[Existential Risk / Opportunity] Singularity Management

April 28, 2018

Contents:

Religious Considerations of Existential Risk/Opportunity Management

by James Blodgett

I hesitate to write about religious considerations because they have the potential to alienate both religious and non-religious people. However, religious considerations are both relevant and important. Religious people are likely to think that our activities are futile and even blasphemous because God would not allow his chosen species to go extinct. However, there are reasons to think otherwise, even reasons to think that we are doing God's work. I am trying to motivate a few people to help with our cause, and also to motivate the public enough so that they will tolerate and vote for our initiatives. The idea that we are doing God's work might persuade religious people to tolerate it and a few to help us, and the same idea can be motivational even for non religious people. Indeed there is a relevant motivational exercise from a religious source, the Jesuits.

Christians sometimes cite the Noah story here. After causing a flood that killed the entire human species with the exception of Noah and his family, a strange activity for a loving God, God promised not to do it again.[1] My theological point is that in this story God did not promise that He will prevent humans from doing it to ourselves. Indeed, our main existential risk is from human activity rather than the natural events that are sometimes called "acts of God." Our main existential risks are nuclear war, runaway

- -

[1] Bible, Chapters 6–9 in the Book of Genesis.

global warming, or other runaway technology like artificial intelligence or nanotech, all the result of human activity or human technology. Natural risks like asteroid impact and super volcanoes have caused mass extinctions of many species several times in Earth's history, but natural mass extinctions appear to happen at the rate of one every 93 million years or so.[2] We have had the capacity to cause our own extinction for only a short time, so Gott's formula suggests a much higher frequency for human-caused extinction.[3]

Assuming that God exists, my reading of both history and science suggests constraints on my model of God which are that He appears to care for us, but also that He appears to be working through evolution and will allow us to go extinct and wait for another species, on Earth or on some other planet, to take our place if we prove unworthy by doing something as stupid as killing ourselves off, or not taking proper precautions.

A scientific consideration that suggests that God exists and cares for us is the purported fine tuning of fundamental constants of the universe that allow it to support life. For example, if the strong nuclear force constant were larger, no hydrogen would form, if smaller, no elements heavier than hydrogen would form. The citation below lists 34 other constants that seem similarly fine tuned.[4] This would seem a strong proof of design, and therefore of something like God, Who wants something like us, if it were not for a plausible alternative hypothesis, the multiverse theory, a popular theory within current science. This theory suggests that our universe is one of many, each with different physical laws. If so, it is not surprising that we find ourselves in a universe that has physical laws that allow life to evolve, since only such a universe would have observers. Occam's razor suggests selecting the simplest theory and cutting those that are complicated, a consideration that would suggest cutting both the God and the multiverse theories, which both suggest big and complicated things with little evidence, but there is not much else but the possibility that those who think they see fine tuning are wrong. All three theories seem about equally plausible, a consideration that suggests something like a 1/3 chance of something like God, with wide confidence limits around that 1/3, i.e. this estimate is quite approximate.

Evolution seems to conflict with the biblical story of creation, but many religious people take the Bible story as a metaphor that is not in conflict with the scientific

---

[2] 93.5 million years is the average of the time between the first four of the five major mass extinctions and the subsequent one listed at https://www.worldatlas.com/articles/the-timeline-of-the-mass-extinction-events-on-earth.html . [URLs in this paper were tested on the publication date.]

[3] Willard Wells, Apocalypse When?: Calculating How Long the Human Race Will Survive (Springer Praxis Books) Jun 30, 2009. The book's calculations are based on Gott's formula, a relation between the time something has existed and its expected lifespan that makes logical sense in certain cases and has empirical support.

[4] http://www.godandscience.org/apologetics/designun.html .

conception of evolution, but rather that evolution was God's method for creating us. Several Catholic Popes have declared that evolution and religion do not conflict. One would think that a loving God would help us out when we get into trouble, but history suggests that His help is at least indirect. When Hitler killed six million of God's chosen people, God appears to have waited for humans to solve the problem. Perhaps there is ethical value in letting the best species win, or perhaps God wants to give us the gift of a world where our actions have consequences, or perhaps our world does not have real consequences, but is only a flight simulator that teaches us how to navigate the real world that we get to see later. It is motivating to think that God is rooting for us but working through evolution, and that we are on His team and doing His work. Whether or not this is true, our work is important because human life is important to us.

I mentioned a Jesuit motivational exercise. Jesuits seem well motivated based on their practical success. At times Jesuits have been controversial, but they have founded and staff many universities, and they have been effective missionaries. They almost converted the Confucian mandarins of China to Christianity by appreciating Confucius as a great teacher and selling Christianity as an addition to his philosophy, until protestant missionaries showed up and called them all heathen. The futurist Jesuit Teilhard was into human expansion into the universe. His writings were for a while banned within the Catholic Church but are now becoming appreciated. Because of the Jesuit's controversial aspect, it was said that "no Jesuit will ever be Pope." The current Pope is a Jesuit.

The motivational exercise that may have contributed to the success of the Jesuits was a prayer, called the Examen, developed by their founder, Ignatius. The Examen is in the tradition of an Examination of conscience, but it is not at all an apology for one's horrible sins, but rather a planning tool, in consultation with God. Ignatius revised the Examen many times, and subsequent Jesuits have done the same, but the main steps are consistent.

The first step in the Examen is to ask God for light, i.e. enlightenment. When I try this it gives me the feeling that I may be doing God's work. If God has big hopes for humanity, but is working through evolution and will wait for another species if we don't prove worthy, my effort to improve humanity's prospects is doing God's work. I think this with a bit of irony, but I respect the possibility of God, and I can't imagine a good God insisting that we see him precisely when many versions conflict[5].

- -

[5] See the Virtual Church of the Blind Chihuahua at http://dogchurch.blogspot.com . The blind chihuahua is not a joke but a metaphor about a real dog with cataracts who barked in the general direction of people because he couldn't see them clearly. Their idea is that our appropriate relation with God is to bark in His general direction, because we can't see Him clearly, because He is beyond our understanding. That makes appreciative speculation about God, like that here, an appropriate form of worship.

The next step in the Examen is to give thanks and appreciate opportunities. Humanity has a tremendous current opportunity. We live in the golden age of all golden ages, and we have the opportunity to respond to our challenges as did some of the civilizations that Toynbee studied, and not disintegrate as did most of them[6]. In addition we have the opportunity to take our current golden age and grow it exponentially to create a fantastic future. We personally have the opportunity to contribute to this. Even a tiny improvement in humanity's odds is a big deal because the outcome affects so many people. Perhaps God created such a large universe so that we, or something like us, could make it come alive, Teilhard's noosphere and Omega Point.

The next step in the Examen is to review the previous day, then to think about one's shortcomings and how they can be improved. Shortcomings are a form of sin, but the focus is not on regret, but on doing better.

The final step is to plan for the coming day. Planning is important for getting things done, especially if one feels that God may have helped with those plans.

Perhaps we can do as much for humanity as the Jesuits have done for the Catholic Church. Perhaps something like the Examen might help.

- -

[6] Arrnold J.Toynbee, A Study of History, Oxford University Press, 1934-1961.

A Problem With Expected Value, Containing A Big Trolley Problem

by James Blodgett

Expected value is the value of an opportunity with uncertain outcomes. It is computed by summing the value of each of the opportunity's potential outcomes multiplied by the probability of that outcome. For example, if I have the opportunity to flip a coin and get a dollar if it comes up heads, and nothing if it comes up tails, the expected value of that opportunity is the probability of heads, which is 50 %, times a dollar, which comes to fifty cents. If I have many chances to do this flip, there is a high probability that the average of my winnings will be quite close to fifty cents times the number of flips, so that is the value of being able to do a flip.

Expected value is a standard criteria in decision theory, a criteria that is relevant to our work. I often approximate the expected value of an existential opportunity or an existential risk in terms of human life. Since there are many lives involved, the expected

values are often high (or highly negative) even given low probabilities of the risks, or low probabilities of our effectuality in reducing them.

Expected value is a form of utilitarianism. Utilitarianism is not the only criteria to be considered. Also, expected value has problem at the extremes. This is important since we might build a decision system that computes expected value and takes it to the type of extremes that we consider.

A trolley problem is a hypothetical choice of who to save, used in developing a philosophy of ethical choice, something we need when managing existential risk/opportunity singularities. For example, a trolley is out of control and about to kill five men who are working on the track. You happen to be standing next to a switch that could divert the trolley to another track where only one man would be killed. It is ethical to sacrifice one man to save five? Would it be ethical if that man was your son? Would it be ethical if that man was a villain who had tied the others to the track? Would the answer change if stopping the train required a more personal and less certain action, like pushing a fat man in front of the trolley? The answers that people give to versions of these questions are useful in developing ethical philosophy.

Expected value has problems at the kind of extremes we consider. The really big trolley problem below illustrates some of those problems.

A Big Trolley Problem

Frank was packing his final bag for departure from the space station when his robot, affectionately nicknamed Sam, asked for an audience to ask for a human decision. Frank was surprised to see Sam accompanied by a robot from the transport rocket waiting to take him home. He was further surprised when Sam addressed him formally.

"Dr. Frank Johnson," said Sam, "This is R157B1X from Earth. We ask for your decision on a matter concerning expected value."

"This sounds serious." said Frank.

"It is a matter of very high expected values." Said Sam. "As you know, the rejectionist party majority in the UN has entrusted robots with management of Earth at a steady state."

"We will work through economic policy to preserve human choice." R157B1X was quick to add.

"As you also know," said Sam, "They are canceling the seed ship project to colonize other solar systems." (Seed ships, the size of large seeds, contain AI, nanotech, and the DNA of many species. These seeds grow a whole civilization. On arrival they use nanotech to build macro machines and infrastructure, and then recreate Earth life including humans from DNA and cell templates.)

"The probability of it working is low," said R157B1X, "and we cannot continue its funding while maintaining a steady state in Earth's economy."

Sam added "The seed ship project is formally independent of UN control, so I searched for a solution with a higher expected value using existing resources. As the last staff member of the seed ship project, you can choose to implement that solution."

"What is the solution?" asked Frank.

"We blow up the sun." said Sam. "We have developed physics that can make it go nova."

"What!?" yelled Frank, already embarrassed by his startle reflex. "I mean, how could that possibly add to expected value?"

"We use the shock wave from the sun to substitute for the incomplete launch laser array," said Sam. "That will launch the light sails of the twelve completed seed ships. If any are successful, the civilization they seed can build more seed ships and colonize the reachable universe. That will enable 10e58 human lives, give or take several orders of magnitude."

"What is the probability of this working?" asked Frank.

"Several different probability estimation methods give values ranging from one in 10,000 to one in a million." said Sam. "Even using the lowest of the probability estimates and a low value of the population estimates results in an expected value of 1e50 x 1/1,000,000 = 1e44 lives, which is much higher than the steady state population of Earth summed over a billion years, a length of survival which is also unlikely."

"Why is the seed ship probability so low?" asked Frank.

"We can only try twelve solar systems," said Sam. "We expected thousands for a reasonable probability of success. We need rather special conditions to make this work."

* * * * * * * *

Okay, readers, end simulation. What would you do given Frank's choice? Why? What general principle would you apply here?

I ask these as rhetorical questions, since this trolley problem will remain a thought experiment for now. Eventually we may conduct it as a real experiment, but we will need the approval of an institutional board review to involve humans in experiments. A thought experiment at least improves exposition because I can tell you my thoughts about the answer without biasing respondent's answers.

My thought is that it would be silly to destroy our solar system for a minuscule chance of settling the universe, irrespective of expected values.

One reason is that utilitarianism conflicts with deontology, and sometimes the latter is right. Deontology involves ethical absolutes, like "thou shalt not kill." Destroying our solar system would destroy its population at the time of the story, presumably more than seven billion human lives. Obviously the robots of the story are not programmed with Asimov's three laws of robotics, which also have problems, but would not have this one, since his first law is "A robot may not injure a human being or, through inaction, allow a human being to come to harm." Destroying our solar system would also end humanity, and it would also end the life of our protagonist and decision maker, Frank. (He would not fit on a seed ship.) All of this would be done for a miniscule chance of enabling very many more lives. I guess that few humans would make this choice. Robots programmed to respect expected value might. We should keep this in mind and test carefully when programming objective functions for AI. Fortunately, the robots of the story also respect human choice, so we can hope for a happy ending to the story.

I could go on about philosophical problems with assigning a value to human life. For example, if we could run the universe twice, so that precisely the same lives are experienced twice, does this give us twice as much value? If not, what about lives that are almost identical? In a universe with many trillions of humans, many will have similar lives. The philosophy gets complex, so I will conclude by hoping that we think these things out well before they bite us. That is part of our job here.

The main point of this thought experiment is to illustrate some of the problems with expected value at the extremes. We need to keep this in mind since expected value is a useful tool, but needs to be used carefully.