Existential Risk / Opportunity Singularity Management

January 31, 2022

Contents:

A Philosophy that Needs Improvement

by James Blodgett

When I was in high school, I tried out for a school play. It was a play about Don Quixote. It did not have the sophistication of Man of La Mancha, but its lack of sophistication made it an ideal play for high school ham actors. I tried out for a minor part. They gave me the lead. It was type casting. Don Quixote was a strange character, somewhat of a nut case. I looked the part, especially with makeup. I wore a crepe hair beard and put some white in my hair and lines in my face. I wondered if I would ever look like that in real life. Years later I did. Over the years I went through stages, first clean shaven, then a moustache, then a Don Quixote goatee, then a full white Santa Claus beard. For a few Christmases, I wore a Santa Claus hat while wrapping presents. During Covid times I started cutting my own hair, and it got a bit wild. I started thinking about trying out to play the lead in An Act of God, often depicted with rather wild long white hair. During the Omicron spike, I noticed that my beard was preventing a tight mask seal, perhaps endangering myself and others, so I shaved off my beard. My self cut hair still gives me somewhat of a weirdo look.

I mention this because, on one level, Don Quixote was crazy, but on another level he was noble because he was trying to save the world. I am trying to save the world too. I am not crazy, but the rational version is to try to do something that is quite improbable. i.e. managing singularities. How can we manage something as humongous as a singularity? One answer is expected value, i.e., value times probability, a good way to

bet. If the cost of enabling a singularity is less than its expected value, that is a good reason for making the attempt. However, expected value has problems at the extremes. See my big trolley problem in the April 2018 issue of EROSM. Expected value is also very difficult to compute because most actions can trigger many potential outcomes, the probability of those outcomes is rarely known precisely, and both probability and value can usually be estimated only roughly. Some of those potential outcomes can be bad , and sometimes their negative expected value might outweigh the positive expected value of the desired outcomes. Because of the difficulty with estimation, it is hard to know for sure if this is the case.

I try to think carefully about potential outcomes, especially bad ones. I don't want to be the guy who tells a super intelligent computer to make a lot of paperclips, so it uses its super intelligence to develop magic physics that turns the entire universe, including humans, into paperclips. It might seem best to do nothing, but we have to do something, because doing nothing is also a thing we might do, and doing nothing is usually not the best choice.

So here is the challenge for readers: see if you can help me improve my philosophy. Our SIG website, mentioned above as a source for this publication, has a contact email. Click on "contact us" on the left panel. I am interested in your opinion. If you have a really good contribution, submit it for potential publication.