The golden rule

Paul Christiano Follow
Nov 29, 2014 · 6 min read

Most of my moral intuitions are well-encapsulated by the maxim: "do unto others as you would have them do unto you." This is a principle which has extremely broad intuitive appeal, and so it seems worth exploring how I end up with a relatively unusual ethical perspective.

I think there are at least four ways in which my moral views are somewhat unusual:

- I am a committed consequentialist.
- I think that helping twice as many people is roughly twice as good, and I am unusually quantitative and hard-nosed about altruism.
- I think that the welfare of future people matters about as much as the welfare of existing people.
- I think that making happy people is extremely valuable, and would be deeply saddened by a future in which the population was much smaller than it could be.

I think that there are a few modest differences in my understanding of the golden rule that lead to these practical differences:

1. I think about helping twice as many people as if I were helping each person with twice the probability. (Note that from behind a veil of ignorance, if twice as many people are helped I *do* have twice the probability of being helped.)
2. I am very happy to exist. I would accept relatively large decreases in welfare, in exchange for a higher probability of existing.

Consequentialism

When people are considering several policies, some of which might help me, I would really prefer they do the one that helps me the most. I don't care if they are virtuous, or if they violate deontological side constraints, or if they act according to a maxim that could be

universalized, or whatever (except insofar as those things bear on how effectively they help me). I just want to get the things that I want. So when I consider helping other people, I think I should just help them as effectively as possible, rather than being virtuous, satisfying side constraints, acting according to universalizable maxims, etc.

(I should flag some further subtleties here concerning meddling preferences, but they don't change the conclusion and I don't want to get bogged down.)

Aggregation

As a consequence of [1], when I think about giving X to two people or giving Y to one person, I try to think about whether I would prefer receive X with probability 2% or Y with probability 1%. I think that this is a relatively uncommon perspective. My best guess is that the difference is mostly a manifestation of my very quantitative attitude, rather than serious philosophical disagreements.

The future

I consider helping people roughly equally good whether they live now or in the future. I don't like the version of the golden rule that says "Do unto others [who live nearby] as you would have them do unto you," and I don't like the version that says "Do unto others [who are your contemporaries] as you would have them do unto you" either. And I definitely don't like the version that says "Do unto others [who might do back] as you would have them do unto you."

Though it may be question-begging, the reason I don't like these principles is that they violate the golden rule: I want people in history to try to help me live a good life. When I think about people helping *me* I'm not going to say "this food is tasty but I *really* wish that it had been secured by the good will of my neighbor rather than someone living 50 years ago and 10,000 miles away." There is a further thing I find valuable in the goodwill of neighbors and the vitality of a community, but that's just another good that should be weighted up in the calculus. And given the kinds of tradeoffs that are actually available, I tend to think that goods like health and prosperity are just a lot more important than the advantages of being able to experience gratitude in person rather than owing it to someone who lived long ago or far away; our ability to help those who live far away or in the future appears to be so much greater than our ability to help those

around us (at least for readers living in the rich world of the 21st century).

Creating people

As a consequence of [3], when I think about changes that would bring new people into the world with good lives, I tend to think that those changes are valuable. If I imagine someone deciding whether they should take an action that would bring me into existence, I very much want them to take it. In combination with [1], when I think about the comparison between 20 billion people existing under conditions X or 10 billion people existing under conditions Y, I try to think about whether I would prefer exist under conditions X with probability 2% or under conditions Y with probability 1%. (Carl Shulman has written a post that takes this idea to one logical extreme.)

The kinds of tradeoffs that I find myself considering in my own do-gooding seem to be modest losses in quality of life for existing people in exchange for a small improvement in the probability of many billions of billions of people existing and having rich and valuable lives. Scaled up, I think the tradeoffs look something like making existing people 10% poorer in exchange for increasing by 0.1% the probability of a prosperous and very (very) large future. On the kind of calculus outlined above, I am quite happy with this tradeoff.

While this view is still related to my very quantitative outlook, I think disagreement about this view is much more likely to be due to philosophical disagreements.

One common objection to this view is "if no one exists, no one will be sad about their non-existence." I consider this objection very weak. If someone does exist, someone will be happy about their existence. So even if non-existence is not bad, existence is good.

A problem I consider more serious is that this view seems to be very sensitive to the outlook of possible people regarding existence. I would be happy to make my life significantly worse in exchange for a higher probability of existing, but many people feel differently. Moreover, this difference seems to be more about a difference in outlook, rather than a difference in quality of life. This seems particularly problematic given that for nearly any outlook, we can imagine possible people with that outlook. Should we be happy to create any possible people who are enthusiastic about existing, and unhappy to create any possible people who aren't? What if the actual

character of their experiences were almost the same, and only their attitudes about non-existence differed?

For example, regardless of the actual content of individuals' experiences, we might expect natural selection to produce people who would prefer to exist. So if we think there is *any* kind of experience that would be a net negative, but which could have been produced by natural selection, then we might run into a conflict.

I don't want to get into a long discussion here, but my own best guess is that if someone would prefer existing than to not existing (knowing all of the facts, reflecting appropriately, and so on), then all else equal I would prefer that person to exist. These intuitions become weaker as we move further away from the kinds of minds I am familiar with, and aren't that strong to begin with, but I'll postpone a longer discussion for some future day.

Utilitarianism

Incidentally, though I am a committed consequentialist I am not a hedonistic utilitarian. This perspective sometimes puts me at odds with a few acquaintances. I haven't listed this view above because I think that in the world at large, hedonistic utilitarianism is extremely unusual; however, it *is* another moral issue where I feel like my view is driven by the golden rule. I have preferences for things other than my own experiences of pleasure and pain, and even if I did not I can recognize that *if* I had other preferences, I would prefer that others respect those preferences.

. . .

Originally published at rationalaltruist.com on August 23, 2014.