[Existential Risk / Opportunity] Singularity Management

January 2018

Contents:

An Interview With Paul Werbos

Introduction & questions by James Blodgett

Introduction

Our interview subject in this issue is Paul Werbos. He has many ideas about existential risks and opportunities. I will ask some general questions, but focus on one of his recent concerns: a potential runaway version of global warming that is an extinction risk.

Paul has an extensive back story that I can only sketch here. For more, Google him and check his Wikipedia article and his website at www.werbos.com . At a time when neural networks were producing discouraging results, Paul wrote a Ph.D. thesis on back propagation that revived the field. He became a Program Director for the National Science Foundation, and approved funding for many important grants, a position that made him knowledgeable about almost everything. He retired recently. He is currently working on a new version of quantum theory intended to fix a mismatch between its theory of measurement and what it assumes about the evolution of the wave function over time.

Paul is amazingly prolific. He has many contacts and he posts on several discussion groups. I am familiar with his posts on the Lifeboat Foundation discussion on Yahoo Groups, and also on the Power Satellite Economics discussion on Google Groups.

Another of his concerns is with reusable rockets, important to enable low cost access to space. He is worried that a hot structures technology developed for reentry by Boeing so rockets can be recovered and reused, the only one of many competing technologies that worked well, is being lost as the Boeing group retires.

Paul's name worked wonders for me at a World Future Society convention in Toronto in 2012. I was there to conduct a meet up among Lifeboat Foundation attendees, and I also had just published the lead paper in the then current issue of World Futures Review. At the convention I happened to meet the editor of that publication. He invited me to lunch with several others, one of whom was Jerome Glenn, cofounder of the Millennium Project. I mentioned Lifeboat, and Jerome made a mildly disparaging remark, not surprising because Lifeboat is diverse, with many people with varying interests, some of which can seem fairly nutty. I defended by defending diversity and by mentioning a couple of top-level Lifeboat members, one of whom was Paul Werbos. Jerome was immediately impressed, and spoke of Paul's many interests. Jerome is now a Lifeboat member.

An important one of Paul's concerns is that global warming may be much worse than expected. Ocean currents are driven by the buoyancy of water, which changes with temperature and with the amount of dissolved salt. At extreme northern and southern latitudes surface water freezes, and most of its salt is expelled. The water below the ice is cold and salty because it has received the expelled salt. Both cold and salinity make the water heavier and make it sink to the bottom of the ocean, carrying dissolved oxygen. This drives a current that travels along the ocean bottom, rising at the tropics and traveling back as warm water along the surface. Global warming can shut down these currents. Currents near Antarctica have already shut down. Shut down of these currents means less oxygen in the deep ocean. That encourages a form of bacteria that produces hydrosulfuric acid, H2S, a gas that smells like rotten eggs and can kill at low concentrations. One of several theories about several global extinctions of many species in Earth's prehistory is that they were caused by this process. Paul has reasons to credit the H2S theory.

Paul discusses this in the following paper posted on his website: http://www.werbos.com/Atacama.pdf . The first part of the paper is a proposal to build massive solar arrays in a South American desert. He explains in Section 3 that he wants the solar power to reduce carbon emissions because of global warming and H2S, and explains some of the reason why this might be a problem. Solar power could reduce burning of fossil fuels that produce carbon dioxide that contributes to global warming. That makes this paper more than just a discussion of the problem, but also a potential solution. He is also actively contributing to plans for space based solar power, and he advocates for more research to increase understanding of the risk and how to avoid it.

I see Paul as a hero for trying to get this solved. He is working at a level where he just might get it done. I present him and this concern here because H2S could be an existential risk, therefore it is our concern too. We should know about potential risks, and we may be able to help.

Interview:

Blodgett: What is your estimate of the probability of the H2S scenario? Why?

Werbos: Definitive measurements, discussed in Peter Ward's book Under A Green Sky, show that H2S (and the resulting radiation) have reached levels high enough to kill every human on earth, if humans had been there at the time, five to ten times in the past history of the earth. Recently Newsweek published a new report confirming the worst: Sidney Pereira, Oxygen Is Disappearing From The World's Oceans At An Alarmingly Rapid Pace, Newsweek, 1,15, 2018. This report cites a recent scientific paper: Denise Breitburg et al, Declining oxygen in the global ocean and coastal waters, Science, Vol. 359, Issue 6371, 1/5/2018.

My best guess is that the microbes which produce H2S will start to proliferate in the Pacific Ocean about 40 years from now, because that is when the NOAA data [National Oceanic and Atmospheric Administration] on the oxygen-carrying deep layers of the Pacific show it reaching zero, and because the Pacific is already full of the nutrients needed by these bacteria, thanks to agricultural runoff. California, Peru and Chile should start to smell like sewers very soon after that, because of the way that upwelling works in the Pacific, but I would guess that mass death would begin just a few decades after that, most likely because of what the sulfur compounds start to do in the atmosphere, from acid rain to new holes in the ozone layer.

If we do nothing, I see lots of uncertainty about the timing, but not about mass death.

Like Peter Ward, I once hoped that cyanobacteria in the Pacific might block this phenomenon, which they can do in areas of lesser H2S production. But as I look at the chemistry, and at the experience in the Black Sea and the Chesapeake, I am not so hopeful about that now. It's worth looking into further, but I wouldn't count on it when our lives are at stake.

Blodgett: What are ways to address H2S?

Werbos: That's the big question, since doing nothing most likely means that we all die.

The first need, of course, is to learn more about this problem than anyone on earth (including me) knows as yet. We need a more precise understanding of what causes these archaea to proliferate. Exactly what is the "surface" dividing the danger zone, from the safe zone, along the dimensions of oxygen, temperature and nutrients in solution?

Experiments in aquaria could tell us a lot, if the folks doing the experiments know what Woods Hole knows about how to do assays of these archaea.

Just as important is new research to try to create, assess and improve options for geoengineering -- for brute force band-aids intended to switch the Antarctic currents back on as soon as possible, as cheaply as possible. The Teller/Wood/Caldeira scheme is estimated to cost less than a billion or two dollars per year, more than a hundred times less than what the Waxman climate change bill was expected to cost. But is it possible to do it without putting sulfates into the Pacific, stimulating the very archaea we want to hold back? What is the most cost-effective safe way to proceed? Can we reduce launch costs enough to create other options like mirrors in space? Could we preserve the ozone layer, in the worst case, to buy us more time? Could we find other options? We need a very focused research effort as soon as possible. We don't need to stop all global warming; it is enough to reverse the horrible stuff at the poles, which is also threatening to cause sea level rise much greater and sooner than we expected just a few years ago.

There are also ways to accelerate renewable energy, both on earth and in space, but we shouldn't let those things get in the way of saving our skin here and now.

Blodgett: If you were appointed commissioner of a new existential risk reduction agency with massive but not unlimited funding, what would you fund? How would you prioritize?

Werbos: H2S would certainly be one of the top 3-5 items on my list, and I would do just what I said above. I would also work with IT [Information technology] folks to develop better platforms for deeper international dialogue and sharing of science and technology aimed at this and other threats. Misuse of IT (including AI but also including over-empowered artificial stupidity) and misuse of nuclear technology (especially nuclear proliferation and terrorism) would be two other items at the top of my list. All of these are tricky and complex challenges, requiring lots more dialogue and more effective thinking. I give a quick summary of this, with URL citations, on my six overview slides posted at http://www.werbos.com/IT_big_picture.pdf .

I am very grateful to "the old NSF" [National Science Foundation] and others for giving me a unique opportunity to get really deep into so many issues, which need to be connected to see the real strategic picture. I would like to see more people having such an opportunity, perhaps as a track starting from the PhD in Applied Mathematics (my own starting point), with a lot more social support. Restoring the old NSF would be a good halfway step towards doing that.

Review of Turchin & Denkeberger, Global Catastrophic and Existential Risks Communication Scale

Review by James Blodgett

A new paper in Futures suggests a simple scale for communicating the relative significance of existential risks. The paper is by Alexei Turchin, who was interviewed in the October 2016 issue of EROSM, and his coauthor David Denkenberger. The scale is simple, six color-coded categories, but it seems quite useful for several purposes. It is useful for popularization. The public knows a small amount about existential risks, but relatively little about their relative probability or importance. Several similar scales already exist, for example VEI (volcanic explosivity index), DEFCON (defense readiness condition), the Saffir–Simpson hurricane scale, etc, and they are quite useful in communicating with the public. The scale also seems useful for debate among experts. For example, Turchin & Denkenberger rate the risk from AI as red, the second highest level and the highest current risk, and the risk from particle colliders as green, the second to the lowest level and the lowest current risk. They base these classifications on a review of the literature. As an example of possible debate about this, I know something about both risks, and I could suggest that they lower the first and raise the second by a category or two, and I could write a few pages about why. The scale also seems useful as a way that governments, voters, and philanthropists might prioritize prevention activities.

The paper is also useful as a review of the literature. Turchin & Denkenberger discuss several aspects of risk importance. For example, they discuss probability, costs and benefits of remediation, limits in ability to estimate these, and extent of damage if the risk is actualized, from destruction of human civilization to destruction of all life in the universe and worse. They also discuss several existential risks in order to classify them. They cite 72 sources as part of this discussion. Carefully reading and thinking about their paper and following up by reading a reasonable number of their relevant sources is about equivalent to a graduate seminar in this area.

The paper is forthcoming in Futures, but it is available from Futures online now at https://doi.org/10.1016/j.futures.2018.01.003 . Its availability there can be free if your library or institution has access, but otherwise costs money. A preprint is available for free at https://philpapers.org/rec/TURGCA .

IEEE Standards for AI--An Opportunity for Public Comment

by James Blodgett

The IEEE (Institute of Electrical and Electronics Engineers) is the parent organization of the IEEE Standards Association, which develops global standards in a wide range of industries. They are currently developing something like standards for

autonomous and intelligent systems. What they are creating is more like philosophies and guidelines than standards. We are probably not yet ready to develop hard standards. Their current document is titled "Ethically Aligned Design, Version 2." This is available for download at https://ethicsinaction.ieee.org . They are inviting public comments by March 12, 2018 at 5P.M. (EST). Some of our SIG members might want to comment.

This document, 263 pages in length, was developed by committees of The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, committees which have several hundred participants.

The IEEE has several groups working on Ethically Aligned Design projects. These working groups are not limited to IEEE members, so we could get involved, albeit only various classes of IEEE members can vote on standard drafts. Also, there are classes of membership for which SIG members might be eligible. See https://ethicsinaction.ieee.org .

I am somewhat ambivalent about all of this. We can certainly use good thought in this area, and the IEEE material contains some good thought. It also contains many citations, which include important AI critics such as Bostrom and Yudkowsky. Most of its recommendations seem worthwhile. We should go in this direction. However, I am ambivalent about whether something as amorphous as a nice recommendation is adequate protection against the AI that Turchin sees as our biggest existential risk, albeit AI is also an existential opportunity and can be protective. Also, the IEEE consists of engineers who design AI systems, and therefore have somewhat of a conflict of interest. Balancing that conflict is one reason why SIG members might have something to contribute. I applaud IEEE's attempt to encourage diversity of participation and public comment, but IEEE members still determine final recommendations. Sometimes professionals transcend such conflicts of interest. However, even if their recommendations are exemplary, effectuality is still an issue. It seems difficult to assure that such recommendations will be followed by every single group and person that builds AI, or interacts with AI in a way that teaches something to any of its many versions. We may have to hope that exponential growth hits limits, so that AIs never acquire the power to turn the universe into paperclips, or that exponential growth is so unlimited that it gives unlimited AI the unlimited ethics to use that power for unlimited good, whatever that means. In a way the IEEE is similar to our SIG. Our SIG has done a few things that might matter, but we could still question our effectuality. I justify our efforts by the concept that we have been able to at least tweak the odds by a small amount, perhaps more, more likely toward the good because that is our aim, increasing probabilities in that direction because of that likelihood (Unless there are reasons why we don't see things correctly). Even a small tweak is worthwhile because of expected value, which equals probability times value, since the value in this equation is the future of our species. The same philosophy would seem to apply to the IEEE.