[Existential Risk / Opportunity] Singularity Management

January 2016

Contents:

Existential Risk, Willard Wells, and Me.

by James Blodgett, with comments [in red] by Willard Wells

This essay is, in part, a public airing of personal feelings. I usually write like a policy analyst. This time I have skin in the game. But I want that skin to transcend personal. I want to get to what the South Africans called "truth and reconciliation." When a wing falls off of an airplane while you are flying in it, the normal reaction is to scream in terror. However, if you are a right-stuff test pilot flying that plane, you had better "be afraid to panic" and "fly the plane."

> Think fast--
> spinning wildly--
> too fast to bail out-- use remaining wing to slow spin--
> NOW--hit eject button--BANG--I'm out!

[Like your stories/example, this one and the volcano.]

The subject of global risk can be even more terrifying than a wing falling off, albeit it doesn't hit one quite as fast. When it does bite, it is time to take a calming breath and to do your job as well as possible.

Willard Wells is a personal friend and colleague, and also an important thinker on existential risk. He is a physicist who has written two books that use a modification of Gott's formula to estimate the probability of civilization collapsing, and (as a separate calculation), of humanity going extinct. His results are like a wing falling off. The odds, as he calculates them, are not good. Willard is not a member of our SIG, but we did invite him to give a talk about his first book at the Mensa Annual Gathering at Reno in 2012. Both Willard and I are members of the Lifeboat Foundation. For a while I was chair of their Grantsmanship Committee, helping others apply for grants at other foundations. In that capacity I worked with Willard trying to develop a grant for another of Willard's interests, promotion of survival colonies as Noah's Arks to get some survivors past the collapse of civilization his equations predict and help avoid his other probabilistic prediction, human extinction. We didn't get a grant, but the Lifeboat Foundation does maintain a list of members interested in survival colonies. Just a few weeks ago, the Lifeboat Foundation published his second book, **Prospects for Human Survival**. Click here to order.

This essay started out as a review of Willard's second book, and as an interview in which I ask him questions about his method. Then a wing fell off. (Metaphorically, of course.) A year ago, I had addressed Willard directly in a post to a Lifeboat discussion group on Yahoo Groups. A copy of that post is appended, below. I said that Willard's work might justify taking extreme measures, for example, creation of a super artificial intelligence tasked to take over the world in a friendly way and protect us from ourselves. (I also asked if it is possible to take over the world in a friendly way.) I presented this as an extreme measure, a measure that is an existential risk in itself and that has other problems, but that might nevertheless be justified if it promised a solution to a greater risk. While reading Willard's book, I was surprised, pleased, and also appalled to see Willard quote that post to tout me as an advocate for creation of an AI overlord. Even the alternate term, AI nanny, would have sounded better, but, to be fair, I hadn't used that term. I hadn't called it an AI overlord, either.

Scholars get promoted when they have a high impact factor, measured by the number of citations of their work in scholarly journals. So I was partly pleased that Willard had quoted my work. However, global risk reduction deals with really big risks. One of the potential pitfalls is that attempting to chain the monster may backfire. This time, I may have woken the monster. It is time to take a breath and fly the plane.

To avoid waking monsters, I have generally focused, when considering global risk, on areas where the good direction seems fairly clear, not the case with AI. I am intrigued by AI, but also scared of it. It could be used in several ways to help solve existential risk, but it is an existential risk itself. I have not had to think much about an AI nanny since it is not even an option yet since the technology is not ready. I was appalled when hearing about the topic of AI nanny (before reading about it in Willard's book) because I realized

that an AI nanny could be thought of as being DESIGNED to take over the world, an activity I had thought reserved for mad scientists and James Bond villains. [Yes, if the good guys don't, the bad guys will.] "Taking over the world" is not mentioned explicitly, but something like that seems necessary as part of an AI nanny's assumed capabilities. To be fair, the idea apparently developed not out of ambition for world domination, but as a way to protect against it; thus the word "nanny". As part of the concept, an AI nanny is NOT an overlord, but is supposed to use its power only to protect us. [Oh? I think of a nanny as the children's overlord, and adult humans will be like children in the confusing fast-paced world of the near future.] The problem is that, in order to protect against world domination using this method, it is necessary to invoke some level of it, since an AI nanny must be powerful in order to protect us from other powerful entities. My post to Willard was about the only time I have addressed the AI nanny issue. I thought I made clear by my question in parenthesis in my post that there are problems with the idea. Willard used proper ellipsis to delete that question in his quote, since it detracted from what he saw as my main point. Now that Willard has called me out on the issue, I had better think about it.

An AI nanny might be a good idea--a risk, but perhaps worth taking--if the human species chose to go there by democratic/representative procedures, signed the appropriate treaties, gave it the appropriate approval, and spent billions doing it right. This seems unlikely to happen anytime soon. [Meanwhile, we're dead!] **[With some probability, if you are right. But you might be wrong. Meanwhile, I don't see any sure and safe way of making it happen anyway, although we might get there by something like what's described below.]** It just might happen in the future given big problems that motivate giving up some human control, and also given lots of confidence-building experience with AI. At best, an AI nanny might be like the US Supreme Court, making sure that a constitution developed by humans is implemented appropriately. Perhaps an AI nanny could be tested first as a form of arbitration chosen by both parties. However, there are big problems with the approach. In addition to human reluctance to let robots rule us even if they do so fairly and in a friendly manner, there is also great difficulty in assuring that result. Yudkowsky has written extensively on the difficulty of assuring AI friendliness, even assuming that they follow our instructions precisely. We will assume that the AI is smarter than us, since if not, a human council would seem more appropriate in the role of nanny. If the AI is smarter than us, it might use that smartness to start an exponential process involving design of enhancements and become even smarter than that, and if it acquires super intelligence, it might be able to use that to develop super physics that gives it other super powers, as we humans have done in comparison with animals. Given super powers, it becomes like a magic genie. Even if it follows instructions precisely, you had better state your wishes carefully. If told to make lots of paperclips, it might turn the universe, including us, into paperclips. [At this stage of its evolution, I think the nanny will have common sense + artificial emotions

akin to love for humanity.] An AI nanny would be tasked with protecting us from this, but it would also be a super intelligence and so might do the same thing itself.

Even if the AI nanny is friendly towards us and devoted to the task of protecting us from unfriendly AI, it is not clear that it could do that job. If it is appropriately constrained to operate using friendly and legal methods, this could be an impediment that would slow it down, and that would put it at a disadvantage in a conflict with a bad AI that is not so constrained. Even if our AI nanny discovers magic physics that gives it super powers, it might be possible for other AI (or even humans) to discover the same or other physics that enables other super powers. For example, we are very lucky that nuclear weapons are very difficult to construct. If someone (AI or human) discovers technology that does equivalent or greater damage but can be constructed in a garage workshop, it is difficult to see how that could be countered. It would not be trivial, and perhaps not possible, for our AI nanny to scan the world for threats (how exactly does the scanner work?), and find all of them, and counter all of them if found. The AI nanny could counter threats by aggressively suppressing research by AIs and by humans, but that does not sound like a friendly result.

Another potential failure mode for an AI nanny is that it might keep us from reaching our potential. This depends on how one defines "us" and what one considers important potential. For example, "us" might be defined as humans with present biology. Anything too far from current human design might not be good in several ways. On the other hand, perhaps we have the potential to develop better biology, even perhaps by such natural methods as sexual selection. Despite the hypothesis that nerds are unpopular, I suggest that one of the drivers of the evolution of human intelligence has been sexual selection. Perhaps, just by choosing interesting mates, we are fated to become much more intelligent that we are at present. Would we want an AI nanny to interfere with this process in order to maintain our current biology? Another way we might become better in some ways is to expand the definition of human to include AIs designed to be like humans We might be able to achieve more of the utilitarian objective of "the greatest good for the greatest number" by including in that "number" not only biological humans, but implementations of humans as AIs. "Human" AIs might be supportable using far fewer resources than biological humans, so that we could support many more of them and make "the greatest number" much larger. Of course, using resources for myriad "human" AIs could be a disaster, despite human AIs passing IQ and Turing tests, if their internal workings do not include human-like qualia, or, from a religious point of view, a human soul. A potential failure mode for us is that we might not make the right choice in this definition of what is human. If a so-called human AI does not have human consciousness, it does not seem to have the same moral status, and we might as well build a trillion refrigerators. A potential failure mode for an AI nanny is that it might not make the right choice either, especially if we rush to equip it with some definition of friendliness tied to the wrong definition of human. (Another potential failure mode is

that human AI could have more and better qualia that our biological version and we fail to implement it.) My personal reaction to this is to want to postpone a final definition of "human" until we have a lot more data. Actually developing an AI with artificial qualia will give us interesting data, albeit solipsists will never be able to totally validate that data. One of the reasons I am concerned about development of an AI nanny is that developing its objective function may force us to make this type of choices before we have thought enough. I am well aware of the contingency that we may have to satisfice before we know everything.

Willard makes the point that AI research is not illegal, and suggests that someone simply construct a super AI, tell it (very carefully, I hope) to be friendly, and unleash it on the world. However, the developer of an AI would find himself immersed in lawsuits if his AI started to infringe on other's rights. Worse, development of an AI specifically designed to take over the world might quite legitimately be taken as an act of war in some quarters. Even given an AI so designed, its success seems problematic unless carefully tested in lesser exploits. Even if it reports itself ready for the job, it may be overestimating its abilities as humans often do in such circumstances. If the project is secret the test exploits would have to be secret, limiting their scope. If the AI is secret, it also could not be vetted by other researchers, an important safeguard in normal science. If the AI succeeds in taking over the world, the lack of legal approval would then not slow it down because it would then be in the position to write its own laws, but this is not exactly a friendly result. If it is appropriately constrained to operate using friendly and legal methods, the slowing down effect would be extreme if the AI did not have the world approval that would facilitate those laws being adjusted to accommodate its mission. For example, if scanning the world for threats involves swarms of nanobots photographing everything, that would seem to violate privacy laws. (Also, could nanobots get into a closed and weather-stripped garage?) [Yes, in semitropical climates bugs get into everything.] **[Yes, but this example is just a placeholder for many ways bad guys and bad AIs might find to avoid surveillance. A super AI nanny might counter those ways, but that assumes that a counter is possible, and that seems to require that a perfect method of surveillance is physically possible. It may not be. For example, the CIA etc, with considerable resources, took a long time to find Bin Laden.]** Even given that the AI is successful and on the job of protecting us from unfriendly AI, this does not assure that it will win against an unfriendly AI that stumbles across magic technology that our AI has not achieved or against which our AI cannot protect.

An AI nanny might be worth trying in extremis as a last desperate measure. Willard's results suggest extremis. Can we rely on Willard's results to make existential judgments?

I have an MS in statistics, so I can check some of his work. I am not enough of an expert to judge the validity of some of his engineering simplifications or multidimensional numerical integrations, and I have not had time to check everything I could check, but I understand his basic approach. Gott's estimator calculates the probability distribution for survival time of an entity based on the time it has been in existence. Given an entity like a business firm or a species, i.e. an entity that does not grow old and die with a more-or-less fixed lifespan, those that have been around for a while have demonstrated robustness and so are more likely to stay around for a while longer. Assuming no reason to think otherwise, the probability that they are being observed in the first one percent of their lifetime, or in the last one percent of their lifetime, is in both cases one percent. Based on this type of math, we can calculate a complete probability distribution for survival at each time in the future. Willard and others have applied this to data on business firm survival with good results. The human species has a long history, so a naive interpretation is that we can hope for a long future. However, that depends on what one defines as human. Did Neanderthal count? If the relevant humans are those who have developed technology that enables human extinction, then Gott's formula gives a pessimistic result.

Willard's modification is to adjust the time parameter to control for exposure to technological hazards as technology increases. As he says, the lifetime of a clay pigeon in a shooting gallery is based not on its time hanging on the wall, but on the number of shots being fired. Before we developed technology that could cause our own extinction, our exposure to that kind of risk was zero, so our long prehistory is not protective. However, Willard's exposure parameter makes the analysis no longer completely top down, since it builds in estimates of the potency of future technology. Such estimates are a matter of judgment. One line of thought is that technological potency is expanding exponentially, as we have seen since the industrial revolution and as the singularity people expect. However, perhaps they are wrong. In nature, exponential growth often hits limits. Perhaps we have already discovered most useful inventions. Perhaps future science will not be as dangerously potent as we worry that it might be. Some people think they see declining marginal returns to science and technology. As an example, Edison invented the phonograph in a flash of insight. In order to develop the light bulb, he had to make another invention, the industrial research lab. He hired hundreds of assistants and technicians to help him test what is sometimes cited as 3,000 filaments, and also to design the required infrastructure. The integrated circuit, with very roughly an impact on our lives in the range of the light bulb, required many workers and many labs to reach its present state of development. These are rough points on a declining curve of technological effectiveness versus developmental effort. So exponential increase may be the wrong estimate. This is why I gave it small statistical weight. On the basis you your review and another, I'll probably make it even smaller in the next edition.

Willard sees one component of exposure as growing exponentially. The exponential part of his probability calculation starts small, but grows so fast that it overwhelms the rest. However, he hasn't adjusted for the possibility that exponential growth will not continue. The contribution of the exponential component of probability to overall probability cannot exceed the probability that the exponential component actually actualizes. This criticism is not devastating for Willard's results because even without the exponential component, his results are still alarming. However, even the construction of the non exponential component is an attempt to estimate the potency of future technology, and despite Willard's efforts, an estimate of that potency is not totally a top-down estimate mathematically derived from first principles. Willard gives a good picture of his sensible judgment calls in selecting data sets on which to base estimates, sensible but not infallible. It would help to do a sensitivity analysis, in which variables are estimated at the top and bottom of their plausible ranges to see how much difference that might make. An extended sensitivity analysis would consist of many analyses based on many different plausible models.

On one level I think Willard is 100% right. We live in an interesting time, interestingly dangerous. People and decision makers have some sense of that, but they are not adequately aware. It would help to get Willard's results onto the intellectual agenda. It would help to test and polish those results. However, I am reluctant to take Willard's results as a justification for extreme action before there is no other choice, since Willard's models and judgment calls have some potential to be wrong. For me the value of Willard's results is not that they prove that the end is near, but that they remind us that that is plausible. That reminder can be used to motivate activities that would be protective against some existential risks but that are also useful for other good reasons, activities like massive space industrialization as pictured plausibly by O'Neill and Metzger. Should it also be used to motivate last ditch efforts? Perhaps, but this is a beautiful time in a prospering world. Most recent prophecies of doom have turned out to be wrong. I am not yet ready to push a button that implements a desperate risk in order to get out of our present circumstances. [Fair enough.] In a sense fortunately, we don't yet have to consider implementing the AI nanny version of that desperate risk because there is no button to push. The technology to make a super AI is of concern because it might be stumbled upon, but it is not yet ready for the implementation of an AI nanny, so there is no button to push that would reliably invoke that technology. That and other buttons may become available in the future, so it is important to vet our philosophy about such things. Meanwhile there is other technology that would help. Willard is right that diverse survival habitats, on Earth and in space, would increase our odds of survival.

Interview with Willard Wells

Interview of Willard Wells by James Blodgett

Willard, what is your take on all of this?

First, when the time does come to push the button, I think it will be very urgent. Spending time by getting approval through democratic processes would most likely be deadly.

Second, I think geniuses that develop superhuman AI will be somewhat contemptuous of authority and not very respectful of slow democratic political processes, maybe a bit arrogant. This personality seems to go with the territory. Like Steve Jobs, perhaps. Apple was caught practicing aggressive tax evasion. When Jobs got cancer, he thought he could outwit his physicians, and then he died as a result.

I described your method very briefly. Did I do justice to it? Is there some part you would like to describe in more detail? (But still briefly, we can't duplicate your book here.)

Yes, you did justice to it.

I used to see collapse of civilization as possibly leading to human extinction because a collapse of population could put us metaphorically "on the endangered species list" and make us less robust to other troubles. However, I now agree with your idea that collapse of civilization could be protective at least in regard to extinction because it would eliminate most of the technology that creates most of the existential risk. How do you see the interplay between these two principles? When is each most likely to be active?

Humans from the Inuit in the arctic to the Yanomami in tropical jungle have inhabited a great many different niches, many of them uncivilized, so I don't see collapse of civilization as a cause of extinction. Of course it would take these primitive peoples centuries to move into the fully civilized niche, but they would find vast new territory covered with exotic artifacts that would stimulate their curiosity and speed the process.

I question the value of initiating some minor collapse of some aspect of civilization as vaccination against a larger collapse, as you seem to suggest. Really? That would cause lots of trouble, and I don't think we can reliably predict the result. For example, some nut causing minor disaster and using this as a justification might immunize the world against our attempts at global risk reduction, since we can sound like similar nuts.

As I am learning from the AI overlord issue, it is not only in the case of a super AI genie that we had better be careful with our words.

Do you mean like wiping out the Internet for a month or so? Yes, that would cause lots of trouble. However, I continue to believe that we need a small minority of hacker types just to test our robustness. Otherwise, we get complacent and think our civilization is robust when in fact it becomes very fragile from numerous interconnections. For example, the Internet tells us which brands of any commodity are best, and so most brands go out of business, and we lose their potential for innovation and for backup in case of emergency.

I had a bit of experience with Gott's method before seeing your use of it. I like it, but I found it fairly sensitive to assumptions. You work is an example. Your second book refines the assumptions of the first, and comes up with a somewhat different result. Gott used different assumptions and qot a quite different result, albeit you had good reasons to question Gott's assumptions. You talk of the distribution we could make of survival times if we knew the history of all human-like species in the galaxy. What do you think would be the variability of a similar distribution: the distribution of survival estimates if we had a million people like you making similar estimates drawing from the metaphorical galaxy of reasonable assumptions?

As a wild guess, a factor of 3. Of course with a million estimators, there would be thousands of outliers! I'm already thinking of a pdf that I wish I had used in the book. However, when I go back in my calculation mode, I may rediscover why I didn't use it.

Blodgett's Post On Lifeboat Foundation Discussion Section On Yahoo Groups, Quoted (In Part) In Wells' Book

by James Blodgett

Mon Sep 8, 2014 11:04 am (PDT)
Posted by: James R. Blodgett
Re: Wither civilization

Hello Willard. I am glad I mentioned your name. You had told me about your new result, a higher risk to civilization, but I didn't understand its derivation. It is an important result because of what should be a basic principle of decision theory—should be and probably is, but I made up this version. My rhetoric would be improved if I knew the terminology of the standard version, so someone clue me in if you recognize it.

My principle is the relationship between whether we should take a risk, and the background level of risk. It is clear from a little parable. The parable posits a sleepy island village with a small airfield, serviced by a weekly passenger flight. The last plane blew one of several tires on landing, so there is 1/100 chance the entire landing gear will blow on its next use. This is much too high odds for passenger flight, so the plane is parked on the tarmac awaiting a replacement tire. Meanwhile the island volcano erupts, and a flood of lava is headed toward the village. So everyone jumps on the plane and takes off.

If Willard is right, it turns the precautionary principle on its head. It may be a reason to justify risky measures that have some probability of solving the problem. For example, it might be a reason to push forward on AI research despite incomplete safety measures, and hope to develop a friendly version and hope that it can develop magic super science that allows it to take over the world in a friendly way and save us from ourselves. (Is it possible to take over the world in a friendly way?) Space tech might save us via dispersion but has similar failure modes that might similarly be worth ignoring if the background risk is high.

Solutions cost money as well as risk. Willard's results might justify spending more money on possible solutions.

If Willard is right, his results should be on the intellectual agenda. Also, we should vet his results to see if there is some way he might be wrong.

Proposed Annual Gathering Presentation by David L. Minger

David Minger, a member of our Special Interest Group, proposes to give a talk on existential threats to humans at the 2016 Mensa Annual Gathering in San Diego.