simply ask "Why that chain rather than some other?"

Another way that something can be *should*, is if there's a general rule that makes it *should*. If your belief pool starts out with the general rule "All children X: It is better for X to live than to die", then it is quite a short step to "It is better for Stephanie to live than to die". Ah, but why save all children? Because they may all become trade partners or scientists? But then where did *that* general rule come from?

If *should*-ness only comes from *should*-ness—from a *should*-consequence, or from a *should*-universal—then how does anything end up *should* in the first place?

Now human beings have argued these issues for thousands of years and maybe much longer. We do not hesitate to continue arguing when we reach a terminal value (something that has a charge of *should*-ness independently of its consequences). We just go on arguing about the universals.

I usually take, as my archetypal example, the undoing of slavery: Somehow, slaves' lives went from having no value to having value. Nor do I think that, back at the dawn of time, anyone was even trying to argue that slaves were better off being slaves (as it would be latter argued). They'd probably have looked at you like you were crazy if you even tried. Somehow, we got from there, to here...

And some of us would even hold this up as a case of moral progress, and look at our ancestors as having made a *moral error.* Which seems easy enough to describe in terms of *should*-ness: Our ancestors *thought* that they should enslave defeated enemies, but they were mistaken.

But all our philosophical arguments ultimately seem to ground in statements that no one has bothered to justify—except perhaps to plead that they are *self-evident,* or that any *reasonable* mind must surely agree, or that they are *a priori* truths, or some such. Perhaps, then, *all* our moral beliefs are as erroneous as that old bit about slavery? Perhaps we have entirely misperceived the flowing streams of *should?*

So I once believed was plausible; and one of the arguments I wish I could go back and say to myself, is, "If you know nothing at all about should-ness, then how do you know

that the procedure, 'Do whatever Emperor Ming says' is not the entirety of should-ness? Or even worse, perhaps, the procedure, 'Do whatever maximizes inclusive genetic fitness' or 'Do whatever makes you personally happy'." The point here would have been to make my past self see that in *rejecting* these rules, he was asserting a kind of knowledge—that to say, "This is *not* morality," he must reveal that, despite himself, he knows something about morality or meta-morality. Otherwise, the procedure "Do whatever Emperor Ming says" would seem *just* as plausible, as a guiding principle, as his current path of "Rejecting things that seem unjustified." Unjustified—according to what criterion of *justification?* Why trust the principle that says that moral statements need to be justified, if you know nothing at all about morality?

What indeed would distinguish, *at all,* the question "What is right?" from "What is wrong?"

What is "right", if you can't say "good" or "desirable" or "better" or "preferable" or "moral" or "should"? What happens if you try to carry out the operation of replacing the symbol with what it stands for?

If you're guessing that I'm trying to inveigle you into letting me say: "Well, there are just some things that are baked into the *question,* when you start asking questions about *morality,* rather than wakalixes or toaster ovens", then you would be right. I'll be making use of that later, and, yes, will address "But why *should* we ask that question?"

Okay, now: morality-goggles off, reduction-goggles on.

Those who remember Possibility and Could-ness, or those familiar with simple search techniques in AI, will realize that the "should" label is behaving like the inverse of the "could" label, which we previously analyzed in terms of "reachability". Reachability spreads *forward* in time: if I could reach the state with the button pressed, I could reach the state with the string pulled; if I could reach the state with the string pulled, I could reach the state with the switch flipped.

Where the "could" label and the "should" label collide, the algorithm produces a plan.

> Now, as I say this, I suspect that at least some readers may find themselves fearing that I am about to reduce *should*-ness to a *mere* artifact of a way that a planning system feels from inside. Once again I urge you to check Changing Your Metaethics, if this starts to happen. Remember above all the Moral Void: Even if there were no morality, you could still choose to help people rather than hurt them. This, above all, holds in place what you hold precious, while your beliefs about the nature of morality change.
>
> I do not intend, with this post, to take away anything of value; it will all be given back before the end.

Now this algorithm is not very sophisticated, as AI algorithms go, but to apply it in full generality—to learned information, not just ancestrally encountered, genetically programmed situations—is a rare thing among animals. Put a food reward in a transparent box. Put the matching key, which looks unique and uniquely corresponds to that box, in another transparent box. Put the unique key to *that* box in another box. Do this with five boxes. Mix in another sequence of five boxes that doesn't lead to a food reward. Then offer a choice of two keys, one of which starts the sequence of five boxes leading to food, one of which starts the sequence leading nowhere.

Chimpanzees can learn to do this, but so far as I know, no non-primate species can pull that trick.

And as smart as chimpanzees are, they are not quite as good as humans at inventing plans—plans such as, for example, planting in the spring to harvest in the fall.

So what else are humans doing, in the way of planning?

It is a general observation that natural selection seems to *reuse* existing complexity, rather than creating things from scratch, whenever it *possibly* can—though not always in the *same way* that a human engineer would. It is a function of the enormous time required for evolution to create machines with many interdependent parts, and the vastly shorter time required to create a mutated copy of something already evolved.

What else are humans doing? Quite a bit, and some of it I don't understand—there are plans humans make, that no modern-day AI can.

But *one* of the things we are doing, is reasoning about "right-ness" the same way we would reason about any other observable property.

Are animals with bright colors often poisonous? Does the delicious nid-nut grow only in the spring? Is it usually a good idea to take with a waterskin on long hunts?

It seems that Martha and Fred have an obligation to take care of their child, and Jane and Bob are obligated to take care of their child, and Susan and Wilson have a duty to care for their child. Could it be that parents in general must take care of their children?

By representing right-ness as an attribute of objects, you can recruit a whole previously evolved system that reasons about the attributes of objects. You can save quite a lot of planning time, if you decide (based on experience) that *in general* it is a good idea to take a waterskin on hunts, from which it follows that it must be a good idea to take a waterskin on hunt #342.

Is this damnable for a Mind Projection Fallacy—treating properties of the mind as if they were out there in the world?

Depends on how you look at it.

This business of, "It's been a good idea to take waterskins on the last three hunts, maybe it's a good idea in general, if so it's a good idea to take a waterskin on this hunt", does seem to *work*.

Let's say that your mind, faced with any countable set of objects, automatically and perceptually tagged them with their remainder modulo 5. If you saw a group of 17 objects, for example, they would look *remainder-2-ish.* Though, if you didn't have any notion of *what* your neurons were doing, and perhaps no notion of modulo arithmetic, you would only see that the group of 17 objects had the same *remainder*-ness as a group of 2 objects. You might not even know how to count—your brain doing the whole thing automatically, subconsciously and neurally—in which case you would just have five different words for the *remainder*-ness attributes that we would call 0, 1, 2, 3, and 4.

If you look out upon the world you see, and guess that *remainder*-ness is a separate and additional attribute of things—like the attribute of having an electric charge—or

like a tiny little XML tag hanging off of things—then you will be wrong. But this does not mean it is nonsense to talk about *remainder*-ness, or that you must automatically commit the Mind Projection Fallacy in doing so. So long as you've got a well-defined way to compute a property, it can have a well-defined output and hence an empirical truth condition.

If you're looking at 17 objects, then their *remainder*-ness is, indeed and truly, 2, and not 0, 3, 4, or 1. If I tell you, "Those red things you told me to look at are *remainder-2-ish*", you have indeed been told a falsifiable and empirical property of those red things. It is just not a separate, additional, physically existent attribute.

And as for reasoning *about* derived properties, and which other inherent or derived properties they correlate to—I don't see anything inherently fallacious about that.

One may notice, for example, that things which are 7 modulo 10 are often also 2 modulo 5. *Empirical* observations of this sort play a large role in mathematics, suggesting theorems to prove. (See Polya's *How To Solve It*.)

Indeed, virtually all the experience we have, is derived by complicated neural computations from the raw physical events impinging on our sense organs. By the time you *see* anything, it has been extensively processed by the retina, lateral geniculate nucleus, visual cortex, parietal cortex, and temporal cortex, into a very complex sort of derived computational property.

If you thought of a property like *redness* as residing strictly *in an apple*, you would be committing the Mind Projection Fallacy. The apple's surface has a reflectance which sends out a mixture of wavelengths that impinge on your retina and are processed with respect to ambient light to extract a summary color of *red...* But if you tell me that the apple is red, rather than green, and make no claims as to whether this is an ontologically fundamental physical attribute of the apple, then I am quite happy to agree with you.

So as long as there is a stable computation involved, or a stable process—even if you can't consciously verbalize the specification—it often makes a great deal of sense to talk about properties that are not fundamental. And reason about them, and remember where they have been found in the past, and guess where they will be found next.

> (In retrospect, that should have been a separate post in the Reductionism sequence. "Derived Properties", or "Computational Properties" maybe. Oh, well; I promised you morality this day, and this day morality you shall have.)

Now let's say we want to make a little machine, one that will save the lives of children. (This enables us to save more children than we could do without a machine, just like you can move more dirt with a shovel than by hand.) The machine will be a planning machine, and it will reason about events that may or may not have the property, *leads-to-child-living.*

A simple planning machine would just have a pre-made model of the environmental process. It would search forward from its actions, applying a label that we might *call* "reachable-from-action-ness", but which might as well say "Xybliz" internally for all that it matters to the program. And it would search backward from scenarios, situations, in which the child lived, labeling these "leads-to-child-living". If situation X leads to situation Y, and Y has the label "leads-to-child-living"—which might just be a little flag bit, for all the difference it would make—then X will inherit the flag from Y.

When the two labels meet in the middle, the leads-to-child-living flag will quickly trace down the stored path of reachability, until finally some particular sequence of actions ends up labeled "leads-to-child-living". Then the machine automatically executes those actions—that's just what the machine does.

Now this machine is not complicated enough to feel existential angst. It is not complicated enough to commit the Mind Projection Fallacy. It is not, in fact, complicated enough to *reason abstractly* about the property "leads-to-child-living-ness". The machine—as specified so far—does not notice if the action "jump in the air" turns out to always have this property, or never have this property. If "jump in the air" always led to situations in which the child lived, this could greatly simplify future planning—but only if the machine were sophisticated enough to notice this fact and use it.

If it is a fact that "jump in the air" "leads-to-child-living-ness", this fact is composed of empirical truth and logical truth. It is an *empirical* truth that if the world is such that if you perform the (ideal abstract) algorithm "trace back from situations where the child lives", then it will be a *logical* truth about the output of this (ideal abstract) algorithm that it labels the "jump in the air" action.

(You cannot always define this fact in *entirely* empirical terms, by looking for the physical real-world coincidence of jumping and child survival. It might be that "stomp left" *also* always saves the child, and the machine in fact stomps left. In which case the fact that jumping in the air *would have* saved the child, is a counterfactual extrapolation.)

Okay, now we're ready to bridge the levels.

As you must surely have guessed by now, this *should*-ness stuff is how the human decision algorithm feels from inside. It is not an extra, physical, ontologically fundamental attribute hanging off of events like a tiny little XML tag.

But it is a *moral* question what we should do about that—how we should react to it.

To adopt an attitude of complete nihilism, because we *wanted* those tiny little XML tags, and they're *not physically there,* strikes me as the wrong move. It is like supposing that the absence of an XML tag, equates to the XML tag *being there*, saying in its tiny brackets *what value we should attach,* and having value zero. And then this value zero, in turn, equating to a moral imperative to wear black, feel awful, write gloomy poetry, betray friends, and commit suicide.

No.

So what would I say instead?

The force behind my answer is contained in The Moral Void and The Gift We Give To Tomorrow. I would try to save lives "even if there were no morality", as it were.

And it seems like an awful shame to—after so many millions and hundreds of millions of years of evolution—after the moral miracle of so much cutthroat genetic competition producing intelligent minds that love, and hope, and appreciate beauty, and create beauty—after coming so far, to throw away the Gift of morality, *just because our brain happened to represent morality in such fashion as to potentially mislead us when we reflect on the nature of morality.*

This little accident of the Gift doesn't seem like a good reason to throw away the Gift; it certainly isn't a inescapable logical justification for wearing black.

Why not keep the Gift, but adjust the way we reflect on it?

So here's my metaethics:

I earlier asked,

> What is "right", if you can't say "good" or "desirable" or "better" or "preferable" or "moral" or "should"? What happens if you try to carry out the operation of replacing the symbol with what it stands for?

I answer that if you try to replace the symbol "should" with *what it stands for*, you end up with quite a large sentence.

For the much simpler save-life machine, the "should" label stands for leads-to-child-living-ness.

For a human this is a much huger blob of a computation that looks like, "Did everyone survive? How many people are happy? Are people in control of their own lives? ..." Humans have complex emotions, have many values—the thousand shards of desire, the godshatter of natural selection. I would say, by the way, that the huge blob of a computation is not just my present terminal values (which I don't really *have*—I am not a consistent expected utility maximizers); the huge blob of a computation includes the specification of those moral arguments, those justifications, that would sway me if I heard them. So that I can regard my present values, as an approximation to the ideal morality that I would have if I heard all the arguments, to whatever extent such an extrapolation is coherent.

No one can write down their big computation; it is not just too large, it is also unknown to its user. No more could you print out a listing of the neurons in your brain. You never *mention* your big computation—you only *use* it, every hour of every day.

Now why might one *identify* this enormous abstract computation, with what-is-*right*?

If you identify rightness with this *huge computational property*, then moral judgments are subjunctively objective (like math), subjectively objective (like probability), and capable of being true (like counterfactuals).

You will find yourself saying, "If I wanted to kill someone—even if I thought it was right to kill someone—that wouldn't make it right." Why? Because what is *right* is a huge computational property—an *abstract* computation—*not* tied to the state of anyone's brain, including your own brain.

This distinction was introduced earlier in 2-Place and 1-Place Words. We can treat the word "sexy" as a 2-place function that goes out and hoovers up someone's sense of sexiness, and then eats an object of admiration. Or we can treat the word "sexy" as *meaning* a 1-place function, a *particular* sense of sexiness, like Sexiness_20934, that only accepts one argument, an object of admiration.

Here we are treating morality as a *1-place function.* It does not accept a person as an argument, spit out whatever cognitive algorithm they use to choose between actions, and then apply that algorithm to the situation at hand. When I say *right,* I mean a certain *particular* 1-place function that just asks, "Did the child live? Did anyone else

get killed? Are people happy? Are they in control of their own lives? Has justice been served?" ... and so on through many, many other elements of rightness. (And perhaps those arguments that might persuade me otherwise, which I have not heard.)

Hence the notion, "Replace the symbol with what it stands for."

Since what's *right* is a 1-place function, if I subjunctively imagine a world in which someone has slipped me a pill that makes me want to kill people, then, in this subjunctive world, it is not *right* to kill people. That's not merely because I'm judging with my current brain. It's because when I say *right,* I am referring to a 1-place function. Rightness doesn't go out and hoover up the current state of my brain, in this subjunctive world, before producing the judgment "Oh, wait, it's now okay to kill people." When I say *right*, I don't *mean* "that which my future self wants", I *mean* the function that looks at a situation and asks, "Did anyone get killed? Are people happy? Are they in control of their own lives? ..."

And once you've defined a particular abstract computation that says what is *right*—or even if you haven't defined it, and it's computed in some part of your brain you can't perfectly print out, but the computation is *stable*—more or less—then as with any other derived property, it makes sense to speak of a moral judgment being *true.* If I say that today was a good day, you've learned something empirical and falsifiable about my day—if it turns out that actually my grandmother died, you will suspect that I was originally lying.

The apparent objectivity of morality has just been explained—and *not* explained *away.* For indeed, if someone slipped me a pill that made me want to kill people, nonetheless, it would not be *right* to kill people. Perhaps I would actually kill people, in that situation—but that is because something other than morality would be controlling my actions.

Morality is not just subjunctively objective, but subjectively objective. I experience it as something I cannot change. Even after I know that it's myself who computes this 1-place function, and not a rock somewhere—even after I know that I will not find any star or mountain that computes this function, that only upon me is it written—even so, I find that I wish to save lives, and that even if I could change this by an act of will, I would not choose to do so. I do not wish to reject joy, or beauty, or freedom. What else would I do instead? I do not wish to reject the Gift that natural selection accidentally barfed into me. This is the principle of The Moral Void and The Gift We Give To Tomorrow.

Our origins may seem unattractive, our brains untrustworthy.

But love has to enter the universe somehow, starting from non-love, or love cannot enter time.

And if our brains are untrustworthy, it is only our own brains that say so. Do you sometimes think that human beings are not very nice? Then it is you, a human being, who says so. It is you, a human being, who judges that human beings could do better. You will not find such written upon the stars or the mountains: they are not minds, they cannot think.

In this, of course, we find a justificational strange loop through the meta-level. Which is unavoidable so far as I can see—you can't argue morality, or any kind of goal optimization, into a rock. But note the exact structure of this strange loop: *there is no general moral principle which says that you should do what evolution programmed*

you to do. There is, indeed, no general principle to trust your moral intuitions! You can find a moral intuition within yourself, describe it—quote it—consider it deliberately and in the full light of your entire morality, and reject it, on grounds of other arguments. What counts as an argument is also built into the rightness-function.

Just as, in the strange loop of rationality, there is no general principle in rationality to trust your brain, or to believe what evolution programmed you to believe—but indeed, when you ask which parts of your brain you need to rebel against, you do so using your current brain. When you ask whether the universe is simple, you can consider the *simple* hypothesis that the universe's apparent simplicity is explained by its actual simplicity.

Rather than trying to unwind ourselves into rocks, I proposed that we should use the *full strength* of our current rationality, in reflecting upon ourselves—that no part of ourselves be immune from examination, and that we use all of ourselves that we currently believe in to examine it.

You would do the same thing with morality; if you consider that a part of yourself might be considered harmful, then use your *best* current guess at what is *right*, your full moral strength, to do the considering. Why *should* we want to unwind ourselves to a rock? Why *should* we do less than our best, when reflecting? You can't unwind past Occam's Razor, modus ponens, or morality *and it's not clear why you should try.*

For any part of rightness, you can always imagine another part that overrides it—it would not be right to drag the child from the train tracks, if this resulted in everyone on Earth becoming unable to love—or so I would judge. For every part of rightness you examine, you will find that it cannot be the sole and perfect and only criterion of rightness. This may lead to the incorrect inference that there is something beyond, some perfect and only criterion from which all the others are derived—but that does not follow. The whole is the sum of the parts. We ran into an analogous situation with free will, where no part of ourselves seems perfectly decisive.

The classic dilemma for those who would trust their moral intuitions, I believe, is the one who says: "Interracial marriage is repugnant—it disgusts me—and that is my moral intuition!" I reply, "There is no general rule to obey your intuitions. You just *mentioned* intuitions, rather than *using* them. Very few people have legitimate cause to *mention* intuitions—Friendly AI programmers, for example, delving into the cognitive science of things, have a legitimate reason to mention them. Everyone else just has ordinary moral arguments, in which they *use* their intuitions, for example, by saying, 'An interracial marriage doesn't hurt anyone, if both parties consent'. I do not say, 'And I have an intuition that anything consenting adults do is right, and all intuitions must be obeyed, therefore I win.' I just offer up that argument, and any others I can think of, to weigh in the balance."

Indeed, evolution that made us cannot be trusted—so there is no general principle to trust it! Rightness is not defined in terms of automatic correspondence to any possible decision we actually make—so there's no general principle that says you're infallible! Just do what is, ahem, *right*—to the best of your ability to weigh the arguments you have heard, and ponder the arguments you may not have heard.

If you were hoping to have a perfectly trustworthy system, or to have been created in correspondence with a perfectly trustworthy morality—well, I can't give *that* back to you; but even most religions don't try that one. Even most religions have the human psychology containing elements of sin, and even most religions don't *actually* give you

an effectively executable and perfect procedure, though they may tell you "Consult the Bible! It always works!"

If you hoped to find a source of morality outside humanity—well, I can't give that back, but I can ask once again: Why would you even *want* that? And what good would it do? Even if there were some great light in the sky—something that could tell us, "Sorry, happiness is bad for you, pain is better, now get out there and kill some babies!"—it would still be your own decision to follow it. You cannot evade responsibility.

There isn't enough mystery *left* to justify *reasonable doubt* as to whether the causal origin of morality is something outside humanity. We have evolutionary psychology. We know where morality came from. We pretty much know how it works, in broad outline at least. We know there are no little XML value tags on electrons (and indeed, even if you found them, why *should* you pay attention to what is written there?)

If you hoped that morality would be universalizable—sorry, that one I *really* can't give back. Well, unless we're just talking about humans. Between neurologically intact *humans,* there is indeed much cause to hope for overlap and coherence; and a great and reasonable doubt as to whether any present disagreement is *really* unresolvable, even it seems to be about "values". The obvious reason for hope is the psychological unity of humankind, and the intuitions of symmetry, universalizability, and simplicity that we execute in the course of our moral arguments. (In retrospect, I should have done a post on Interpersonal Morality before this...)

If I tell you that three people have found a pie and are arguing about how to divide it up, the thought "Give one-third of the pie to each" is bound to occur to you—and if the three people are humans, it's bound to occur to them, too. If one of them is a psychopath and insists on getting the whole pie, though, there may be nothing for it but to say: "Sorry, fairness is not 'what everyone thinks is fair', fairness is everyone getting a third of the pie". You might be able to resolve the remaining disagreement by politics and game theory, short of violence—but that is not the same as coming to agreement on values. (Maybe you could persuade the psychopath that taking a pill to be more human, if one were available, would make them happier? Would you be justified in forcing them to swallow the pill? These get us into stranger waters that deserve a separate post.)

If I define rightness to include the space of arguments that move me, then when you and I argue about *what is right,* we are arguing our *approximations* to what we would come to believe if we knew all empirical facts and had a million years to think about it—and that might be a lot closer than the present and heated argument. Or it might not. This gets into the notion of 'construing an extrapolated volition' which would be, again, a separate post.

But if you were stepping outside the human and hoping for moral arguments that would persuade any possible mind, even a mind that just wanted to maximize the number of paperclips in the universe, then sorry—the space of possible mind designs is too large to permit universally compelling arguments. You are better off treating your intuition that your moral arguments ought to persuade others, as applying only to other humans who are more or less neurologically intact. Trying it on human psychopaths would be dangerous, yet perhaps possible. But a paperclip maximizer is just not the sort of mind that would be moved by a *moral* argument. (This will definitely be a separate post.)

Once, in my wild and reckless youth, I tried dutifully—I thought it was my duty—to be ready and willing to follow the dictates of a great light in the sky, an external objective morality, when I discovered it. I questioned everything, even altruism toward human lives, even the value of happiness. Finally I realized that there was no foundation but humanity—no evidence pointing to even a reasonable doubt that there was anything else—and indeed I shouldn't even *want* to hope for anything else—and indeed would have no moral cause to follow the dictates of a light in the sky, even if I found one.

I didn't get back *immediately* all the pieces of myself that I had tried to deprecate—it took time for the realization "There is nothing else" to sink in. The notion that humanity could just... you know... live and have fun... seemed much too good to be true, so I mistrusted it. But eventually, it sank in that there really *was* nothing else to take the place of beauty. And then I got it back.

So you see, it all really *does* add up to moral normality, very exactly in fact. You go on with the same morals as before, and the same moral arguments as before. There is no sudden Grand Overlord Procedure to which you can appeal to get a perfectly trustworthy answer. You don't know, cannot print out, the great rightness-function; and even if you could, you would not have enough computational power to search the entire specified space of arguments that might move you. You will just have to argue it out.

I suspect that a fair number of those who propound metaethics do so in order to have it add up to some new and unusual moral—else why would they bother? In my case, I bother because I am a Friendly AI programmer and I have to make a physical system outside myself do what's right; for which purpose metaethics becomes very important indeed. But for the most part, the effect of my proffered metaethic is threefold:

- Anyone worried that reductionism drains the meaning from existence can stop worrying;
- Anyone who was rejecting parts of their human existence based on strange metaethics—i.e., "Why should I care about others, if that doesn't help me maximize my inclusive genetic fitness?"—can welcome back all the parts of themselves that they once exiled.
- You can stop arguing about metaethics, and go back to whatever ordinary moral argument you were having before then. This knowledge will help you avoid metaethical *mistakes* that mess up moral arguments, but you can't actually use it to *settle debates* unless you can build a Friendly AI.

And, oh yes—*why* is it *right* to save a child's life?

Well... you could ask "Is this event that just happened, right?" and find that the child had survived, in which case you would have discovered the nonobvious empirical fact about the world, that it had come out right.

Or you could start out already knowing a complicated state of the world, but still have to apply the rightness-function to it in a nontrivial way—one involving a complicated moral argument, or extrapolating consequences into the future—in which case you would learn the nonobvious logical / computational fact that rightness, applied to this situation, yielded thumbs-up.

In both these cases, there are nonobvious facts to learn, which seem to *explain* why what just happened is *right.*

But if you ask "Why is it good to be happy?" and then replace the symbol 'good' with what it stands for, you'll end up with a question like "Why does happiness match {happiness + survival + justice + individuality + ...}?" This gets computed so fast, that it scarcely seems like there's anything there to be explained. It's like asking "Why does 4 = 4?" instead of "Why does 2 + 2 = 4?"

Now, I bet that feels quite a bit like what happens when I ask you: "Why is happiness good?"

Right?

And that's also my answer to Moore's Open Question. Why is this big function I'm talking about, *right*? Because when I say "that big function", and you say "right", we are dereferencing two different pointers to the same unverbalizable abstract computation. I mean, that big function I'm talking about, happens to be the same thing that labels things *right* in your own brain. You might reflect on the pieces of the quotation of the big function, but you would start out by using your sense of *right*-ness to do it. If you had the perfect empirical knowledge to taboo both "that big function" and "right", substitute what the pointers stood for, and write out the full enormity of the resulting sentence, it would come out as... sorry, I can't resist this one... A=A.

Part of *The Metaethics Sequence*

Next post: "Interpersonal Morality"

Previous post: "Setting Up Metaethics"

Interpersonal Morality

Followup to: The Bedrock of Fairness

Every time I wonder if I really need to do so much prep work to explain an idea, I manage to forget some minor thing and a dozen people promptly post objections.

In this case, I seem to have forgotten to cover the topic of how morality applies to more than one person at a time.

Stop laughing, it's not quite as dumb an oversight as it sounds. Sort of like how some people argue that macroeconomics should be constructed from microeconomics, I tend to see interpersonal morality as constructed from personal morality. (And definitely not the other way around!)

In "The Bedrock of Fairness" I offered a situation where three people discover a pie, and one of them *insists* that they want half. This is actually toned down from an older dialogue where five people discover a pie, and one of them—regardless of any argument offered—insists that they want the *whole* pie.

Let's consider the latter situation: Dennis wants the whole pie. Not only that, Dennis says that it is "fair" for him to get the whole pie, and that the "right" way *to resolve this group disagreement* is for him to get the whole pie; and he goes on saying this no matter what arguments are offered him.

This group is not going to agree, no matter what. But I would, nonetheless, say that the *right* thing to do, the *fair* thing to do, is to give Dennis one-fifth of the pie—the other four combining to hold him off by force, if necessary, if he tries to take more.

> A terminological note:
>
> In this series of posts I have been using "morality" to mean something more like "the sum of all values and valuation rules", not just "values that apply to interactions between people".
>
> The ordinary usage would have that jumping on a trampoline is not "morality", it is just some selfish fun. On the other hand, giving someone else a turn to jump on the trampoline, is more akin to "morality" in common usage; and if you say "Everyone should take turns!" that's definitely "morality".
>
> But the thing-I-want-to-talk-about includes the Fun Theory of a single person jumping on a trampoline.
>
> Think of what a disaster it would be if all fun were removed from human civilization! So I consider it quite *right* to jump on a trampoline. Even if one would not say, in ordinary conversation, "I am jumping on that trampoline because I have a moral obligation to do so." (Indeed, that sounds rather dull, and not at all fun, which is another important element of my "morality".)
>
> Alas, I do get the impression that in a standard academic discussion, one would use the term "morality" to refer to the sum-of-all-valu(ation rul)es that I am talking about. If there's a standard alternative term in moral philosophy then do *please* let me know.

> If there's a better term than "morality" for the sum of all values and valuation rules, then this would free up "morality" for interpersonal values, which is closer to the common usage.

Some years ago, I was pondering what to say to the old cynical argument: If two monkeys want the same banana, in the end one will have it, and the other will cry morality. I think the particular context was about whether the word "rights", as in the context of "individual rights", meant anything. It had just been vehemently asserted (on the Extropians mailing list, I think) that this concept was meaningless and ought to be tossed out the window.

Suppose there are two people, a Mugger and a Muggee. The Mugger wants to take the Muggee's wallet. The Muggee doesn't want to give it to him. A cynic might say: "There is nothing more to say than this; they disagree. What use is it for the Muggee to claim that he has an individual_right to keep his wallet? The Mugger will just claim that he has an individual_right to take the wallet."

Now today I might introduce the notion of a 1-place versus 2-place function, and reply to the cynic, "Either they do not mean the same thing by *individual_right*, or at least one of them is very mistaken about what their common morality implies." At most one of these people is controlled by a good approximation of what I name when I say "morality", and the other one is definitely not.

But the cynic might just say again, "So what? That's what *you* say. The Mugger could just say the opposite. What meaning is there in such claims? What difference does it make?"

So I came up with this reply: "Suppose that *I* happen along this mugging. I will decide to side with the Muggee, not the Mugger, because I have the notion that the Mugger is interfering with the Muggee's individual_right to keep his wallet, rather than the Muggee interfering with the Mugger's individual_right to take it. And if a fourth person comes along, and must decide whether to allow my intervention, or alternatively stop me from treating on the Mugger's individual_right to take the wallet, then they are likely to side with the idea that I can intervene against the Mugger, in support of the Muggee."

Now this does not work as a metaethics; it does not work to define the word *should.* If you fell backward in time, to an era when no one on Earth thought that slavery was wrong, you *should* still help slaves escape their owners. Indeed, the era when such an act was done in heroic defiance of society and the law, was not so very long ago.

But to defend the notion of individual_rights against the charge of *meaninglessness,* the notion of third-party interventions and fourth-party allowances of those interventions, seems to me to coherently cash out *what is asserted* when we assert that an individual_right exists. To assert that someone has a *right* to keep their wallet, is to assert that third parties *should* help them keep it, and that fourth parties *should* applaud those who thus help.

This perspective does make a good deal of what is said about individual_rights into nonsense. "Everyone has a right to be free from starvation!" Um, who are you talking to? Nature? Perhaps you mean, "If you're starving, and someone else has a hamburger, I'll help you take it." If so, you should say so clearly. (See also The Death of Common Sense.)

So that is a notion of individual_rights, but what does it have to do with the more general question of interpersonal morality?

The notion is that you can construct interpersonal morality out of individual morality. Just as, in this particular example, I constructed the notion of *what is* asserted by talking about an individual_right, by making it an assertion about whether third parties should decide, for themselves, to intefere; and whether fourth parties should, individually, decide to applaud the interference.

Why go to such lengths to define things in individual terms? Some people might say: "To assert the existence of a right, is to say what *society* should do."

But societies don't always agree on things. And then you, as an individual, will have to decide what's *right* for *you* to do, in that case.

"But individuals don't always agree within themselves, either," you say. "They have emotional conflicts."

Well... you *could* say that and it would sound wise. But generally speaking, neurologically intact humans will end up *doing some particular thing.* As opposed to flopping around on the floor as their limbs twitch in different directions under the temporary control of different personalities. Contrast to a government or a corporation.

A human brain is a coherently adapted system whose parts have been together optimized for a common criterion of fitness (more or less). A group is not functionally optimized as a group. (You can verify this very quickly by looking at the sex ratios in a maternity hospital.) *Individuals* may be optimized to do well out of their collective interaction—but that is quite a different selection pressure, the adaptations for which do not always produce group agreement! So if you want to look at a coherent decision system, it really is a good idea to look at one human, rather than a bureaucracy.

I myself am one person—admittedly with a long trail of human history behind me that makes me what I am, maybe more than any thoughts I ever thought myself. But still, at the end of the day, *I* am writing this blog post; it is not the negotiated output of a consortium. It is quite easy for me to imagine being faced, as an individual, with a case where the local group does not agree within itself—and in such a case I must decide, as an individual, what is *right.* In general I must decide what is right! If I go along with the group that does not absolve me of responsibility. If there are any countries that think differently, they can write their own blog posts.

This perspective, which does not exhibit undefined behavior in the event of a group disagreement, is one reason why I tend to treat interpersonal morality as a special case of individual morality, and not the other way around.

Now, with that said, interpersonal morality is a *highly distinguishable* special case of morality.

As humans, we don't just hunt in groups, we argue in groups. We've probably been arguing linguistically in adaptive political contexts for long enough—hundreds of thousands of years, maybe millions—to have adapted specifically to that selection pressure.

So it shouldn't be all that surprising if we have moral intuitions, like *fairness*, that apply specifically to the morality of groups.

One of these intuitions seems to be *universalizability.*

If Dennis just strides around saying, "I want the whole pie! Give me the whole pie! What's *fair* is for me to get the whole pie! Not you, me!" then that's not going to persuade anyone else in the tribe. Dennis has not managed to frame his desires in a form which enable them to leap from one mind to another. His desires will not take wings and become interpersonal. He is not likely to leave many offspring.

Now, the evolution of interpersonal moral intuitions, is a topic which (he said, smiling grimly) deserves its own blog post. And its own academic subfield. (Anything out there besides *The Evolutionary Origins of Morality*? It seemed to me very basic.)

But I do think it worth noting that, rather than trying to manipulate 2-person and 3-person and 7-person interactions, some of our moral instincts seem to have made the leap to N-person interactions. We just think about *general moral arguments*. As though the values that leap from mind to mind, take on a life of their own and become something that you can reason about. To the extent that everyone in your environment *does* share some values, this will work as adaptive cognition. This creates moral intuitions that are not just *interpersonal* but *transpersonal*.

Transpersonal moral intuitions are not necessarily false-to-fact, so long as you don't expect your arguments cast in "universal" terms to sway a rock. There really is such a thing as the psychological unity of humankind. Read a morality tale from an entirely different culture; I bet you can figure out what it's *trying* to argue *for,* even if you don't agree with it.

The problem arises when you try to apply the universalizability instinct to say, "If this argument could not persuade an UnFriendly AI that tries to maximize the number of paperclips in the universe, then it must not be a good argument."

There are No Universally Compelling Arguments, so if you try to apply the universalizability instinct universally, you end up with no morality. Not even universalizability; the paperclip maximizer has no intuition of universalizability. It just chooses that action which leads to a future containing the maximum number of paperclips.

There are some things you just can't have a moral conversation with. There is not that within them that could respond to your arguments. You should think twice and maybe three times before ever saying this about one of your fellow humans—but a paperclip maximizer is another matter. You'll just have to override your moral instinct to regard anything labeled a "mind" as a little floating ghost-in-the-machine, with a hidden core of perfect emptiness, which could surely be persuaded to reject its mistaken source code if you just came up with the right argument. If you're going to preserve universalizability as an intuition, you can try extending it to all humans; but you can't extend it to rocks or chatbots, nor even powerful optimization processes like evolutions or paperclip maximizers.

The question of how much *in-principle agreement* would exist *among human beings* about *the transpersonal portion of their values*, given perfect knowledge of the facts and perhaps a much wider search of the argument space, is not a matter on which we can get much evidence by observing the prevalence of moral agreement and disagreement in today's world. Any disagreement might be something that the truth

could destroy—dependent on a different view of how the world is, or maybe just dependent on having not yet heard the right argument. It is also possible that knowing more could dispel illusions of moral agreement, not just produce new accords.

But does that question really make much difference in day-to-day moral reasoning, if you're *not* trying to build a Friendly AI?

Part of *The Metaethics Sequence*

Next post: "Morality as Fixed Computation"

Previous post: "The Meaning of Right"

Morality as Fixed Computation

Toby Ord commented:

> Eliezer, I've just reread your article and was wondering if this is a good quick summary of your position (leaving apart how you got to it):
>
> 'I should X' means that I would attempt to X were I fully informed.

Toby's a pro, so if he didn't get it, I'd better try again. Let me try a different tack of explanation—one closer to the historical way that I arrived at my own position.

Suppose you build an AI, and—leaving aside that AI goal systems cannot be built around English statements, and all such descriptions are only dreams—you try to infuse the AI with the action-determining principle, "Do what I want."

And suppose you get the AI design close *enough*—it doesn't just end up tiling the universe with paperclips, cheesecake or tiny molecular copies of satisfied programmers—that its utility function actually assigns utilities as follows, to the world-states we would describe in English as:

```
<Programmer weakly desires 'X',   quantity 20 of X exists>:  +20
<Programmer strongly desires 'Y', quantity 20 of X exists>:  0
<Programmer weakly desires 'X',   quantity 30 of Y exists>:  0
<Programmer strongly desires 'Y', quantity 30 of Y exists>:  +60
```

You perceive, of course, that this destroys the world.

...since if the programmer initially weakly wants 'X' and X is hard to obtain, the AI will modify the programmer to strongly want 'Y', which is easy to create, and then bring about lots of Y. Y might be, say, iron atoms—those are highly stable.

Can you patch this problem? No. As a general rule, it is not possible to patch flawed Friendly AI designs.

If you try to bound the utility function, or make the AI not care about how *much* the programmer wants things, the AI still has a motive (as an *expected* utility maximizer) to make the programmer want something that can be obtained with a very high degree of certainty.

If you try to make it so that the AI can't modify the programmer, then the AI can't talk to the programmer (talking to someone modifies them).

If you try to rule out a specific class of ways the AI could modify the programmer, the AI has a motive to superintelligently seek out loopholes and ways to modify the programmer indirectly.

As a general rule, it is not possible to patch flawed FAI designs.

We, ourselves, do not imagine the future and judge, that any future in which our brains want something, and that thing exists, is a good future. If we did think this way, we would say: "Yay! Go ahead and modify us to strongly want something cheap!" But we do *not* say this, which means that this AI design is *fundamentally* flawed: it will choose things very unlike what we would choose; it will judge

desirability very differently from how we judge it. This core disharmony cannot be patched by ruling out a handful of specific failure modes.

There's also a duality between Friendly AI problems and moral philosophy problems—though you've got to structure that duality in exactly the right way. So if you prefer, the core problem is that the AI will choose in a way very unlike the structure of what is, y'know, actually *right*—never mind the way we choose. Isn't the whole point of this problem, that merely *wanting* something doesn't *make* it right?

So this is the paradoxical-seeming issue which I have analogized to the difference between:

> A calculator that, when you press '2', '+', and '3', tries to compute:
> "What is 2 + 3?"
>
> A calculator that, when you press '2', '+', and '3', tries to compute:
> "What does this calculator output when you press '2', '+', and '3'?"

The Type 1 calculator, as it were, *wants* to output 5.

The Type 2 "calculator" could return any result; and in the act of returning that result, it *becomes* the correct answer to the question that was internally asked.

We ourselves are like unto the Type 1 calculator. But the putative AI is being built as though it were to reflect the Type 2 calculator.

Now imagine that the Type 1 calculator is trying to build an AI, only the Type 1 calculator doesn't *know* its own question. The calculator continually asks the question by its very nature, it was born to ask that question, created already in motion around that question—but the calculator has no insight into its own transistors; it cannot print out the question, which is extremely complicated and has no simple approximation.

So the calculator wants to build an AI (it's a pretty smart calculator, it just doesn't have access to its own transistors) and have the AI give the right answer. Only the calculator can't print out the question. So the calculator wants to have the AI look at the calculator, where the question is written, and answer the question that the AI will discover implicit in those transistors. But this cannot be done by the cheap shortcut of a utility function that says "All X: <calculator asks 'X?', answer X>: utility 1; else: utility 0" because that actually mirrors the utility function of a Type 2 calculator, not a Type 1 calculator.

This gets us into FAI issues that I am not going into (some of which I'm still working out myself).

However, when you back out of the details of FAI design, and swap back to the perspective of moral philosophy, then *what we were just talking about* was the dual of the moral issue: "But if what's 'right' is a mere preference, then anything that anyone wants is 'right'."

Now I did argue against that particular concept in some detail, in The Meaning of Right, so I am not going to repeat all that...

But the key notion is the idea that what we name by 'right' is a *fixed* question, or perhaps a *fixed framework*. We can encounter moral arguments that modify our terminal values, and even encounter moral arguments that modify what we count as a

moral argument; nonetheless, it all grows out of a particular starting point. We do not experience ourselves as embodying the question "What will I decide to do?" which would be a Type 2 calculator; anything we decided would thereby become right. We experience ourselves as asking the embodied question: "What will save my friends, and my people, from getting hurt? How can we all have more fun? ..." where the "..." is around a thousand other things.

So 'I should X' does not mean that I would attempt to X were I fully informed.

'I should X' means that X answers the question, "What will save my people? How can we all have more fun? How can we get more control over our own lives? What's the funniest jokes we can tell? ..."

And I may not *know* what this question *is,* actually; I may not be able to print out my current guess nor my surrounding framework; but I know, as all non-moral-relativists instinctively know, that the question *surely* is not just "How can I do whatever I want?"

When these two formulations begin to seem as entirely distinct as "snow" and snow, then you shall have created distinct buckets for the quotation and the referent.

Added: This was posted automatically and the front page got screwed up somehow. I have no idea how. It is now fixed and should make sense.

Inseparably Right; or, Joy in the Merely Good

Followup to: The Meaning of Right

I fear that in my drive for full explanation, I may have obscured the punchline from my theory of metaethics. Here then is an attempted rephrase:

There is no pure ghostly essence of goodness apart from things like truth, happiness and sentient life.

What do you value? At a guess, you value the life of your friends and your family and your Significant Other and yourself, all in different ways. You would probably say that you value human life in general, and I would take your word for it, though Robin Hanson might ask how you've acted on this supposed preference. If you're reading this blog you probably attach some value to truth for the sake of truth. If you've ever learned to play a musical instrument, or paint a picture, or if you've ever solved a math problem for the fun of it, then you probably attach real value to good art. You value your freedom, the control that you possess over your own life; and if you've ever really helped someone you probably enjoyed it. You might not think of playing a video game as a great sacrifice of dutiful morality, but I for one would not wish to see the joy of complex challenge perish from the universe. You may not think of telling jokes as a matter of interpersonal morality, but I would consider the human sense of humor as part of the gift we give to tomorrow.

And you value many more things than these.

Your brain assesses these things I have said, or others, or more, depending on the specific event, and finally affixes a little internal representational label that we recognize and call "good".

There's no way you can detach the little label from what it stands for, and still make ontological or moral sense.

Why might the little 'good' label *seem* detachable? A number of reasons.

Mainly, that's just how your mind is structured—the labels it attaches internally seem like extra, floating, ontological properties.

And there's no *one* value that determines whether a complicated event is good or not—and no five values, either. No matter what rule you try to describe, there's always something left over, some counterexample. Since no single value defines goodness, this can make it seem like all of them together couldn't define goodness. But when you add them up all together, there is nothing else left.

If there's no detachable property of goodness, what does this mean?

It means that the question, "Okay, but what makes happiness or self-determination, *good?*" is either very quickly answered, or else malformed.

The concept of a "utility function" or "optimization criterion" is detachable when talking about optimization processes. Natural selection, for example, optimizes for

inclusive genetic fitness. But there are possible minds that implement any utility function, so you don't get any advice there about what you *should* do. You can't ask about utility apart from any utility function.

When you ask "But which utility function *should* I use?" the word *should* is something inseparable from the dynamic that labels a choice "should"—inseparable from the reasons like "Because I can save more lives that way."

Every time you say *should,* it includes an implicit criterion of choice; there is no should-ness that can be abstracted away from any criterion.

There is no separable right-ness that you could abstract from pulling a child off the train tracks, and attach to some other act.

Your values can change in response to arguments; you have metamorals as well as morals. So it probably does make sense to think of an idealized good, or idealized right, that you would assign if you could think of all possible arguments. Arguments may even convince you to change your criteria of what counts as a persuasive argument. Even so, when you consider the total trajectory arising out of that *entire framework,* that *moral frame of reference,* there is no separable property of justification-ness, apart from any particular criterion of justification; no final answer apart from a starting question.

I sometimes say that morality is "created already in motion".

There is no perfect argument that persuades the ideal philosopher of perfect emptiness to attach a perfectly abstract label of 'good'. The notion of the perfectly abstract label is incoherent, which is why people chase it round and round in circles. What would distinguish a perfectly empty label of 'good' from a perfectly empty label of 'bad'? How would you tell which was which?

But since every supposed criterion of goodness that we describe, turns out to be wrong, or incomplete, or changes the next time we hear a moral argument, it's easy to see why someone might think that 'goodness' was a thing apart from any criterion at all.

Humans have a cognitive architecture that easily misleads us into conceiving of goodness as something that can be detached from any criterion.

This conception turns out to be incoherent. Very sad. I too was hoping for a perfectly abstract argument; it appealed to my universalizing instinct. But...

But the question then becomes: is that little fillip of human psychology, more important than everything else? Is it more important than the happiness of your family, your friends, your mate, your extended tribe, and yourself? If your universalizing instinct is frustrated, is that worth abandoning life? If you represented rightness wrongly, do pictures stop being beautiful and maths stop being elegant? Is that one tiny mistake worth forsaking the gift we could give to tomorrow? Is it even really worth all that much in the way of existential angst?

Or will you just say "Oops" and go back to life, to truth, fun, art, freedom, challenge, humor, moral arguments, and all those other things that in their sum and in their reflective trajectory, are the entire and only meaning of the word 'right'?

Here is the strange habit of thought I mean to convey: Don't look to some surprising unusual twist of logic for your justification. Look to the living child, successfully dragged off the train tracks. There you will find your justification. What ever should be more important than that?

I could dress that up in computational metaethics and FAI theory—which indeed is whence the notion first came to me—but when I translated it all back into human-talk, that is what it turned out to say.

If we cannot take joy in things that are merely good, our lives shall be empty indeed.

Part of *The Metaethics Sequence*

Next post: "Sorting Pebbles Into Correct Heaps"

Previous post: "Morality as Fixed Computation"

Sorting Pebbles Into Correct Heaps

Once upon a time there was a strange little species—that might have been biological, or might have been synthetic, and perhaps were only a dream—whose passion was sorting pebbles into correct heaps.

They couldn't tell you *why* some heaps were correct, and some incorrect. But all of them agreed that the most important thing in the world was to create correct heaps, and scatter incorrect ones.

Why the Pebblesorting People cared so much, is lost to this history—maybe a Fisherian runaway sexual selection, started by sheer accident a million years ago? Or maybe a strange work of sentient art, created by more powerful minds and abandoned?

But it mattered so drastically to them, this sorting of pebbles, that all the Pebblesorting philosophers said in unison that pebble-heap-sorting was the very meaning of their lives: and held that the only justified reason to eat was to sort pebbles, the only justified reason to mate was to sort pebbles, the only justified reason to participate in their world economy was to efficiently sort pebbles.

The Pebblesorting People all agreed on that, but they didn't always agree on which heaps were correct or incorrect.

In the early days of Pebblesorting civilization, the heaps they made were mostly small, with counts like 23 or 29; they couldn't tell if larger heaps were correct or not. Three millennia ago, the Great Leader Biko made a heap of 91 pebbles and proclaimed it correct, and his legions of admiring followers made more heaps likewise. But over a handful of centuries, as the power of the Bikonians faded, an intuition began to accumulate among the smartest and most educated that a heap of 91 pebbles was incorrect. Until finally they came to know what they had done: and they scattered all the heaps of 91 pebbles. Not without flashes of regret, for some of those heaps were great works of art, but incorrect. They even scattered Biko's original heap, made of 91 precious gemstones each of a different type and color.

And no civilization since has seriously doubted that a heap of 91 is incorrect.

Today, in these wiser times, the size of the heaps that Pebblesorters dare attempt, has grown very much larger—which all agree would be a most great and excellent thing, if only they could ensure the heaps were really *correct.* Wars have been fought between countries that disagree on which heaps are correct: the Pebblesorters will never forget the Great War of 1957, fought between Y'ha-nthlei and Y'not'ha-nthlei, over heaps of size 1957. That war, which saw the first use of nuclear weapons on the Pebblesorting Planet, finally ended when the Y'not'ha-nthleian philosopher At'gra'len'ley exhibited a heap of 103 pebbles and a heap of 19 pebbles side-by-side. So persuasive was this argument that even Y'not'ha-nthlei reluctantly conceded that it was best to stop building heaps of 1957 pebbles, at least for the time being.

Since the Great War of 1957, countries have been reluctant to openly endorse or condemn heaps of large size, since this leads so easily to war. Indeed, some Pebblesorting philosophers—who seem to take a tangible delight in shocking others with their cynicism—have entirely denied the existence of pebble-sorting *progress;* they suggest that opinions about pebbles have simply been a random walk over time, with no coherence to them, the illusion of progress created by condemning all

dissimilar pasts as incorrect. The philosophers point to the disagreement over pebbles of large size, as proof that there is nothing that makes a heap of size 91 really *incorrect*—that it was simply fashionable to build such heaps at one point in time, and then at another point, fashionable to condemn them. "But... 13!" carries no truck with them; for to regard "13!" as a persuasive counterargument, is only another convention, they say. The Heap Relativists claim that their philosophy may help prevent future disasters like the Great War of 1957, but it is widely considered to be a philosophy of despair.

Now the question of what makes a heap correct or incorrect, has taken on new urgency; for the Pebblesorters may shortly embark on the creation of self-improving Artificial Intelligences. The Heap Relativists have warned against this project: They say that AIs, not being of the species *Pebblesorter sapiens*, may form their own culture with entirely different ideas of which heaps are correct or incorrect. "They could decide that heaps of 8 pebbles are correct," say the Heap Relativists, "and while ultimately they'd be no righter or wronger than us, still, *our* civilization says we shouldn't build such heaps. It is not in our interest to create AI, unless all the computers have bombs strapped to them, so that even if the AI thinks a heap of 8 pebbles is correct, we can force it to build heaps of 7 pebbles instead. Otherwise, KABOOM!"

But this, to most Pebblesorters, seems absurd. Surely a sufficiently powerful AI—especially the "superintelligence" some transpebblesorterists go on about—would be able to see *at a glance* which heaps were correct or incorrect! The thought of something with a brain the size of a planet, thinking that a heap of 8 pebbles was correct, is just too absurd to be worth talking about.

Indeed, it is an utterly futile project to constrain how a superintelligence sorts pebbles into heaps. Suppose that Great Leader Biko had been able, in his primitive era, to construct a self-improving AI; and he had built it as an expected utility maximizer whose utility function told it to create as many heaps as possible of size 91. Surely, when this AI improved itself far enough, and became smart enough, then it would see at a glance that this utility function was incorrect; and, having the ability to modify its own source code, it would *rewrite its utility function* to value more reasonable heap sizes, like 101 or 103.

And certainly not heaps of size 8. That would just be *stupid.* Any mind that stupid is too dumb to be a threat.

Reassured by such common sense, the Pebblesorters pour full speed ahead on their project to throw together lots of algorithms at random on big computers until some kind of intelligence emerges. The whole history of civilization has shown that richer, smarter, better educated civilizations are likely to agree about heaps that their ancestors once disputed. Sure, there are then larger heaps to argue about—but the further technology has advanced, the larger the heaps that have been agreed upon and constructed.

Indeed, intelligence itself has always correlated with making correct heaps—the nearest evolutionary cousins to the Pebblesorters, the Pebpanzees, make heaps of only size 2 or 3, and occasionally stupid heaps like 9. And other, even less intelligent creatures, like fish, make no heaps at all.

Smarter minds equal smarter heaps. Why would that trend break?

Moral Error and Moral Disagreement

Followup to: Inseparably Right, Sorting Pebbles Into Correct Heaps

Richard Chappell, a pro, writes:

> "When Bob says "Abortion is wrong", and Sally says, "No it isn't", they are disagreeing with each other.
>
> I don't see how Eliezer can accommodate this. On his account, what Bob asserted is true iff abortion is prohibited by the morality_Bob norms. How can Sally disagree? There's no disputing (we may suppose) that abortion is indeed prohibited by morality_Bob...
>
> Since there is moral disagreement, whatever Eliezer purports to be analysing here, it is not morality."

The phenomena of moral disagreement, moral error, and moral progress, on terminal values, are *the* primary drivers behind my metaethics. Think of how simple Friendly AI would be if there were no moral disagreements, moral errors, or moral progress!

Richard claims, "There's no disputing (we may suppose) that abortion is indeed prohibited by morality_Bob."

We may *not* suppose, and there *is* disputing. Bob does not have direct, unmediated, veridical access to the output of his own morality.

I tried to describe morality as a "computation". In retrospect, I don't think this is functioning as the Word of Power that I thought I was emitting.

Let us read, for "computation", "idealized abstract dynamic"—maybe that will be a more comfortable label to apply to morality.

Even so, I would have thought it obvious that computations may be the subjects of mystery and error. Maybe it's not as obvious outside computer science?

Disagreement has two prerequisites: the possibility of agreement and the possibility of error. For two people to agree on something, there must be something they are agreeing *about*, a referent held in common. And it must be possible for an "error" to take place, a conflict between "P" in the map and not-P in the territory. Where these two prerequisites are present, Sally can say to Bob: "That thing we were just both talking about—you are in error about it."

Richard's objection would seem in the first place to rule out the possibility of moral error, from which he derives the impossibility of moral agreement.

So: does my metaethics rule out moral error? Is there no disputing that abortion is indeed prohibited by morality_Bob?

This is such a strange idea that I find myself wondering what the heck Richard could be thinking. My best guess is that Richard, perhaps having not read all the posts in this sequence, is taking my notion of morality_Bob to refer to a *flat, static list of*

valuations explicitly asserted by Bob. "Abortion is wrong" would be on Bob's list, and there would be no disputing that.

But on the contrary, I conceive of morality_Bob as something that *unfolds* into Bob's morality—like the way one can describe in 6 states and 2 symbols a Turing machine that will write 4.640×10^{1439} **1**s to its tape before halting.

So morality_Bob refers to a compact folded specification, and not a flat list of outputs. But still, how could Bob be wrong about the output of his own morality?

In manifold obvious and non-obvious ways:

Bob could be empirically mistaken about the state of fetuses, perhaps believing fetuses to be aware of the outside world. (Correcting this might change Bob's instrumental values but not terminal values.)

Bob could have formed his beliefs about what constituted "personhood" in the presence of confusion about the nature of consciousness, so that if Bob were fully informed about consciousness, Bob would not have been tempted to talk about "the beginning of life" or "the human kind" in order to define personhood. (This changes Bob's expressed terminal values; afterward he will state different general rules about what sort of physical things are ends in themselves.)

So those are the obvious moral errors—instrumental errors driven by empirical mistakes; and erroneous generalizations about terminal values, driven by failure to consider moral arguments that are valid but hard to find in the search space.

Then there are less obvious sources of moral error: Bob could have a list of mind-influencing considerations that he considers morally valid, and a list of other mind-influencing considerations that Bob considers morally invalid. Maybe Bob was raised a Christian and now considers that cultural influence to be invalid. But, unknown to Bob, when he weighs up his values for and against abortion, the influence of his Christian upbringing comes in and distorts his summing of value-weights. So Bob believes that the output of his current validated moral beliefs is to prohibit abortion, but actually this is a leftover of his childhood and not the output of those beliefs at all.

(Note that Robin Hanson and I seem to disagree, in a case like this, as to exactly what degree we should take Bob's word about what his morals are.)

Or Bob could believe that the word of God determines moral truth and that God has prohibited abortion in the Bible. Then Bob is making metaethical mistakes, causing his mind to malfunction in a highly general way, and add moral generalizations to his belief pool, which he would not do if veridical knowledge of the universe destroyed his current and incoherent metaethics.

Now let us turn to the disagreement between Sally and Bob.

You could suggest that Sally is saying to Bob, "Abortion is allowed by morality_Bob", but that seems a bit oversimplified; it is not psychologically or morally realistic.

If Sally and Bob were *un*realistically sophisticated, they might describe their dispute as follows:

Bob: "Abortion is wrong."

Sally: "Do you think that this is something of which most humans ought to be persuadable?"

Bob: "Yes, I do. Do you think abortion is right?"

Sally: "Yes, I do. And I don't think that's because I'm a psychopath by common human standards. I think most humans would come to agree with me, if they knew the facts I knew, and heard the same moral arguments I've heard."

Bob: "I think, then, that we must have a moral disagreement: since we both believe ourselves to be a shared moral frame of reference on this issue, and yet our moral intuitions say different things to us."

Sally: "Well, it is not *logically necessary* that we have a genuine disagreement. We might be mistaken in believing ourselves to mean the same thing by the words *right* and *wrong*, since neither of us can introspectively report our own moral reference frames or unfold them fully."

Bob: "But if the meaning is similar up to the third decimal place, or sufficiently similar in some respects that it ought to be delivering similar answers on *this particular* issue, then, even if our moralities are not in-principle *identical,* I would not hesitate to invoke the intuitions for transpersonal morality."

Sally: "I agree. Until proven otherwise, I am inclined to talk about this question as if it is the same question unto us."

Bob: "So I say 'Abortion is wrong' without further qualification or specialization on what *wrong* means unto me."

Sally: "And I think that abortion is right. We have a disagreement, then, and at least one of us must be mistaken."

Bob: "Unless we're *actually* choosing differently *because* of in-principle unresolvable differences in our moral frame of reference, as if one of us were a paperclip maximizer. In that case, we would be mutually mistaken in our belief that when we talk about doing what is right, we mean the same thing by *right*. We would agree that we have a disagreement, but we would both be wrong."

Now, this is not exactly what most people are explicitly thinking when they engage in a moral dispute—but it is how I would cash out and naturalize their intuitions about transpersonal morality.

Richard also says, "Since there is moral disagreement..." This seems like a prime case of what I call naive philosophical realism—the belief that philosophical intuitions are direct unmediated veridical passports to philosophical truth.

It so happens that I agree that there *is* such a thing as moral disagreement. Tomorrow I will endeavor to justify, in fuller detail, how this statement can possibly make sense in a reductionistic natural universe. So I am not disputing this particular *proposition.* But I note, in passing, that Richard cannot justifiably assert the existence of moral disagreement as an *irrefutable premise* for discussion, though he could consider it as an *apparent datum.* You cannot take as irrefutable premises, things that you have not explained exactly; for then what is it that is certain to be true?

I cannot help but note the resemblance to Richard's assumption that "there's no disputing" that abortion is indeed prohibited by morality_Bob—the assumption that Bob has direct veridical unmediated access to the final unfolded output of his own morality.

Perhaps Richard means that we *could* suppose that abortion is indeed prohibited by morality_Bob, and allowed by morality_Sally, there being at least two possible minds for whom this would be true. Then the two minds might be mistaken about believing themselves to disagree. Actually they would simply be directed by different algorithms.

You cannot have a disagreement about which algorithm *should* direct your actions, without first having the same meaning of *should*—and no matter how you try to phrase this in terms of "what ought to direct your actions" or "right actions" or "correct heaps of pebbles", in the end you will be left with the empirical fact that it is possible to construct minds directed by any coherent utility function.

When a paperclip maximizer and a pencil maximizer do different things, they are not *disagreeing* about anything, they are just different optimization processes. You cannot detach should-ness from any specific criterion of should-ness and be left with a pure empty should-ness that the paperclip maximizer and pencil maximizer can be said to *disagree* about—unless you cover "disagreement" to include differences where two agents have nothing to say to each other.

But this would be an extreme position to take with respect to your fellow humans, and I recommend against doing so. Even a psychopath would still be in a common moral reference frame with you, if, fully informed, they would decide to take a pill that would make them non-psychopaths. If you told me that my ability to care about other people was neurologically damaged, and you offered me a pill to fix it, *I* would take it. Now, perhaps some psychopaths would not be persuadable in-principle to take the pill that would, by our standards, "fix" them. But I note the possibility to emphasize what an extreme statement it is to say of someone:

"We have nothing to argue about, we are only different optimization processes."

That should be reserved for paperclip maximizers, not used against humans whose arguments you don't like.

Part of *The Metaethics Sequence*

Next post: "Abstracted Idealized Dynamics"

Previous post: "Sorting Pebbles Into Correct Heaps"

Abstracted Idealized Dynamics

Followup to: Morality as Fixed Computation

I keep trying to describe morality as a "computation", but people don't stand up and say "Aha!"

Pondering the surprising inferential distances that seem to be at work here, it occurs to me that when I say "computation", some of my listeners may not hear the Word of Power that I thought I was emitting; but, rather, may think of some complicated boring unimportant thing like Microsoft Word.

Maybe I should have said that morality is an *abstracted idealized dynamic*. This might not have meant anything to start with, but at least it wouldn't sound like I was describing Microsoft Word.

How, oh how, am I to describe the awesome import of this concept, "computation"?

Perhaps I can display the inner nature of computation, in its most general form, by showing how that inner nature manifests in something that seems very unlike Microsoft Word—namely, morality.

Consider certain features we might wish to ascribe to that-which-we-call "morality", or "should" or "right" or "good":

• It seems that we sometimes think about morality in our armchairs, without further peeking at the state of the outside world, and arrive at some previously unknown conclusion.

Someone sees a slave being whipped, and it doesn't occur to them right away that slavery is wrong. But they go home and think about it, and imagine themselves in the slave's place, and finally think, "No."

Can you think of anywhere else that something like this happens?

Suppose I tell you that I am making a rectangle of pebbles. You look at the rectangle, and count 19 pebbles on one side and 103 dots pebbles on the other side. You don't know right away how many pebbles there are. But you go home to your living room, and draw the blinds, and sit in your armchair and think; and without further looking at the physical array, you come to the conclusion that the rectangle contains 1957 pebbles.

Now, I'm not going to say the word "computation". But it seems like that-which-is "morality" should have the property of *latent development of answers*—that you may not know right away, everything that you have sufficient in-principle information to know. All the ingredients are present, but it takes additional time to bake the pie.

You can specify a Turing machine of 6 states and 2 symbols that unfolds into a string of 4.6×10^{1439} **1**s after 2.5×10^{2879} steps. A machine I could describe aloud in ten seconds, runs longer and produces a larger state than the whole observed universe to date.

When you distinguish between the program *description* and the program's *executing state,* between the process specification and the final outcome, between the question and the answer, you can see why even certainty about a program description does not imply human certainty about the executing program's outcome. See also Artificial Addition on the difference between a compact specification versus a flat list of outputs.

Morality, likewise, is something that unfolds, through arguments, through discovery, through thinking; from a bounded set of intuitions and beliefs that animate our initial states, to a potentially much larger set of specific moral judgments we may have to make over the course of our lifetimes.

• When two human beings both think about the same moral question, even in a case where they both start out uncertain of the answer, it is not unknown for them to come to the same conclusion. It seems to happen more often than chance alone would allow—though the biased focus of reporting and memory is on the shouting and the arguments. And this is so, even if both humans remain in their armchairs and do not peek out the living-room blinds while thinking.

Where else does this happen? It happens when trying to guess the number of pebbles in a rectangle of sides 19 and 103. Now this does not prove by Greek analogy that morality is multiplication. If A has property X and B has property X it does not follow that A is B. But it seems that morality ought to have the property of *expected agreement about unknown latent answers*, which, please note, generally implies that *similar questions are being asked in different places.*

This is part of what is conveyed by the Word of Power, "computation": the notion of similar questions being asked in different places and having similar answers. Or as we might say in the business, the same computation can have multiple instantiations.

If we know the structure of calculator 1 and calculator 2, we can decide that they are "asking the same question" and that we ought to see the "same result" flashing on the screen of calculator 1 and calculator 2 after pressing the Enter key. We decide this in advance of seeing the actual results, which is what makes the concept of "computation" predictively useful.

And in fact, we can make this deduction even without knowing the exact circuit diagrams of calculators 1 and 2, so long as we're told that the circuit diagrams are the same.

And then when we see the result "1957" flash on the screen of calculator 1, we know that the same "1957" can be expected to flash on calculator 2, and we even expect to count up 1957 pebbles in the array of 19 by 103.

A hundred calculators, performing the same multiplication in a hundred different ways, can be expected to arrive at the same answer—and this is not a vacuous expectation adduced after seeing similar answers. We can form the expectation in *advance* of seeing the actual answer.

Now this does not show that morality is in fact a little electronic calculator. But it highlights the notion of something that *factors out* of different physical phenomena in different physical places, even phenomena as physically different as a calculator and an array of pebbles—a common answer to a common question. (Where is this factored-out thing? Is there an Ideal Multiplication Table written on a stone tablet somewhere outside the universe? But we are not concerned with that for now.)

Seeing that one calculator outputs "1957", we infer that *the answer*—the *abstracted* answer—is 1957; and from there we make our predictions of what to see on all the other calculator screens, and what to see in the array of pebbles.

So that-which-we-name-morality seems to have the further properties of *agreement about developed latent answers,* which we may as well think of in terms of *abstract answers;* and note that such agreement is unlikely in the absence of *similar questions.*

• We sometimes look back on our own past moral judgments, and say "Oops!" E.g., "Oops! Maybe in retrospect I shouldn't have killed all those guys when I was a teenager."

So by now it seems easy to extend the analogy, and say: "Well, maybe a cosmic ray hits one of the transistors in the calculator and it says '1959' instead of 1957—that's an error."

But this notion of "error", like the notion of "computation" itself, is more subtle than it appears.

Calculator Q says '1959' and calculator X says '1957'. Who says that calculator Q is wrong, and calculator X is right? Why not say that calculator X is wrong and calculator Q is right? Why not just say, "the results are different"?

"Well," you say, drawing on your store of common sense, "if it was just those two calculators, I wouldn't know for sure which was right. But here I've got nine other calculators that all say '1957', so it certainly seems *probable* that 1957 is the correct answer."

What's this business about "correct"? Why not just say "different"?

"Because if I have to predict the outcome of any other calculators that compute 19 x 103, or the number of pebbles in a 19 x 103 array, I'll predict 1957—or whatever observable outcome corresponds to the abstract number 1957."

So perhaps 19 x 103 = 1957 only most of the time. Why call the answer 1957 the *correct* one, rather than the mere fad among calculators, the majority vote?

If I've got a hundred calculators, all of them rather error-prone—say a 10% probability of error—then there is no *one* calculator I can point to and say, "This is the standard!" I might pick a calculator that would happen, on this occasion, to vote with ten other calculators rather than ninety other calculators. This is why I have to *idealize* the answer, to talk about this *ethereal* thing that is not associated with any particular physical process known to me—not even arithmetic done in my own head, which can also be "incorrect".

It is this ethereal process, this idealized question, to which we compare the results of any one particular calculator, and say that the result was "right" or "wrong".

But how can we obtain information about this perfect and un-physical answer, when all that we can ever observe, are merely physical phenomena? Even doing "mental" arithmetic just tells you about the result in your own, merely physical brain.

"Well," you say, "the pragmatic answer is that we can obtain extremely strong evidence by looking at the results of a hundred calculators, even if they are only 90% likely to be correct on any one occasion."

But wait: When do electrons or quarks or magnetic fields ever make an "error"? If no individual particle can be mistaken, how can any collection of particles be mistaken? The concept of an "error", though humans may take it for granted, is hardly something that would be mentioned in a fully reductionist view of the universe.

Really, what happens is that we have a certain model in mind of the calculator—the model that we looked over and said, "This implements 19 * 103"—and then other physical events caused the calculator to depart from this model, so that the final outcome, while physically lawful, did not correlate with that mysterious abstract thing, and the other physical calculators, in the way we had in mind. Given our mistaken beliefs about the physical process of the first calculator, we would look at its output '1959', and make mistaken predictions about the other calculators (which do still hew to the model we have in mind).

So "incorrect" cashes out, naturalistically, as "physically departed from the model that I had of it" or "physically departed from the idealized question that I had in mind". A calculator struck by a cosmic ray, is not 'wrong' in any physical sense, not an unlawful event in the universe; but the outcome is not the answer to the question you had in mind, the question that you believed empirically-falsely the calculator would correspond to.

The calculator's "incorrect" answer, one might say, is an answer to a different question than the one you had in mind—it is an empirical fact about the calculator that it implements a different computation.

• The 'right' act or the 'should' option sometimes seem to depend on the state of the physical world. For example, should you cut the red wire or the green wire to disarm the bomb?

Suppose I show you a long straight line of pebbles, and ask you, "How many pebbles would I have, if I had a rectangular array of six lines like this one?" You start to count, but only get up to 8 when I suddenly blindfold you.

Now you are not completely ignorant of the answer to this question. You know, for example, that the result will be even, and that it will be greater than 48. But you can't answer the question until you know how many pebbles were in the original line.

But mark this about the question: It wasn't a question about anything you could directly see in the world, at that instant. There was not in fact a rectangular array of pebbles, six on a side. You *could* perhaps lay out an array of such pebbles and count the results—but then there are more complicated computations that we could run on the unknown length of a line of pebbles. For example, we could treat the line length as the start of a Goodstein sequence, and ask whether the sequence halts. To physically play out this sequence would require many more pebbles than exist in the universe. Does it make sense to ask if the Goodstein sequence which starts with the length of this line of pebbles, "would halt"? Does it make sense to talk about *the answer,* in a case like this?

I'd say yes, personally.

But meditate upon the etherealness of the answer—that we talk about idealized abstract processes that never really happen; that we talk about what *would* happen if the law of the Goodstein sequence came into effect upon this line of pebbles, even though the law of the Goodstein sequence will never physically come into effect.

It is the same sort of etherealness that accompanies the notion of a proposition that 19 * 103 = 1957 which factors out of any particular physical calculator and is not identified with the result of any particular physical calculator.

Only now that etherealness has been mixed with physical things; we talk about the effect of an ethereal operation on a physical thing. We talk about what would happen if we ran the Goodstein process on *the number of pebbles in this line here,* which we have not counted—we do not know exactly how many pebbles there are. There is no tiny little XML tag upon the pebbles that says "Goodstein halts", but we still think—or at least I still think—that it makes sense to say of the pebbles that they have the property of their Goodstein sequence terminating.

So computations can be, as it were, idealized abstract *dynamics*—idealized abstract applications of idealized abstract laws, iterated over an imaginary causal-time that could go on for quite a number of steps (as Goodstein sequences often do).

So when we wonder, "*Should* I cut the red wire or the green wire?", we are not multiplying or simulating the Goodstein process, in particular. But we are wondering about something that is not physically immanent in the red wires or the green wires themselves; there is no little XML tag on the green wire, saying, "This is the wire that *should* be cut."

We may not know which wire defuses the bomb, but say, "Whichever wire does in fact defuse the bomb, that is the wire that *should* be cut."

Still, there are no little XML tags on the wires, and we may not even have any way to look inside the bomb—we may just have to guess, in real life.

So if we try to cash out this notion of a definite wire that *should* be cut, it's going to come out as...

...some rule that would tell us which wire to cut, if we knew the exact state of the physical world...

...which is to say, some kind of idealized abstract process into which we feed the state of the world as an input, and get back out, "cut the green wire" or "cut the red wire"...

...which is to say, the output of a computation that would take the world as an input.

• And finally I note that from the twin phenomena of *moral agreement* and *moral error,* we can construct the notion of *moral disagreement.*

This adds nothing to our understanding of "computation" as a Word of Power, but it's helpful in putting the pieces together.

Let's say that Bob and Sally are talking about an abstracted idealized dynamic they call "Enamuh".

Bob says "The output of Enamuh is 'Cut the blue wire'," and Sally says "The output of Enamuh is 'Cut the brown wire'."

Now there are several non-exclusive possibilities:

Either Bob or Sally could have committed an error in applying the rules of Enamuh—they could have done the equivalent of mis-multiplying known inputs.

Either Bob or Sally could be mistaken about some empirical state of affairs upon which Enamuh depends—the wiring of the bomb.

Bob and Sally could be talking about different things when they talk about Enamuh, in which case both of them are committing an error when they refer to Enamuh_Bob and Enamuh_Sally by the same name. (However, if Enamuh_Bob and Enamuh_Sally differ in the sixth decimal place in a fashion that doesn't change the output about which wire gets cut, Bob and Sally can quite legitimately gloss the difference.)

Or if Enamuh itself is defined by some other abstracted idealized dynamic, a Meta-Enamuh whose output is Enamuh, then either Bob or Sally could be mistaken about Meta-Enamuh in any of the same ways they could be mistaken about Enamuh. (But in the case of morality, we have an abstracted idealized dynamic that includes a specification of how it, itself, changes. Morality is *self*-renormalizing—it is not a guess at the product of some different and outside source.)

To sum up:

- Morality, like computation, involves *latent development of answers*;
- Morality, like computation, permits *expected agreement of unknown latent answers*;
- Morality, like computation, reasons about *abstract results apart from any particular physical implementation*;
- Morality, like computation, *unfolds from bounded initial state* into something *potentially much larger*;
- Morality, like computation, can be viewed as *an idealized dynamic that would operate on the true state of the physical world*—permitting us to speak about idealized answers of which we are physically uncertain;
- Morality, like computation, lets us to speak of such un-physical stuff as "error", by *comparing a physical outcome to an abstract outcome*—presumably in a case where there was previously reason to believe or desire that the physical process was isomorphic to the abstract process, yet this was not actually the case.

And so with all that said, I hope that the word "computation" has come to convey something other than Microsoft Word.

Part of *The Metaethics Sequence*

Next post: "'Arbitrary'"

Previous post: "Moral Error and Moral Disagreement"

"Arbitrary"

Followup to: Inseparably Right; or, Joy in the Merely Good, Sorting Pebbles Into Correct Heaps

One of the experiences of following the Way is that, from time to time, you notice a new word that you have been using without really understanding. And you say: "What does this word, 'X', really mean?"

Perhaps 'X' is 'error', for example. And those who have not yet realized the importance of this aspect of the Way, may reply: "Huh? What do you mean? Everyone knows what an 'error' is; it's when you get something wrong, when you make a mistake." And you reply, "But those are only synonyms; what can the term 'error' mean in a universe where particles only ever do what they do?"

It's not meant to be a rhetorical question; you're meant to go out and answer it. One of the primary tools for doing so is Rationalist's Taboo, when you try to speak without using the word or its synonyms—to replace the symbol with the substance.

So I ask you therefore, what is this word "arbitrary"? Is a rock arbitrary? A leaf? A human?

How about sorting pebbles into prime-numbered heaps? How about maximizing inclusive genetic fitness? How about dragging a child off the train tracks?

How can I tell exactly which things are arbitrary, and which not, in this universe where particles only ever do what they do? Can you tell me exactly what property is being discriminated, without using the word "arbitrary" or any direct synonyms? Can you open up the box of "arbitrary", this label that your mind assigns to some things and not others, and tell me what kind of algorithm is at work here?

Having pondered this issue myself, I offer to you the following proposal:

> A piece of cognitive content feels "arbitrary" if it is the kind of cognitive content that we expect to come with attached justifications, and those justifications are not present in our mind.

You'll note that I've performed the standard operation for guaranteeing that a potentially confusing question has a real answer: I substituted the question, "How does my brain label things 'arbitrary'?" for "What is this mysterious property of arbitrariness?" This is not necessarily a sleight-of-hand, since to explain something is not the same as explaining it away.

In this case, for nearly all everyday purposes, I would make free to proceed from "arbitrary" to arbitrary. If someone says to me, "I believe that the probability of finding life on Mars is $6.203 * 10^{-23}$ to four significant digits," I would make free to respond, "That sounds like a rather arbitrary number," not "My brain has attached the subjective arbitrariness-label to its representation of the number in your belief."

So as it turned out in this case, having answered the question "What is 'arbitrary'?" turns out not to affect the way I use the word 'arbitrary'; I am just more aware of what the arbitrariness-sensation indicates. I am aware that when I say, "$6.203 * 10^{-23}$ sounds like an arbitrary number", I am indicating that I would expect some

justification for assigning that particular number, and I haven't heard it. This also explains why the precision is important—why I would question that particular number, but not someone saying "Less than 1%". In the latter case, I have some idea what might justify such a statement; but giving a very precise figure implies that you have some kind of information I don't know about, either that or you're being silly.

"Ah," you say, "but what do you mean by 'justification'? Haven't you failed to make any progress, and just passed the recursive buck to another black box?"

Actually, no; I told you that "arbitrariness" was a sensation produced by the *absence* of an *expected* X. Even if I don't tell you anything more about that X, you've learned something about the cognitive algorithm—opened up the original black box, and taken out two gears and a *smaller* black box.

But yes, it makes sense to continue onward to discuss this mysterious notion of "justification".

Suppose I told you that "justification" is what tells you whether a belief is reasonable. Would this tell you anything? No, because there are no extra gears that have been factored out, just a direct invocation of "reasonable"-ness.

Okay, then suppose instead I tell you, "Your mind labels X as a justification for Y, whenever adding 'X' to the pool of cognitive content would result in 'Y' being added to the pool, or increasing the intensity associated with 'Y'." How about that?

"Enough of this buck-passing tomfoolery!" you may be tempted to cry. But wait; this really does factor out another couple of gears. We have the idea that different propositions, to the extent they are held, can create each other in the mind, or increase the felt level of intensity—credence for beliefs, desire for acts or goals. You may have already known this, more or less, but stating it aloud is still progress.

This may not provide much satisfaction to someone inquiring into morals. But then someone inquiring into morals may well do better to just think moral thoughts, rather than thinking about metaethics or reductionism.

On the other hand, if you were building a Friendly AI, and trying to explain to that FAI what a human being means by the term "justification", then the statement I just issued might help the FAI narrow it down. With some additional guidance, the FAI might be able to figure out where to look, in an empirical model of a human, for representations of the sort of *specific* moral content that a human inquirer-into-morals would be interested in—what *specifically* counts or doesn't count as a justification, in the eyes of that human. And this being the case, you might not have to explain the specifics exactly correctly at system boot time; the FAI knows how to find out the rest on its own. My inquiries into metaethics are not directed toward the same purposes as those of standard philosophy.

Now of course you may reply, "Then the FAI finds out what the human *thinks* is a "justification". But is that formulation of 'justification', really *justified?*" But by this time, I hope, you can predict my answer to that sort of question, whether or not you agree. I answer that we have just witnessed a strange loop through the meta-level, in which you use justification-as-justification to evaluate the quoted form of justification-as-cognitive-algorithm, which algorithm may, perhaps, happen to be your own, &c. And that the feeling of "justification" cannot be coherently detached from the specific algorithm we use to decide justification in particular cases; that there is no pure

empty essence of justification that will persuade any optimization process regardless of its algorithm, &c.

And the upshot is that differently structured minds may well label different propositions with their *analogues* of the internal label "arbitrary"—though only one of these labels is what *you mean* when you say "arbitrary", so you and these other agents do not really have a disagreement.

Part of *The Metaethics Sequence*

Next post: "Is Fairness Arbitrary?"

Previous post: "Abstracted Idealized Dynamics"

Is Fairness Arbitrary?

Followup to: The Bedrock of Fairness

In "The Bedrock of Fairness", Xannon, Yancy, and Zaire argue over how to split up a pie that they found in the woods. Yancy thinks that 1/3 each is fair; Zaire demands half; and Xannon tries to compromise.

Dividing a pie fairly isn't as trivial a problem as it may sound. What if people have different preferences for crust, filling, and topping? Should they each start with a third, and trade voluntarily? But then they have conflicts of interest over how to divide the surplus utility generated by trading...

But I would say that "half for Zaire" surely isn't fair.

I confess that I originally wrote Zaire as a foil—this is clearer in an earlier version of the dialog, where Zaire, named Dennis, demands the whole pie—and was surprised to find some of my readers taking Zaire's claim seriously, perhaps because I had Zaire say "I'm hungry."

Well, okay; I believe that when I write a dialogue, the reader has a right to their own interpretation. But I did intend that dialogue to illustrate a particular point:

You can argue about how to divide up the pie, or even argue how to argue about dividing up the pie, you can argue over what is fair... but there finally comes a point when you hit bedrock. If Dennis says, "No, the *fair* way to argue is that *I* get to dictate everything, and I now hereby dictate that I get the whole pie," there's nothing left to say but "Sorry, that's just not what *fairness* is—you can try to take the pie and I can try to stop you, but you can't convince that *that* is fair."

A "fair division" is not the same as "a division that compels everyone to admit that the division is fair". Dennis can always just refuse to agree, after all.

But more to the point, when you encounter a pie in the forest, in the company of friends, and you try to be *fair,* there's a certain particular thing you're trying to do—the term "fair" is not perfectly empty, it cannot attach to just anything. Metaphorically speaking, "fair" is not a hypothesis equally compatible with any outcome.

Fairness expresses notions of concern for the other agents who also want the pie; a goal to take their goals into account. It's a separate question whether that concern is pure altruism, or not wanting to make them angry enough to fight. Fairness expresses notions of symmetry, equal treatment—which might be a terminal value unto you, or just an attempt to find a convenient meeting-point to avoid an outright battle.

Is it fair to take into account what *other* people think is "fair", and not just what *you* think is "fair"?

The obvious reason to care what other people think is "fair", is if they're being moved by *similar considerations*, yet arriving at different conclusions. If you think that the Other's word "fair" means what you think of as *fair,* and you think the Other is being honest about what they think, then you ought to pay attention just by way of fulfilling your *own* desire to be fair. It is like paying attention to an honest person who means the same thing you do by "multiplication", who says that 19 * 103 might not be 1947.

The attention you pay to that suggestion, is not a favor to the other person; it is something you do if *you* want to get the multiplication right—*they're* doing *you* a favor by correcting you.

Politics is more subject to bias than multiplication. And you might think that the Other's reasoning is corrupted by self-interest, while yours is as pure as Antarctic snow. But to the extent that you credit the Other's self-honesty, or doubt your own, you would do well to hear what the Other has to say—if *you* wish to be fair.

The second notion of why we might pay attention to what someone else thinks is "fair", is more complicated: it is the notion of *applying fairness to its own quotation,* that is, fairly debating what is "fair". In complicated politics you may have to negotiate a negotiating procedure. Surely it wouldn't be fair if Dennis just got to say, "The fair resolution procedure is that I get to decide what's fair." So why should *you* get to just decide what's fair, then?

Here the attention you pay to the other person's beliefs about "fairness", is a favor that *you* do to *them,* a concession that you expect to be met with a return concession.

But when you set out to fairly discuss what is "fair" (note the strange loop through the meta-level), that doesn't put *everything* up for grabs. A zeroth-order fair division of a pie doesn't involve giving away the *whole* pie to Dennis—just giving identical portions to all. Even though Dennis wants the whole thing, and asks for the whole thing, the zeroth-order fair division only gives Dennis a symmetrical portion to everyone else's. Similarly, a first-order fair attempt to resolve a dispute about what is "fair", doesn't involve conceding everything to the Other's viewpoint without reciprocation. That wouldn't be fair. Why give everything away to the Other, if you receive nothing in return? Why give Dennis the whole first-order pie?

On some level, then, there has to be a possible demand which would be too great—a demand exceeding what may be *fairly* requested of you. This is part of the content of fairness; it is part of what you are setting out to do, when you set out to be fair. Admittedly, one should not be too trigger-happy about saying "That's too much!" We human beings tend to overestimate the concessions we have made, and underestimate the concessions that others have made to us; we tend to underadjust for the Other's point of view... even so, if *nothing* is "too much", then you're not engaging in *fairness*.

Fairness might call on you to hear out what the Other has to say; fairness may call on you to exert an effort to really truly consider the Other's point of view—but there is a limit to this, as there is a limit to all fair concessions. If all Dennis can say is "I want the whole pie!" over and over, there's a limit to how long fairness requires you to ponder this argument.

You reach the bedrock of fairness at the point where, no matter who questions whether the division is fair, no matter who refuses to be persuaded, no matter who offers further objections, and regardless of your awareness that you yourself may be biased... Dennis still isn't getting the whole pie. If there are others present who are also trying to be fair, and Dennis is not already dictator, they will probably back you rather than Dennis—this is one sign that you can trust the line you've drawn, that it really is time to say "Enough!"

If you and the others present get together and give Dennis 1/Nth of the pie—or even if *you* happen to have the upper hand, and you unilaterally give Dennis and yourself and

all others each 1/Nth—then you are not being unfair on *any* level; there *is no* meta-level of fairness where Dennis gets the whole pie.

Now I'm sure there are some in the audience who will say, "You and perhaps some others, are *merely* doing things your way, rather than Dennis's." On the contrary: We are merely being fair. It so happens that this fairness is our way, as all acts must be *someone's* way to happen in the real universe. But what we are merely doing, happens to be, being *fair.* And there is no level on which it is unfair, because there is no level on which fairness requires unlimited unreciprocated surrender.

I don't believe in unchangeable bedrock—I believe in self-modifying bedrock. But I do believe in bedrock, in the sense that everything has to start somewhere. It can be turtles all the way up, but not turtles all the way down.

You cannot define fairness *entirely* in terms of "That which everyone agrees is 'fair'." This isn't just nonterminating. It isn't just ill-defined if Dennis doesn't believe that 'fair' is "that which everyone agrees is 'fair'". It's actually *entirely empty*, like the English sentence "This sentence is true." Is that sentence true? Is it false? It is neither; it doesn't mean anything because it is entirely wrapped up in itself, with no tentacle of relation to reality. If you're going to argue what is fair, there has to *be something* you're arguing *about,* some structure that is baked into the question.

Which is to say that you can't turn "fairness" into an ideal label of pure emptiness, defined *only* by the mysterious compulsion of every possible agent to admit "This is what is 'fair'." Forget the case against universally compelling arguments—just consider the definition itself: *It has absolutely no content, no external references;* it is not just *under*specified, but *entirely un*specified.

But as soon as you introduce any content into the label "fairness" that *isn't* phrased purely in terms of all possible minds applying the label, then you have a foundation on which to stand. It may be self-modifying bedrock, rather than immovable bedrock. But it is still a place to start. A place from which to say: "Regardless of what Dennis says, giving him the whole pie *isn't fair,* because *fairness* is not defined entirely and only in terms of Dennis's agreement."

And you aren't being "arbitrary", either—though the intuitive meaning of that word has never seemed entirely well-specified to me; is a tree arbitrary, or a leaf? But it sounds like the accusation is of pulling some answer out of thin air—which you're *not* doing; you're giving the *fair* answer, not an answer pulled out of thin air. What about when you jump up a meta-level, and look at Dennis's wanting to do it one way, and your wanting a different resolution? Then it's still not arbitrary, because you aren't being *unfair* on that meta-level, either. The answer you pull out is not *merely* an arbitrary answer you invented, but a *fair* answer. You aren't *merely* doing it your way; the way that you are doing it, is the fair way.

You can ask "But why *should* you be fair?"—and that's a separate question, which we'll go into tomorrow. But giving Dennis 1/Nth, we can at least say, is not *merely and only arbitrary* from the perspective of fair-vs.-unfair. Even if Dennis keeps saying "It isn't fair!" and even if Dennis also disputes the 1st-order, 2nd-order, Nth-order meta-fairnesses. Giving N people each 1/Nth is nonetheless a *fair* sort of thing to do, and whether or not we *should* be fair is then a separate question.

Next post: "The Bedrock of Morality: Arbitrary?"

Previous post: "'Arbitrary'"

The Bedrock of Morality: Arbitrary?

Followup to: Is Fairness Arbitrary?, Joy in the Merely Good, Sorting Pebbles Into Correct Heaps

Yesterday, I presented the idea that when only five people are present, having just stumbled across a pie in the woods (a naturally growing pie, that just popped out of the ground) then it is fair to give Dennis only 1/5th of this pie, even if Dennis persistently claims that it is fair for him to get the whole thing. Furthermore, it is meta-fair to follow such a symmetrical division procedure, even if Dennis insists that *he* ought to dictate the division procedure.

Fair, meta-fair, or meta-meta-fair, there is no level of fairness where you're obliged to concede everything to Dennis, without reciprocation or compensation, just because he demands it.

Which goes to say that fairness has a meaning beyond which "that which everyone can be convinced is 'fair'". This is an empty proposition, isomorphic to "Xyblz is that which everyone can be convinced is 'xyblz'". There must be some *specific* thing of which people are being convinced; and once you identify that thing, it has a meaning beyond agreements and convincing.

You're not introducing something *arbitrary*, something un-fair, in refusing to concede everything to Dennis. You are being fair, and meta-fair and meta-meta-fair. As far up as you go, there's no level that calls for unconditional surrender. The stars do not judge between you and Dennis—but it *is* baked into the very question that is asked, when you ask, "What is fair?" as opposed to "What is xyblz?"

Ah, but why *should* you be fair, rather than xyblz? Let us concede that Dennis cannot validly persuade us, on any level, that it is *fair* for him to dictate terms and give himself the whole pie; but perhaps he could argue whether we *should* be fair?

The hidden agenda of the whole discussion of fairness, of course, is that good-ness and right-ness and should-ness, ground out similarly to fairness.

Natural selection optimizes for inclusive genetic fitness. This is not a disagreement with humans about what is good. It is simply that natural selection does not *do* what is good: it optimizes for inclusive genetic fitness.

Well, since some optimization processes optimize for inclusive genetic fitness, instead of what is good, which *should* we do, ourselves?

I know my answer to this question. It has something to do with natural selection being a terribly wasteful and stupid and inefficient process. It has something to do with elephants starving to death in their old age when they wear out their last set of teeth. It has something to do with natural selection never choosing a single act of mercy, of grace, even when it would cost its purpose nothing: not auto-anesthetizing a wounded and dying gazelle, when its pain no longer serves even the adaptive purpose that first created pain. Evolution had to happen sometime in the history of the universe, because that's the only way that intelligence could *first* come into being, without brains to make brains; but now that era is over, and good riddance.

But most of all—why on Earth *would* any human being think that one *ought* to optimize inclusive genetic fitness, rather than what is good? What is even the appeal of this, morally or otherwise? *At all?* I know people who claim to think like this, and I wonder what wrong turn they made in their cognitive history, and I wonder how to get them to snap out of it.

When we take a step back from fairness, and ask if we *should* be fair, the answer may not always be yes. Maybe sometimes we should be merciful. But if you ask if it is *meta-fair* to be fair, the answer will generally be yes. Even if someone else wants you to be unfair in their favor, or claims to disagree about what is "fair", it will still generally be meta-fair to be fair, even if you can't make the Other agree. By the same token, if you ask if we meta-should do what we should, rather than something else, the answer is yes. Even if some other agent or optimization process does not do what is right, that doesn't change what is meta-right.

And this is not "arbitrary" in the sense of rolling dice, not "arbitrary" in the sense that justification is expected and then not found. The accusations that I level against evolution are not *merely* pulled from a hat; they are expressions of morality as I understand it. They are merely moral, and there is nothing mere about that.

In "Arbitrary" I finished by saying:

> The upshot is that differently structured minds may well label different propositions with their *analogues* of the internal label "arbitrary"—though only one of these labels is what *you mean* when you say "arbitrary", so you and these other agents do not really have a disagreement.

This was to help shake people loose of the idea that if any two possible minds can say or do different things, then it must all be arbitrary. Different minds may have different ideas of what's "arbitrary", so clearly this whole business of "arbitrariness" is arbitrary, and we should ignore it. After all, Sinned (the anti-Dennis) just always says "Morality isn't arbitrary!" no matter how you try to persuade her otherwise, so clearly you're just being arbitrary in saying that morality is arbitrary.

From the perspective of a human, saying that one should sort pebbles into prime-numbered heaps is arbitrary—it's the sort of act you'd expect to come with a justification attached, but there isn't any justification.

From the perspective of a Pebblesorter, saying that one p-should scatter a heap of 38 pebbles into two heaps of 19 pebbles is not p-arbitrary at all—it's the most p-important thing in the world, and fully p-justified by the intuitively obvious fact that a heap of 19 pebbles is p-correct and a heap of 38 pebbles is not.

So which perspective should we adopt? I answer that I see no reason at all why I should start sorting pebble-heaps. It strikes me as a completely pointless activity. Better to engage in art, or music, or science, or heck, better to connive political plots of terrifying dark elegance, than to sort pebbles into prime-numbered heaps. A galaxy transformed into pebbles and sorted into prime-numbered heaps would be just plain boring.

The Pebblesorters, of course, would only reason that music is p-pointless because it doesn't help you sort pebbles into heaps; the human activity of humor is not only p-pointless but just plain p-bizarre and p-incomprehensible; and most of all, the human vision of a galaxy in which agents are running around experiencing positive reinforcement *but not sorting any pebbles,* is a vision of an utterly p-arbitrary galaxy

devoid of p-purpose. The Pebblesorters would gladly sacrifice their lives to create a P-Friendly AI that sorted the galaxy on their behalf; it would be the most p-profound statement they could make about the p-meaning of their lives.

So which of these two perspectives do I choose? The human one, of course; not because it is the human one, but because it is *right.* I do not know perfectly what is right, but neither can I plead entire ignorance.

And the Pebblesorters, *who simply are not built to do what is right,* choose the Pebblesorting perspective: not merely because it is theirs, or because they think they can get away with being p-arbitrary, but because that is what is p-right.

And in fact, both we and the Pebblesorters can *agree* on all these points. We can agree that sorting pebbles into prime-numbered heaps is arbitrary and unjustified, but not p-arbitrary or p-unjustified; that it is the sort of thing an agent p-should do, but not the sort of thing an agent should do.

I fully expect that even if there is other life in the universe only a few trillions of lightyears away (I don't think it's local, or we would have seen it by now), that we humans are the only creatures for a long long way indeed who are built to do what is *right*. That may be a moral miracle, but it is not a causal miracle.

There may be some other evolved races, a sizable fraction perhaps, maybe even a majority, who do some right things. Our executing adaptation of compassion is not so far removed from the game theory that gave it birth; it might be a common adaptation. But laughter, I suspect, may be rarer by far than mercy. What would a galactic civilization be like, if it had sympathy, but never a moment of humor? A little more boring, perhaps, by our standards.

This humanity that we find ourselves in, is a great gift. It may not be a great p-gift, but who cares about p-gifts?

So I really must deny the charges of moral relativism: I don't think that human morality is arbitrary at all, and I would expect any logically omniscient reasoner to agree with me on that. We are better than the Pebblesorters, because we care about sentient lives, and the Pebblesorters don't. Just as the Pebblesorters are p-better than us, because they care about pebble heaps, and we don't. Human morality is p-arbitrary, but who cares? P-arbitrariness is arbitrary.

You've just got to avoid thinking that the words "better" and "p-better", or "moral" and "p-moral", are *talking about the same thing*—because then you might think that the Pebblesorters were coming to different conclusions than us about *the same thing*—and then you might be tempted to think that our own morals were arbitrary. Which, of course, they're not.

Yes, I really truly do believe that humanity is better than the Pebblesorters! I am not being sarcastic, I really do believe that. I am not playing games by redefining "good" or "arbitrary", I think I mean the same thing by those terms as everyone else. When you understand that I am genuinely sincere about that, you will understand my metaethics. I really *don't* consider myself a moral relativist—not even in the slightest!

Next post: "You Provably Can't Trust Yourself"

Previous post: "Is Fairness Arbitrary?"

You Provably Can't Trust Yourself

Followup to: Where Recursive Justification Hits Bottom, Löb's Theorem

Peano Arithmetic *seems* pretty trustworthy. We've never found a case where Peano Arithmetic proves a theorem T, and yet T is false in the natural numbers. That is, we know of no case where []T ("T is provable in PA") and yet ~T ("not T").

We also know of no case where first order logic is invalid: We know of no case where first-order logic produces *false conclusions* from *true premises*. (Whenever first-order statements H are true of a model, and we can syntactically deduce C from H, checking C against the model shows that C is also true.)

Combining these two observations, it seems like we should be able to get away with adding a rule to Peano Arithmetic that says:

All T: ([]T -> T)

But Löb's Theorem seems to show that as soon as we do that, everything becomes provable. What went wrong? How can we do *worse* by adding a true premise to a trustworthy theory? Is the premise not true—does PA prove some theorems that are false? Is first-order logic not valid—does it sometimes prove false conclusions from true premises?

Actually, there's nothing wrong with reasoning from the axioms of Peano Arithmetic plus the axiom schema "Anything provable in Peano Arithmetic is true." But the result is a *different* system from PA, which we might call PA+1. PA+1 does not reason from identical premises to PA; something new has been added. So we can evade Löb's Theorem because PA+1 is not trusting *itself*—it is only trusting PA.

If you are not previously familiar with mathematical logic, you might be tempted to say, "Bah! Of course PA+1 is trusting itself! PA+1 just isn't willing to admit it! Peano Arithmetic *already* believes anything provable in Peano Arithmetic—it will *already* output anything provable in Peano Arithmetic as a theorem, *by definition!* How does moving to PA+1 change anything, then? PA+1 is just the same system as PA, and so by trusting PA, PA+1 is really trusting itself. Maybe that dances around some obscure mathematical problem with direct self-reference, but it doesn't evade the charge of self-trust."

But PA+1 and PA really are different systems; in PA+1 it is possible to prove true statements about the natural numbers that are not provable in PA. If you're familiar with mathematical logic, you know this is because some nonstandard models of PA are ruled out in PA+1. Otherwise you'll have to take my word that Peano Arithmetic doesn't fully describe the natural numbers, and neither does PA+1, but PA+1 characterizes the natural numbers slightly better than PA.

The deeper point is the *enormous* gap, the *tremendous* difference, between having a system just like PA except that it trusts PA, and a system just like PA except that it trusts *itself.*

If you have a system that trusts PA, that's no problem; we're pretty sure PA is trustworthy, so the system is reasoning from true premises. But if you have a system that looks like PA—having the standard axioms of PA—but also trusts *itself,* then it is

trusting a self-trusting system, something for which there is no precedent. In the case of PA+1, PA+1 is trusting PA which we're pretty sure is correct. In the case of Self-PA it is trusting Self-PA, which we've never seen before—it's never been tested, despite its *misleading* surface similarity to PA. And indeed, Self-PA collapses via Löb's Theorem and proves everything—so I guess it *shouldn't* have trusted itself after all! All this isn't magic; I've got a nice Cartoon Guide to how it happens, so there's no good excuse for not understanding what goes on here.

I have spoken of the Type 1 calculator that asks "What is 2 + 3?" when the buttons "2", "+", and "3" are pressed; versus the Type 2 calculator that asks "What do I calculate when someone presses '2 + 3'?" The first calculator answers 5; the second calculator can truthfully answer anything, even 54.

But this doesn't mean that all calculators that reason about calculators are flawed. If I build a third calculator that asks "What does the first calculator answer when I press '2 + 3'?", perhaps by calculating out the individual transistors, it too will answer 5. Perhaps this new, reflective calculator will even be able to answer some questions faster, by virtue of proving that some faster calculation is isomorphic to the first calculator.

PA is the equivalent of the first calculator; PA+1 is the equivalent of the third calculator; but Self-PA is like unto the second calculator.

As soon as you start trusting yourself, you become unworthy of trust. You'll start believing *any* damn thing that you think, just because *you* thought it. This wisdom of the human condition is pleasingly analogous to a precise truth of mathematics.

Hence the saying: "Don't believe everything you think."

And the math also suggests, by analogy, how to do better: Don't trust thoughts *because you think them,* but because they *obey specific trustworthy rules.*

PA only starts believing something—metaphorically speaking—when it sees a specific proof, laid out in black and white. If you say to PA—even if you prove to PA—that PA will prove something, PA still won't believe you until it sees the *actual proof.* Now, this might seem to invite inefficiency, and PA+1 will believe you—if you prove that *PA* will prove something, because PA+1 trusts the specific, fixed framework of Peano Arithmetic; not *itself.*

As far as any human knows, PA does happen to be sound; which means that what PA proves is provable in PA, PA *will* eventually prove and *will* eventually believe. Likewise, anything PA+1 can prove that it proves, it will eventually prove and believe. It seems so tempting to just make PA trust *itself*—but then it becomes Self-PA and implodes. Isn't that odd? PA believes everything it proves, but it doesn't believe "Everything I prove is true." PA trusts a fixed framework for how to prove things, and that framework doesn't happen to talk *about* trust in the framework.

You *can* have a system that trusts the PA framework *explicitly,* as well as implicitly: that is PA+1. But the new framework that PA+1 *uses,* makes no *mention* of itself; and the specific proofs that PA+1 demands, make no mention of trusting PA+1, only PA. You might say that PA implicitly trusts PA, PA+1 explicitly trusts PA, and Self-PA trusts itself.

For everything that you believe, you should always find yourself able to say, "I believe because of [specific argument in framework F]", not "I believe because I believe".

Of course, this gets us into the +1 question of why you ought to trust or use framework F. Human beings, not being formal systems, are too reflective to get away with being *unable* to think about the problem. Got a superultimate framework U? Why trust U?

And worse: as far as I can tell, *using* induction is what leads me to *explicitly* say that induction seems to often work, and my *use* of Occam's Razor is implicated in my explicit *endorsement* of Occam's Razor. Despite my best efforts, I have been unable to prove that this is inconsistent, and I suspect it may be valid.

But it does seem that the distinction between *using* a framework and *mentioning* it, or between *explicitly* trusting a fixed framework F and trusting *yourself*, is at least *important* to unraveling foundational tangles—even if Löb turns out not to apply directly.

Which gets me to the reason why I'm saying all this in the middle of a sequence about morality.

I've been pondering the unexpectedly large inferential distances at work here—I thought I'd gotten all the prerequisites out of the way for explaining metaethics, but no. I'm no longer sure I'm even close. I tried to say that morality was a "computation", and that failed; I tried to explain that "computation" meant "abstracted idealized dynamic", but that didn't work either. No matter how many different ways I tried to explain it, I couldn't get across the distinction my metaethics drew between "do the right thing", "do the human thing", and "do my own thing". And it occurs to me that my own background, coming into this, may have relied on having already drawn the distinction between PA, PA+1 and Self-PA.

Coming to terms with metaethics, I am beginning to think, is all about *distinguishing between levels*. I first learned to do this *rigorously* back when I was getting to grips with mathematical logic, and discovering that you could prove complete absurdities, if you lost track even once of the distinction between "believe particular PA proofs", "believe PA is sound", and "believe you yourself are sound". If you believe any particular PA proof, that might sound pretty much the same as believing PA is sound in general; and if you use PA and only PA, then trusting PA (that is, being moved by arguments that follow it) sounds pretty much the same as believing that you yourself are sound. But after a bit of practice with the actual math—I did have to practice the actual math, not just read about it—my mind formed permanent distinct buckets and built walls around them to prevent the contents from slopping over.

Playing around with PA and its various conjugations, gave me the notion of what it meant to trust *arguments within a framework that defined justification*. It gave me practice keeping track of specific frameworks, and holding them distinct in my mind.

Perhaps that's why I expected to communicate more sense than I actually succeeded in doing, when I tried to describe *right* as a framework of justification that involved being moved by particular, specific terminal values and moral arguments; analogous to an entity who is moved by encountering a specific proof from the allowed axioms of Peano Arithmetic. As opposed to a general license to do whatever you prefer, or a morally relativistic term like "utility function" that can eat the values of any given species, or a neurological framework contingent on particular facts about the human brain. You can make good use of such concepts, but I do not identify them with the substance of what is *right.*

Gödelian arguments are inescapable; you can always isolate the framework-of-trusted-arguments if a mathematical system makes sense at all. Maybe the adding-up-to-normality-ness of my system will become clearer, after it becomes clear that you can always isolate the framework-of-trusted-arguments of a human having a moral argument.

Part of *The Metaethics Sequence*

Next post: "No License To Be Human"

Previous post: "The Bedrock of Morality: Arbitrary?"

No License To Be Human

Followup to: You Provably Can't Trust Yourself

Yesterday I discussed the difference between:

- A system that believes—is moved by—any *specific* chain of deductions from the axioms of Peano Arithmetic. (PA, Type 1 calculator)
- A system that believes PA, plus *explicitly* asserts the *general* proposition that PA is sound. (PA+1, meta-1-calculator that calculates the output of Type 1 calculator)
- A system that believes PA, plus explicitly asserts *its own* soundness. (Self-PA, Type 2 calculator)

These systems are formally distinct. PA+1 can prove things that PA cannot. Self-PA is inconsistent, and can prove anything via Löb's Theorem.

With these distinctions in mind, I hope my intent will be clearer, when I say that although I am human and have a human-ish moral framework, I do not think that the fact of *acting in a human-ish* way licenses anything.

I am a self-renormalizing moral system, but I do not think there is any general license to be a self-renormalizing moral system.

And while we're on the subject, I am an epistemologically incoherent creature, trying to modify his ways of thinking in accordance with his current conclusions; but I do not think that reflective coherence implies correctness.

Let me take these issues in reverse order, starting with the general unlicensure of epistemological reflective coherence.

If five different people go out and investigate a city, and draw five different street maps, we should expect the maps to be (mostly roughly) consistent with each other. Accurate maps are necessarily consistent among each other and among themselves, there being only one reality. But if I sit in my living room with my blinds closed, I can draw up one street map from my imagination and then make four copies: these five maps will be consistent among themselves, but not accurate. Accuracy implies consistency but not the other way around.

In Where Recursive Justification Hits Bottom, I talked about whether "I believe that induction will work on the next occasion, because it's usually worked before" is legitimate reasoning, or "I trust Occam's Razor because the simplest explanation for why Occam's Razor often works is that we live in a highly ordered universe". Though we actually *formalized* the idea of scientific induction, starting from an inductive *instinct;* we modified our intuitive understanding of Occam's Razor (Maxwell's Equations are in fact simpler than Thor, as an explanation for lightning) based on the simple idea that "the universe runs on equations, not heroic mythology". So we did not automatically and unthinkingly confirm our assumptions, but rather, *used* our intuitions to *correct* them—seeking reflective coherence.

But I also remarked:

"And what about trusting reflective coherence in general? Wouldn't most possible minds, randomly generated and allowed to settle into a state of reflective coherence, be incorrect? Ah, but *we* evolved by natural selection; we were not generated randomly."

So you are not, *in general,* safe if you reflect on yourself and achieve internal coherence. The Anti-Inductors who compute that the probability of the coin coming up heads on the next occasion, decreases each time they see the coin come up heads, may defend their anti-induction by saying: "But it's never worked before!"

The only reason why our human reflection works, is that we are good enough to make ourselves better—that we had a core instinct of induction, a core instinct of simplicity, that wasn't sophisticated or exactly right, but worked well enough.

A mind that was completely wrong to start with, would have no seed of truth from which to heal itself. (It can't forget everything and become a mind of pure emptiness that would mysteriously do induction correctly.)

So it's not that reflective coherence is licensed *in general*, but that it's a good idea *if* you start out with a core of truth or correctness or good priors. Ah, but who is deciding whether I possess good priors? I am! By reflecting on them! The inescapability of this strange loop is *why* a broken mind can't heal itself—because there is no jumping outside of *all* systems.

I can only plead that, in evolving to perform induction rather than anti-induction, in evolving a flawed but not absolutely wrong instinct for simplicity, I have been blessed with an epistemic gift.

I can only plead that self-renormalization works when *I* do it, even though it wouldn't work for Anti-Inductors. I can only plead that when *I* look over my flawed mind and see a core of useful reasoning, that I am really right, even though a completely broken mind might mistakenly perceive a core of useful truth.

Reflective coherence isn't licensed for all minds. It works for me, because I started out with an epistemic gift.

It doesn't matter if the Anti-Inductors look over themselves and decide that their anti-induction also constitutes an epistemic gift; they're wrong, *I'm* right.

And if that sounds philosophically indefensible, I beg you to step back from philosophy, and conside whether what I have just said is really truly *true*.

(*Using* your own concepts of induction and simplicity to do so, of course.)

Does this sound a little less indefensible, if I mention that PA trusts only proofs from the PA axioms, not proofs from every possible set of axioms? To the extent that I trust things like induction and Occam's Razor, then of course I don't trust anti-induction or anti-Occamian priors—*they wouldn't start working just because I adopted them.*

What I trust isn't a ghostly variable-framework from which I arbitrarily picked one possibility, so that picking any other would have worked as well so long as I renormalized it. *What I trust* is induction and Occam's Razor, which is why I use them to think about induction and Occam's Razor.

(Hopefully I have not just licensed myself to trust myself; only licensed being moved by both implicit and explicit appeals to induction and Occam's Razor. Hopefully this makes me PA+1, not Self-PA.)

So there is no *general, epistemological* license to be a self-renormalizing factual reasoning system.

The reason my system works is because it started out fairly inductive—not because of the naked meta-fact that it's trying to renormalize itself using *any* system; only induction counts. The license—no, the *actual usefulness*—comes from the inductive-ness, not from mere reflective-ness. Though I'm an inductor who says so!

And, sort-of similarly, but not exactly analogously:

There is no general *moral* license to be a self-renormalizing decision system. Self-consistency in your decision algorithms is not that-which-is-right.

The Pebblesorters place the entire meaning of their lives in assembling correct heaps of pebbles and scattering incorrect ones; they don't know what makes a heap correct or incorrect, but they know it when they see it. It turns out that prime heaps are correct, but determining primality is not an easy problem for their brains. Like PA and unlike PA+1, the Pebblesorters are moved by particular and specific arguments tending to show that a heap is correct or incorrect (that is, prime or composite) but they have no explicit notion of "prime heaps are correct" or even "Pebblesorting People can tell which heaps are correct or incorrect". They just know (some) correct heaps when they see them, and can try to figure out the others.

Let us suppose by way of supposition, that when the Pebblesorters are presented with the essence of their decision system—that is, the primality test—they recognize it with a great leap of relief and satisfaction. We can spin other scenarios—Peano Arithmetic, when presented with itself, does not prove itself correct. But let's suppose that the Pebblesorters recognize a wonderful method of systematically producing correct pebble heaps. Or maybe they don't endorse Adleman's test as being the essence of correctness—any more than Peano Arithmetic proves that what PA proves is true—but they do recognize that Adleman's test is a wonderful way of producing correct heaps.

Then the Pebblesorters have a reflectively coherent decision system.

But this does not constitute a disagreement between them and humans about what is *right,* any more than humans, in scattering a heap of 3 pebbles, are disagreeing with the Pebblesorters about which numbers are prime!

The Pebblesorters are moved by arguments like "Look at this row of 13 pebbles, and this row of 7 pebbles, arranged at right angles to each other; how can you see that, and still say that a heap of 91 pebbles is correct?"

Human beings are moved by arguments like "Hatred leads people to play purely negative-sum games, sacrificing themselves and hurting themselves to make others hurt still more" or "If there is not the threat of retaliation, carried out even when retaliation is profitless, there is no credible deterrent against those who can hurt us greatly for a small benefit to themselves".

This is not a minor difference of flavors. When you reflect on the kind of arguments involved here, you are likely to conclude that the Pebblesorters *really are* talking about primality, whereas the humans *really are* arguing about what's right. And I

agree with this, since I am not a moral relativist. I don't think that morality being moral implies any ontologically basic physical rightness attribute of objects; and conversely, I don't think the lack of such a basic attribute is a reason to panic.

I may have contributed to the confusion here by labeling the Pebblesorters' decisions "p-right". But what they are talking about is not a different brand of "right". What they're talking about is prime numbers. There is no general rule that reflectively coherent decision systems are *right;* the Pebblesorters, in *merely* happening to implement a reflectively coherent decision system, are not yet talking about *morality!*

It's been suggested that I should have spoken of "p-right" and "h-right", not "p-right" and "right".

But of course I made a very deliberate decision not to speak of "h-right". That sounds like there is a general license to be human.

It sounds like being human is the essence of rightness. It sounds like the justification framework is "this is what humans do" and not "this is what saves lives, makes people happy, gives us control over our own lives, involves us with others and prevents us from collapsing into total self-absorption, keeps life complex and non-repeating and aesthetic and interesting, dot dot dot etcetera etcetera".

It's possible that the above value list, or your equivalent value list, may not sound like a compelling notion unto you. Perhaps you are only moved to perform *particular* acts that make people happy—not caring all that much yet about this general, explicit, verbal notion of "making people happy is a value". Listing out your values may not seem very valuable to you. (And I'm not even arguing with that judgment, in terms of everyday life; but a Friendly AI researcher has to know the metaethical score, and you may have to judge whether funding a Friendly AI project will make your children happy.) Which is just to say that you're behaving like PA, not PA+1.

And as for that value framework being valuable because it's human—why, it's just the other way around: humans have received a *moral gift*, which Pebblesorters lack, in that we started out interested in things like happiness instead of just prime pebble heaps.

Now this is not actually a case of someone reaching in from outside with a gift-wrapped box; any more than the "moral miracle" of blood-soaked natural selection producing Gandhi, is a real miracle.

It is only when you look out from *within* the perspective of morality, that it seems like a great wonder that natural selection could produce true friendship. And it is only when you look out from within the perspective of morality, that it seems like a great blessing that there are humans around to colonize the galaxies and do something interesting with them. From a purely causal perspective, nothing unlawful has happened.

But from a moral perspective, the wonder is that there are these human brains around that happen to *want* to help each other—a great wonder indeed, since human brains don't *define* rightness, any more than natural selection *defines* rightness.

And that's why I object to the term "h-right". *I am not trying to do what's human. I am not even trying to do what is reflectively coherent for me. I am trying to do what's right.*

It may be that humans argue about what's right, and Pebblesorters do what's prime. But this doesn't change what's right, and it doesn't make what's right vary from planet to planet, and it doesn't mean that the things we do are right in mere virtue of our deciding on them—any more than Pebblesorters *make* a heap prime or not prime by deciding that it's "correct".

The Pebblesorters aren't trying to do what's p-prime any more than humans are trying to do what's h-prime. The Pebblesorters are trying to do what's prime. And the humans are arguing about, and occasionally even really trying to do, what's right.

The Pebblesorters are not trying to create heaps of the sort that a Pebblesorter would create (note circularity). The Pebblesorters don't think that Pebblesorting thoughts have a special and supernatural influence on whether heaps are prime. The Pebblesorters aren't *trying* to do anything explicitly related to Pebblesorters—just like PA isn't trying to prove anything explicitly related to proof. PA just talks about numbers; it took a special and additional effort to encode any notions of proof in PA, to make PA talk about itself.

PA doesn't ask explicitly whether a theorem is provable in PA, before accepting it—indeed PA wouldn't care if it did prove that an encoded theorem was provable in PA. Pebblesorters don't care what's p-prime, just what's prime. And I don't give a damn about this "h-rightness" stuff: there's no license to be human, and it doesn't justify anything.

Part of *The Metaethics Sequence*

Next post: "Invisible Frameworks"

Previous post: "You Provably Can't Trust Yourself"

Invisible Frameworks

Followup to: Passing the Recursive Buck, No License To Be Human

Roko has mentioned his "Universal Instrumental Values" several times in his comments. Roughly, Roko proposes that we ought to adopt as terminal values those things that a supermajority of agents would do instrumentally. On Roko's blog he writes:

> I'm suggesting that UIV provides the cornerstone for a rather new approach to goal system design. Instead of having a fixed utility function/supergoal, you periodically promote certain instrumental values to terminal values i.e. you promote the UIVs.

Roko thinks his morality is more *objective* than mine:

> It also worries me quite a lot that eliezer's post is entirely symmetric under the action of replacing his chosen notions with the pebble-sorter's notions. This property qualifies as "moral relativism" in my book, though there is no point in arguing about the meanings of words.
>
> My posts on universal instrumental values are not symmetric under replacing UIVs with some other set of goals that an agent might have. UIVs are the unique set of values X such that in order to achieve any other value Y, you first have to do X.

Well, and this proposal has a number of problems, as some of the commenters on Roko's blog point out.

For a start, Roko actually says "universal", not "supermajority", but there are no actual universal actions; no matter what the green button *does,* there are possible mind designs whose utility function just says "Don't press the green button." There is no button, in other words, that all possible minds will press. Still, if you defined some prior weighting over the space of possible minds, you could probably find buttons that a supermajority would press, like the "Give me free energy" button.

But to do *nothing* except press such buttons, consists of constantly losing your purposes. You find that driving the car is useful for getting and eating chocolate, or for attending dinner parties, or even for buying and manufacturing more cars. In fact, you realize that *every* intelligent agent will find it useful to travel places. So you start driving the car around without any destination. Roko hasn't noticed this because, by anthropomorphic optimism, he mysteriously only thinks of humanly appealing "UIVs" to propose, like "creativity".

Let me guess, Roko, you don't think that "drive a car!" is a "valid" UIV for some reason? But you did not apply some fixed procedure you had previously written down, to decide whether "drive a car" was a valid UIV or not. Rather you started out feeling a moment of initial discomfort, and then looked for reasons to disapprove. I wonder why the same discomfort didn't occur to you when you considered "creativity".

But let us leave aside the universality, appeal, or well-specified-ness of Roko's metaethics.

Let us consider only Roko's claim that his morality is more *objective* than, say, mine, or this marvelous list by William Frankena that Roko quotes SEP quoting:

> Life, consciousness, and activity; health and strength; pleasures and satisfactions of all or certain kinds; happiness, beatitude, contentment, etc.; truth; knowledge and true opinions of various kinds, understanding, wisdom; beauty, harmony, proportion in objects contemplated; aesthetic experience; morally good dispositions or virtues; mutual affection, love, friendship, cooperation; just distribution of goods and evils; harmony and proportion in one's own life; power and experiences of achievement; self-expression; freedom; peace, security; adventure and novelty; and good reputation, honor, esteem, etc.

So! Roko prefers his Universal Instrumental Values to this, because:

> It also worries me quite a lot that eliezer's post is entirely symmetric under the action of replacing his chosen notions with the pebble-sorter's notions. This property qualifies as "moral relativism" in my book, though there is no point in arguing about the meanings of words.
>
> My posts on universal instrumental values are not symmetric under replacing UIVs with some other set of goals that an agent might have. UIVs are the unique set of values X such that in order to achieve any other value Y, you first have to do X.

It would seem, then, that Roko attaches tremendous *importance* to claims to asymmetry and uniqueness; and tremendous disaffect to symmetry and relativism.

Which is to say that, when it comes to metamoral arguments, Roko is greatly moved to adopt morals by the statement "this goal is universal", while greatly moved to reject morals by the statement "this goal is relative".

In fact, so strong is this tendency of Roko's, that the metamoral argument "Many agents will do X!" is sufficient for Roko to adopt X as a terminal value. Indeed, Roko thinks that we ought to get *all* our terminal values this way.

Is this objective?

Yes and no.

When you evaluate the question "How many agents do X?", the answer does not depend on which agent evaluates it. It does depend on quantities like your weighting over all possible agents, and on the particular way you slice up possible events into categories like "X". But let us be charitable: if you adopt a fixed weighting over agents and a fixed set of category boundaries, the question "How many agents do X?" has a unique answer. In this sense, Roko's meta-utility function is objective.

But of course Roko's meta-utility function is not "objective" in the sense of universal compellingness. It is only Roko who finds the argument "Most agents do X instrumentally" a compelling reason to promote X to a terminal value. I don't find it compelling; it looks to me like losing purpose and double-counting expected utilities. The vast majority of possible agents, in fact, will not find it a compelling argument! A paperclip maximizer perceives no utility-function-changing, metamoral valence in the proposition "Most agents will find it useful to travel from one place to another."

Now this seems like an extremely obvious criticism of Roko's theory. Why wouldn't Roko have thought of it?

Because when Roko feels like he's being *objective,* he's *using* his meta-morality as a fixed given—evaluating the question "How many agents do X?" in different places and times, but not asking any different questions. The answer to his meta-moral question has occurred to him as a variable to be investigated; the meta-moral question itself is off the table.

But—of course—when a Pebblesorter regards "13 and 7!" as a powerful metamoral argument that "heaps of 91 pebbles" should not be a positive value in their utility function, they are asking a question whose answer is the same in all times and all places. They are asking whether 91 is prime or composite. A Pebblesorter, perhaps, would feel the same powerful surge of objectivity that Roko feels when Roko asks the question "How many agents have this instrumental value?" But in this case it readily occurs to Roko to ask "Why care if the heap is prime or not?" As it does not occur to Roko to ask, "Why care if this instrumental goal is universal or not?" Why... isn't it just *obvious* that it matters whether an instrumental goal is universal?

The Pebblesorter's framework is readily visible to Roko, since it differs from his own. But when Roko asks his own question—"Is this goal universally instrumental?"—he sees only the answer, and not the question; he sees only the output as a potential variable, not the framework.

Like PA, that only sees the compellingness of particular proofs that use the Peano Axioms, and does not consider the quoted Peano Axioms as subject matter. It is only PA+1 that sees the framework of PA.

But there is always a framework, every time you are moved to change your morals—the question is whether it will be invisible to you or not. That framework is always implemented in some particular brain, so that the same argument would fail to compel a differently constructed brain—though this does not imply that the framework makes any mention of brains at all.

And this difficulty of the invisible framework is at work, every time someone says, "But of course the correct morality is just *the one that helps you survive / the one that helps you be happy*"—implicit there is a supposed framework of meta-moral arguments that move you. But maybe I don't think that being happy is the one and only argument that matters.

Roko is adopting a special and unusual metamoral framework in regarding "Most agents do X!" as a compelling reason to change one's utility function. Why might Roko find this appealing? Humans, for very understandable reasons of evolutionary psychology, have a universalizing instinct; we think that a valid argument should persuade anyone.

But what happens if we confess that such thinking can be valid? What happens if we confess that a meta-moral argument can (in its invisible framework) use the universalizing instinct? Then we have... just done something very human. We haven't explicitly adopted the rule that all human instincts are good because they are human—but we did use one human instinct to think about morality. We didn't explicitly think that's what we were doing, any more than PA quotes itself in every proof; but we felt that a universally instrumental goal had this appealing quality of objective-ness about that, which is a perception of an intuition that evolved. This doesn't mean that objective-ness is subjective. If you define objectiveness precisely then the question "What is objective?" will have a unique answer. But it does mean that we have just been compelled by an argument that will not compel every possible mind.

If it's okay to be compelled by the appealing objectiveness of a moral, then why not also be compelled by...

...life, consciousness, and activity; health and strength; pleasures and satisfactions of all or certain kinds; happiness, beatitude, contentment, etc.; truth; knowledge and true opinions of various kinds, understanding, wisdom...

Such values, if precisely defined, can be just as objective as the question "How many agents do X?" in the sense that "How much health is in this region here?" will have a single unique answer. But it is humans who care about health, just as it is humans who care about universalizability.

The framework by which we care about health and happiness, as much evolved, and human, and part of the very substance of that which we name *right* whether it is human or not... as our tendency to find universalizable morals appealing.

And every sort of thing that a mind can do will have some framework behind it. Every sort of argument that can compel one mind, will fail to be an argument in the framework of another.

We are in the framework we name *right;* and every time we try to do what is *correct,* what we *should*, what we *must,* what we *ought*, that is the question we are asking.

Which question *should* we ask? What is the *correct* question?

Don't let your framework to *those* questions be invisible! Don't think you've answered them without asking any questions!

There is always the meta-meta-meta-question and it always has a framework.

I, for one, have decided to answer such questions the *right* way, as the alternative is to answer it the *wrong* way, like Roko is doing.

And the Pebblesorters do not disagree with any of this; they do what is objectively prime, not what is objectively right. And the Roko-AI does what is objectively often-instrumental, flying starships around with no destination; I don't disagree that travel is often-instrumental, I just say it is not right.

There is no right-ness that isn't in any framework—no feeling of rightness, no internal label that your brain produces, that can be detached from any method whatsoever of computing it—that just isn't what we're talking about when we ask "What should I do now?" Because if anything labeled *should,* is *right,* then *that* is Self-PA.

Part of *The Metaethics Sequence*

(end of sequence)

Previous post: "No License To Be Human"