

altruism (slight thread necromancy)

The Urth list — 1 messages, 1 voices, 01 Apr 2005

<https://urth.darkrealm.vip/thread/urth/4389>

Exported from an unofficial mirror of a public mailing list. Copyright in each message remains with the person who wrote it. Email addresses are obscured as `user at host`, the form the original archive published. This file was rebuilt from parsed fields rather than retained headers, so it is faithful in content but not byte-exact.

maru — 01 Apr 2005, 18:12

I realize our debate on altruism and ethics is a bit dead, and pointless but I saw this on a another list and it seems pretty relevant to our discussions so I'm forwarding it. Makes for interesting reading, anyway.

~Maru

From <http://sl4.org/wiki/DialogueOnFriendliness>
(Eliezer-2003 era, partially superceded by CollectiveVolition)

Cathryn: All right. Suppose I wished for the genie to grab an ice cream cone from a little girl and give it to me. Now it might be a really delicious and satisfying ice cream cone, but it would still be wrong to take the ice cream cone away from the little girl. Isn't your definition of satisfaction fundamentally selfish?

Dennis: I'll say! *I* should get the ice cream cone.

Bernard: Well, of course, the so-called altruist is also really selfish. It's just that the altruist is made happy by other people's happiness, so he tries to make other people happy in order to increase his own happiness.

Cathryn: That sounds like a silly definition. It sounds like a bunch of philosophers trying to get rid of the inconvenient square peg of altruism by stuffing it into an ill-fitting round hole. That is just not how altruism actually works in real people, Bernard.

Autrey: I wouldn't dismiss the thought entirely. The philosopher Raymond Smullyan once asked: "Is altruism sacrificing your own happiness for the happiness of others, or gaining your happiness through the happiness of others?" I think that's a penetrating question.

Eileen: I would say that altruism is making choices so as to maximize the expected happiness of others. My favorite definition of altruism is one I found in a glossary of Zen: "Altruistic behavior: An act done without any intent for personal gain in any form. Altruism requires that there is no want for material, physical, spiritual, or egoistic gain."

Cathryn: No *spiritual* gain?

Eileen: That's right.

Bernard: That sounds like Zen, all right - self-contradictory, inherently impossible of realization. Different people are made happy by different things, but everyone does what makes them happy. If the

altruist were not made happy by the thought of helping others, he wouldn't do it.

Autrey: I may be made happy by the thought of helping others. That doesn't mean it's the reason I help others.

Cathryn: Yes, how would you account for someone who sacrifices her life to save someone else's? She can't possibly anticipate being happy once she's dead.

Autrey: Some people do.

Cathryn: I don't. And yet there are still things I would give my life for. I think. You can't ever be sure until you face the crunch.

Eileen: There you go, Kathryn. There's your counterexample.

Cathryn: Huh?

Eileen: If your wish is to sacrifice your life so that someone else may live, you can't say "Yes, I'm satisfied" afterward.

Autrey: If you have a genie on hand, you really should be able to think of a better solution than that.

Eileen: Perhaps. Regardless, it demonstrates at least one hole in that definition of volition.

Bernard: It is not a hole in the definition. It is never rational to sacrifice your life for something, precisely because you will not be around to experience the satisfaction you anticipate. A genie should not fulfill irrational wishes.

Autrey: Kathryn knows very well that she cannot feel anything after she dies, and yet there are still things she would die for, as would I. We are not being tricked into that decision, we are making the choice in full awareness of its consequences. To quote Tyrone Pow, "An atheist giving his life for something is a profound gesture." Where is the untrue thing that we must believe in order to make that decision? Where is the inherent irrationality? We do not make that choice in anticipation of feeling satisfied. We make it because some things are more important to us than feeling satisfaction.

Bernard: Like what?

Cathryn: Like ten other people living to fulfill their own wishes. All sentients have the same intrinsic value. If I die, and never get to experience any satisfaction, that's more than made up for by ten other people living to experience their own satisfactions.

Bernard: Okay, what you're saying is that other people's happiness is weighted by your goal system the same as your own happiness, so that when ten other people are happy, you experience ten times as much satisfaction as when you yourself are happy. This can make it rational to sacrifice for other people - for example, you donate a thousand dollars to a charity that helps the poor, because the thousand dollars can create ten times as much happiness in that charity as it could

create if you spent it on yourself. What can never be rational is sacrificing your life, even to save ten other lives, because you won't get to experience the satisfaction.

Cathryn: What? You're saying that you wouldn't sacrifice your own life even to save the entire human species?

Bernard: (Laughs.) Well, I don't always do the rational thing.

Cathryn: Argh. You deserve to be locked in a cell for a week with Ayn Rand.

Autrey: Bernard, I'm not altruistic because I anticipate feeling satisfaction. The reward is that other people benefit, not that I experience the realization that they benefit. Given that, it is perfectly rational to sacrifice my life to save ten people.

Bernard: But you won't ever know those ten people lived.

Autrey: So what? What I value is not "the fact that Autrey knows ten people lived", what I value is "the fact that ten people lived". I care about the territory, not the map. You know, this reminds me of a conversation I once had with Greg Stock. He thought that drugs would eventually become available that could simulate any feeling of satisfaction, not just simple ecstasy - for example, drugs that simulated the feeling of scientific discovery. He then went on to say that he thought that once this happened, everyone would switch over to taking the drugs, because real scientific discovery wouldn't be able to compare.

Cathryn: Yikes. I wouldn't go near a drug like that with a ten-lightyear pole.

Autrey: That's what I said, too - that I wanted to genuinely help people, not just feel like I was doing so. "No," said Greg Stock, "you'd take them anyway, because no matter how much you helped people, the drugs would still make you feel ten times better."

Cathryn: That assumes I'd take the drugs to begin with, which I wouldn't ever do. I don't want to be addicted. I don't want to be transformed into the person those drugs would make me.

Autrey: The strange thing was that Greg Stock didn't seem to mind the prospect. It sounded like he saw it as a natural development.

Cathryn: So where'd the conversation go after that?

Autrey: I wanted to talk about the difference between psychological egoism and psychological altruism. But it was a bit too much territory to cover in the thirty seconds of time I had available.

Dennis: Psychological egoism and psychological altruism? Eh?

Eileen: The difference between a goal system that optimizes an internal state and a goal system that optimizes an external state.

Cathryn: There's a formal difference?

Eileen: Yes.

Bernard: No.

Cathryn: Interesting.

Autrey: In philosophy, this is known as the egoism debate. It's been going on for a while. I don't really agree with the way the arguments are usually phrased, but I can offer a quick summary anyway. You want one?

Dennis: Yeah.

Autrey: Okay. Psychological egoism is the position that all our ultimate ends are self-directed. That is, we can want external things as means to an end, but all our ultimate ends - all things that we desire in themselves rather than for their consequences - are self-directed in the sense that their propositional content is about our own states.

Eileen: Propositional content? Sounds rather GOFAL-ish.

Autrey: Maybe, but it's the way the standard debate is phrased. Anyway, let's say I want it to be the case that I have a chocolate bar. This desire is purely self-directed, since the propositional content mentions me and no other agent. On the other hand, suppose I want it to be the case that Jennie has a candy bar. This desire is other-directed, since the propositional content mentions another person, Jennie, but not myself. Psychological egoism claims that all our ultimate desires are self-directed; psychological altruism says that at least some of our ultimate desires are other-directed.

Bernard: If you want Jennie to have a candy bar, it means that you would be happy if Jennie got a candy bar. Your real end is always happiness.

Autrey: That's known as psychological hedonism, which is a special case of psychological egoism. As Sober and Wilson put it, "The hedonist says that the only ultimate desires that people have are attaining pleasure and avoiding pain... the salient fact about hedonism is its claim that people are motivational solipsists; the only things they care about ultimately are states of their own consciousness. Although hedonists must be egoists, the reverse isn't true. For example, if people desire their own survival as an end in itself, they may be egoists, but they are not hedonists." Another quote from the same authors: "Avoiding pain is one of our ultimate goals. However, many people realize that being in pain reduces their ability to concentrate, so they may sometimes take an aspirin in part because they want to remove a source of distraction. This shows that the things we want as ends in themselves we may also want for instrumental reasons... When psychological egoism seeks to explain why one person helped another, it isn't enough to show that one of the reasons for helping was self-benefit; this is quite consistent with there being another, purely altruistic, reason that the individual had for helping. Symmetrically, to refute egoism, one need not cite examples of helping in which only other-directed motives play a role. If people

sometimes help for both egoistic and altruistic ultimate reasons, then psychological egoism is false."

Dennis: The very notion of altruism is incoherent.

Autrey: That argument is indeed the chief reason why some philosophers espouse psychological hedonism.

Cathryn: Sounds like a lot of silly philosophizing to me. Does it really matter whether I'm considered a "motivational solipsist" or whatever, as long as I actually help people?

Bernard: That's just it! It doesn't make any operational difference - all goal systems operate to maximize their internal satisfaction, no matter what external events cause satisfaction.

Eileen: That's not true; it does make an operational difference. If Autrey values the solipsistic psychological event of knowing he saved ten lives, he will never sacrifice his own life to save ten other lives; if he values those ten lives in themselves, he may. You told him that, remember?

Bernard: Well, I guess Autrey might value the instantaneous happiness of knowing he chose to save ten lives, more than he values all the happiness he might achieve in the rest of his life.

Cathryn: That doesn't sound anything remotely like the way real people think. Square peg, round hole.

Autrey: Do you have anything new to contribute to the debate, Eileen? It's a pretty ancient issue in philosophy.

Eileen: The basic equation for a Bayesian decision system is usually phrased something like $D(a) = \sum U(x)P(x|a)$. This is known as the expected utility equation, and it was derived by von Neumann and Morgenstern in 1944 as a unique constraint on preference orderings for all systems that obey certain consistency axioms -

Dennis: Start over.

Eileen: Okay. Imagine that $D(a)$ stands for the "desirability" of an action A , that $U(x)$ stands for the "utility" of a state of the universe X , and $P(x|a)$ is your assigned "probability" that the state X occurs, given that you take action A . For example, let's say that I show you two spinner wheels, the red spinner and the green spinner. One-third of the red spinner wheel is black, while two-thirds of the green spinner wheel is white. Both spinners have a dial that I'm going to spin around until it settles at random into a red or black area (for the red spinner) or a white or green area (for the green spinner). The red spinner has a one-third chance of turning up black, while the green spinner has a two-thirds chance of turning up white. Let's say that I offer you one of two choices; you can pick the red spinner and get a chocolate ice cream cone if the spinner turns up black, or you can pick the green spinner and get a vanilla ice cream cone if the spinner turns up white.

(And then it gets complicated...)

--

Eliezer S. Yudkowsky <http://singinst.org/>

Research Fellow, Singularity Institute for Artificial Intelligence