[Existential Risk / Opportunity] Singularity Management

July 30, 2018

Contents:

This publication is produced by the Global Risk Reduction Special Interest Group, a SIG within US and International Mensa. Content expressed here does not reflect the opinions of Mensa, which has no opinions. To join Mensa or just see what it is about, visit http://www.us.mensa.org . Past issues of this publication are available at: http://www.global-risk-sig.org/pub.htm .

Progress in Seeking a More Thorough Safety Analysis for China's Supercollider

by James Blodgett and Tom Kerwick

As readers may recall, we published a paper in the November 2015 issue of EROSM titled "Reflecting on China's Ambition to Build the World's Most Powerful Supercollider by 2020." We have worked on the issue since. Tom did most of the significant work.

I (James) know my side of the story best. My first contribution was to find an article that said that the new collider would not be more powerful than the Large Hadron Collider (LHC) at CERN (as we had said,) but only more precise. I wondered if we should publish an erratum. However, I was wrong, because the article I had found was incomplete. It turns out that China has plans for two colliders. The first, which is closer to implementation, is intended to be more precise than the LHC, but not more powerful. The second is intended to be more powerful. At first I thought we should back off and wait until the second was closer to implementation before advocating for safety studies. Later I thought that we should advocate for safety studies for both versions to maintain what we could tout as a tradition since safety studies were done for both the LHC, and also for the earlier Relativistic Heavy Ion Collider at Brookhaven. While I was making these minor recommendations and while I was busy with other things, Tom got something done. Tom reviewed the existing documentation on the Chinese collider. They were doing preliminary safety studies, but had overlooked a specific issue, which should have been included in the analysis. He then posted a well-written technical report (I suggest following the URL below and reading it) titled "The Next Great Supercollider - Beyond the LHC / TECHNICAL NOTE / Environmental Safety Assessment." Find it at: http://www.environmental-safety.webs.com/TechnicalNote-EnvSA03.pdf . Soon after publication, and circulation to the relevant scientific groups, notification was received that China's Institute of High Energy Physics (IHEP) had initiated discussions with their colleagues at CERN on the specific issue raised, and Tom reports that the web server where this paper is stored became temporarily overloaded due to downloads.

The quest of reducing existential risk often moves slowly. It is difficult to make one's voice heard in today's calliope of voices and maze of conflicted interests. But sometimes, when we get our act together, we are heard, and it is encouraging to see another example of this.

Introducing Stuart Armstrong

by James Blodgett

Dr. Armstrong is an Alexander Tamas Fellow in Artificial Intelligence and Machine Learning at the Future of Humanity Institute at Oxford. He has written the book "Smarter Than Us: The Rise of Machine Intelligence," which I have read, and his bio lists 21 other publications. He and I both used to post on the Lifeboat Foundation discussion on Yahoo Groups, and I have exchanged email with him. Before I published the trolley problem in the April 2018 edition of EROSM, I sent him a preliminary copy, mostly to see if scholars who are more qualified than I had things to say about trolley problems that I had missed. His brief article below is his response, used with his permission.

Trolley Problem Comment

by Stuart Armstrong

I like investigating the trolley problem - it shows what underpins our various moral intuitions - but you can take it too far.

My view of the trolley problem is that a) the utilitarian answer is the correct one in the exact formulation of the problem, and b) in the real world, situations like that never come up, so all the considerations that we need to put aside in the trolley problem are relevant again. It is no coincidence that there are laws that would generally prevent the "kill 1 to save 5"; generally these laws are positive, from the utilitarian perspective. People just never have the certainty of outcome, the lack of alternative options, the ability to act with such clarity, the lack of long term consequences, etc...

In practice, people get more utilitarian as the numbers increase. Kill 1 to save 5? A travesty! Kill 1,000 to save 5,000? Now people are more willing to entertain the idea, especially if it's phrased as "killing through neglect" (e.g. reallocation of resources) rather than direct killing. Kill 1,000 to save a million? Generals do that all the time in war, and nobody objects.

Now, on to your example. First of all, some of the same criticisms as to the real trolley problem applies: the set of options is ridiculously constrained (no money for seed ships, but novaing the sun is ok?) The other is that we must maximize expected utility rather than expected lives. How much utility does someone associate with a universe empty of life, versus humanity expanding across the universe, versus a stagnant Earth-bound humanity? I go back and forth on those numbers, but I feel like I'd be willing to risk it for 1/10,000 but not for 1/million (preliminary estimates on my part).

In the real world, I would reject it out of hand, asking the robots to either come up with a better solution, or because of hope that humanity won't be in stagnation forever.

Invitation to Readers to Help Review the Literature on Making Artificial Intelligence Safe

by James Blodgett

I could use some help with this, because I am not sure what to do about it myself. I cut my teeth on the collider issue, where appropriate safety precautions were fairly clear, and for a while were being ignored. Safety considerations for artificial intelligence (AI) have not been ignored; there are a plethora of them. The problem is to sort them out, decide which is best, and advocate for that, in the context of lots of other advocates and some conflicts of interest. It is a good thing that lots of work has already been done, but that is also part of the problem. It is hard to sort through all of it and develop a solid reason to accept some ideas and reject others. Moreover, a mishmash of safety precautions can leave holes, and worst case scenarios require only one hole.

As a quick example of prior work, consider Asimov's three laws of robotics, Nick Bostrom's Superintelligence: Paths, Dangers, Strategies (reviewed in the April 2015 issue of EROSM by James Tankersley), John Havens's Heartificial Intelligence: Embracing Our Humanity to Maximize Machines, Stuart Armstrong's Smarter Than Us: The Rise of Machine Intelligence, the recent Institute of Electrical and Electronics Engineers' (IEEE) Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems, a collection of reports and standards assembled by hundreds of participants, and the Future of Humanity Institute's recent The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation which has 26 authors from six institutions. A full bibliography would include many other works.

My view of humanity's current tactical landscape regarding AI risk is that all of these studies can be sorted into two types of approach. One is alarmist, postulating worst case existential scenarios that with luck may not actualize. However, true safety demands that we consider worst case scenarios that cannot be ruled out. The other is a feel-good approach driven by the conflict of interest that scientists and engineers who have discovered an area of potential technical potency may resist being told that they can't enthusiastically gallop into that area, experimenting in all directions. In the case of AI, this tactical landscape is complicated by the fact that AI is not only a potential existential risk, it is also becoming quite useful and is a potential solution to other existential risks, so unduly restricting AI may not be the safest course. We may have to accept some risk. We only get one shot at this, so the tactical direction to take is existentially important. There are many hands on the steering wheel, sometimes a good thing since group decisions are often better. Advocacy of good ideas can influence some of the other hands. Then there is the problem of getting all to agree enough to cooperate adequately.

Right now I am not capable of making my best potential contribution since I have not yet reviewed the literature well. I am impressed by, but also dubious of, recent IEEE work, which I have so far only looked at briefly. They picked Havens, who wrote a book on AI concerns that is charming but perhaps feel-good in prioritizing ethics and human values over existential risk, as Executive Director of the effort. They recruited hundreds of people for working groups, welcomed participants who were not IEEE members, considered many diverse ethical systems, and included a conference at Asilomar, where biotechnologists famously met in 1975 to develop voluntary guidelines for regulating recombinant DNA experiments. However, the IEEE work, like Havens, doesn't seem to say much about existential risk. Other studies suggest potential solutions. I have yet to see a potential solution that would definitively solve the problem, but there may be one out there. We have to do something, since doing nothing is also an action. Our choice can't exceed the best we can do, so let's come as close as we can to that best. Thinking that is good and persuasive could add real value to the debate and improve humanity's odds. I can't promise that we can solve the problem, but we can try to add something.

I suggest that interested readers engage in a seminar in which each participant studies some issue and gives a presentation to the group, who are encouraged to learn enough about the issue to ask questions and make good comments. We could set it up on some online discussion forum where everybody could post comments. Really good presentations and comments could become articles in EROSM or elsewhere. This might or might not work. There is a chance that few people will volunteer. That is not a reason not to suggest the idea. My general strategy is to try lots of things in the hope that something works. Even one volunteer could help by reading some material and writing a review article for EROSM.

Readers who would like to explore volunteering for this should send an email to the Global Risk SIG contact address at http://www.global-risk-sig.org/contacts.htm with the subject line "AI Literature Review". If you don’t hear from someone in two weeks, try again.