Subject: **Elites and AI**

From: **Luke Muehlhauser** <luke@intelligence.org>
Date: Mon, Jul 8, 2013 at 6:40 PM
To: Jonah Sinick <jsinick@gmail.com>

Hi Jonah,

This is the email thread — with the subject line "Elites and AI" — that I hope to publish in full with the article this email thread will be used to produce.

As we've discussed, I want to further investigate the question I raised in Will the world's elites navigate the creation of AI just fine? and Elites and AI: Stated Opinions.

I'd like to continue the investigation by looking at somewhat-analogous historical cases. As we discussed in person, for now let's focus on historical cases that are analogous on several of the following dimensions:

1. AI may become a major threat in a somewhat unpredictable time.
2. AI may become a threat when the world has very limited experience with it.
3. A good outcome with AI may require solving a difficult global coordination problem.
4. Preparing for the AI threat adequately may require lots of careful work in advance.
5. Elites have strong personal incentives to solve the AI problem.
6. A bad outcome with AI would be a global disaster, a good outcome with AI would have global humanitarian benefit.

Here are some (relatively recent) historical cases to consider in the context of the reference classes suggested by the above list:

1. 2008 financial crisis.
2. Climate change.
3. Iraq War.
4. Leaders being deposed or assassinated.
5. Eradication of smallpox.
6. Nuclear proliferation.
7. Recombinant DNA.

8. Nanotech.
9. Near earth objects.
10. Chloroflourocarbons.
11. Risks to critical infrastructure from solar flares.
12. Cyberterrorism.
13. Swine flu.

Luke Muehlhauser
Executive Director
MIRI

From: **Jonah Sinick** <jsinick@gmail.com>
Date: Tue, Jul 9, 2013 at 1:12 PM
To: Luke Muehlhauser <luke@intelligence.org>

Hi Luke,

Regarding

I'd like to continue the investigation by looking at somewhat-analogous historical cases. As we discussed in person, for now let's focus on historical cases that are analogous on several of the following dimensions:

After the conversation, we decided that what we're trying to do is more subtle than this:

- Some of the criteria, (when taken in isolation), are *necessary* conditions for inclusion of historical cases where the world's elites *did* solve a problem.
- Some of the criteria, (when taken in isolation), are *sufficient* conditions for the inclusion of historical cases where the the world's elites *did not* solve a problem.
- Some of the criteria, (when taken in isolation), are *necessary* conditions for inclusion of historical cases where the world's elites *did not* solve a problem.
- Some of the criteria, (when taken in isolation), are *sufficient* conditions for inclusion of historical cases where the world's elites *did not* solve a problem.

The criteria that are necessary are not necessarily sufficient, and vice versa.

Addressing the criteria in turn:

#1 AI may become a major threat in a somewhat unpredictable time.

We *don't* want to count instances where the world's elites *did* successfully address a major threat that *did* arise at a predictable time as evidence *in favor* of the world's elites successfully navigating the creation of AI, based on this criterion alone.

But we *do* want to consider instances where the world's elites *did not* successfully address a major threat *despite* it having arisen at a predictable time as evidence *against* the world's elites successfully navigating the creation of AI.

#2 AI may become a threat when the world has very limited experience with it.

We *don't* want to count instances where the world's elites *did* successfully address a threat that the world *did* have experience with as evidence *in favor* of the world's elites successfully navigating the creation of AI, based on this criterion alone.

But we *do* want to consider instances where the world's elites *did not* successfully address a threat *despite* the world having experience with it as evidence *against* the world's elites successfully navigating the creation of AI.

#3 A good outcome with AI may require solving a difficult global coordination problem.

We *don't* want to count instances where the world's elites *did* successfully address a threat that *did not* require solving a difficult global coordination problem to address as evidence *in favor* of the world's elites successfully navigating the creation of AI, based on this criterion alone.

But we *do* want to consider instances where the world's elites *did not* successfully address a threat *despite* it not requiring solving a difficult coordination problem to address as evidence *against* the world's elites successfully navigating the creation of AI.

#4 Preparing for the AI threat adequately may require lots of careful work in advance.

We *don't* want to count instances where the world's elites *did* successfully address a threat that *did not* require lots of careful work in advance to address as evidence *in favor* of the world's elites successfully navigating the creation of AI, based on this criterion alone.

But we *do* want to consider instances where the world's elites *did not* successfully address a threat *despite* doing so not requiring a lot of careful work in advance as evidence *against* the world's elites successfully navigating the creation of AI.

#5 Elites have strong personal incentives to solve the AI problem.

We *don't* want to count instances where the world's elites *did not* successfully solve a problem that the world's elites *did not* have strong personal incentives to solve as evidence *against* the world's elites successfully navigating the creation of AI, based on this criterion alone.

But we *do* want to consider instances where the world's elites *did* successfully solve a problem *despite* not having strong personal incentives to solve it as evidence *in favor* of the world's elites successfully navigating the creation of AI.

#6 A bad outcome with AI would be a global disaster, a good outcome with AI would have global humanitarian benefit.

We *don't* want to count instances where the world's elites *did not* successfully address a threat, when the threat would *not* be a global disaster as evidence *against* the world's elites successfully navigating the creation of AI, based on this criterion alone.

But we *do* want to count instances where the world's elites *did* successfully address a threat, when the threat would *not* be a global disaster, as evidence *in favor* of the world's elites successfully navigating the creation of AI.

From: **Jonah Sinick** <jsinick@gmail.com>
Date: Wed, Jul 10, 2013 at 6:28 PM
To: Luke Muehlhauser <luke@intelligence.org>

I spent 3 hours reading about smallpox eradication.

How smallpox eradication does or doesn't fit the criteria

- **#1** Smallpox didn't arrive at an unpredictable time. On the contrary, it had already arrived before the eradication campaign.

- **#2** The world didn't have experience eradicating a disease before smallpox was eradicated, but a number of nations had eliminated smallpox.

- **#3** Smallpox eradication required solving a difficult global coordination problem, but in a way disanalogous to AI safety (see below).

- **#4** Preparing for smallpox eradication required effort in advance in some sense, but the effort had mostly already been exerted before the campaign was announced.

- **#5** Nations without smallpox had incentive to eradicate smallpox so that they didn't have to spend money to immunize citizens so that the virus would not be (re)-introduced to their countries. For example, in 1968, the United States spent about $100 million on routine smallpox vaccinations.

- **#6** Smallpox can be thought of as a global disaster: in 1966 about 2 million people died of smallpox each year.

Main takeways

My impression is that the factors that enabled smallpox eradication are:

- Eradication efforts seem to have been in basically everybody's interest.

  If I read the numbers right, the United States spent 1/3rd as much money *per year* on vaccinating US citizens to prevent the reintroduction of smallpox as *the total cost of the entire eradication campaign*.

  Smallpox caused death and suffering, constituted a burden on health care systems, and reduced human productivity.

- A few technological innovations (the freeze-dried vaccine and the bifurcated needle) that had been developed without a specific view toward eradication.

- The disease transmission being was easy to disrupt

- Actors had a lot of experience with elimination at the national level.

I don't think that the successful eradication of smallpox stands out as being especially relevant to the question of whether the world's elites will deal with AI well.

Some more detailed notes below.

Notes on smallpox eradication

These mostly from a document published by the Center for Global Development:

- The first vaccine was created in 1798, an improved vaccine was created in the 1920's, and vaccines that didn't require cold storage were developed in the 1950's.

- Smallpox was unusually well suited to eradication:

  (i) It wasn't transmitted by insects or animals
  (ii) It was straightforward to diagnose
  (iii) There was a long time lag between getting infected and becoming infectious
  (iv) The disease was sufficiently debilitating so that infectious people had relatively little contact with others
  (v) The vaccine didn't need to be refrigerated
  (vi) Vaccination prevented infection for 10+ years.

- As late as 1966, there were 10-15 million cases of smallpox a year, and 1.5-2 million people died per year.

- The campaign started in 1959.

  The World Health Organization (WHO) didn't make it a high priority.

  There was initially insufficient financial support. The program had trouble establishing a proven track record to bolster support, because case reporting was so low that it was unclear where smallpox prevalence was dropping. The beginning of the campaign coincided with the failure of the malaria eradication campaign, and potential funders had an unfavorable impression of eradication campaigns.

- In 1964, the WHO set up a Smallpox Eradication Unit , with its own staff and budget, and made smallpox eradication one of its major objectives. The United States began providing more support, and this was a key factor in the program's development.

  The WHO began supplying vaccine injectors, specimen collecting kits and training aids to countries that requested them.

  The eradication campaign began using the bifurcated needle, which had been developed in 1961. This needle was very cheap, reusable, and easy to use.

- Between 1967-1973, progress was very rapid: the number of endemic countries dropped from 31 to 5.

- There were some issues of countries sliding back into endemicity, but they were quickly resolved by strengthened efforts.

- Toward the end of the campaign, the focus shifted from national vaccination efforts to actively seeking out cases and containing outbreaks. 10000's of health workers worked in Ethiopia (which had a civil war) to stop smallpox transmission.

- The annual cost of smallpox damage and prevention was about $1.35 billion, and the total cost of the eradication effort was about $300 million.

- An positive unanticipated consequence of the eradication campaign was mainstreaming of routine vaccination in the developed world. Between the time the campaign started and 1990, routine vaccination in the developing world increased from 5% to 80%.

From: **Luke Muehlhauser** <lukeprog@gmail.com>
Date: Wed, Jul 10, 2013 at 6:38 PM
To: Jonah Sinick <jsinick@gmail.com>

A few numbers stand out: How could the entire eradication effort cost $300 million, if the Ethopian effort alone required tens of thousands of health workers, and if the US vaccinations by themselves sometimes cost about $100 million annually?

Luke

From: **Jonah Sinick** <jsinick@gmail.com>
Date: Wed, Jul 10, 2013 at 8:14 PM
To: Luke Muehlhauser <lukeprog@gmail.com>

On the first point: the CDC document says vaccination cost $0.10 per person in endemic areas (in ~1960) and cost $6.50 per person in the United States (in 1968). I don't know why the differential in cost is so high. When I first read it, I assumed that it's because of the differential in cost of labor, but thinking it over again, I find it hard to imagine how that could give rise to such a big differential in cost. I can investigate further if you'd like.

On the second point: the figures are unadjusted for inflation, and $100 in 1973 was worth about $500 today. I believe that people in Ethiopia were living on an amount on the order of $100/year. This is an underestimate for the cost of a health worker, but using it as an input into an order of magnitude calculation gives

(10k people)*(annual wage) = 1 million dollars

From: **Luke Muehlhauser** <lukeprog@gmail.com>
Date: Wed, Jul 10, 2013 at 8:16 PM

To: Jonah Sinick <jsinick@gmail.com>

Okay thanks. I think this is probably deep enough into smallpox for our purposes.

Luke

From: **Jonah Sinick** <jsinick@gmail.com>
Date: Fri, Jul 19, 2013 at 7:22 PM
To: Luke Muehlhauser <lukeprog@gmail.com>

Hi Luke,

Responding on the point of risks to critical infrastructure from solar flares:

I found a document on the OECD risk management website about geomagnetic storms.

I think that the negative expected value coming from the risk is sufficiently small so that it shouldn't be thought of as being a potential global disaster.

There are three historical reference points: a 410 nTs/min geomagnetic storm from 2003, a 640 nTs/min storm in 1989, and a 1760 nTs/min storm from 1859 (pg. 9).

The theoretically derived frequencies of storms of these magnitudes are ~ 1/10 years, ~ 1/50 years and ~1/100k years respectively (pg 19).

The estimated costs of storms storms of the latter two magnitudes are ~$11 billion and ~$3 trillion (pg. 13).

The expected losses coming from storms of the latter two magnitudes are ~$200 million/year and ~ $30 million/year respectively.

Even if one is suspicious of the 1/100k years frequency for the most severe ones and uses a ~1/1k year frequency instead, one only gets negative expected value of $3 billion/year, which is large enough so that it should be addressed, but still a small perturbation on the economy as a whole.

I think that the problem is too small to be in the same reference class as AI risk. As for how well it's being addressed, the report discusses measures that are in place and room for improvement on pages 40 - 47, and the picture is mixed.

Jonah

From: **Jonah Sinick** <jsinick@gmail.com>
Date: Mon, Jul 22, 2013 at 3:08 PM
To: Luke Muehlhauser <lukeprog@gmail.com>

Responding on climate change:

The book *The Discovery of Global Warming* has an associated website here. It offers a summary of the history of climate change science here.

Some points:

- People started to see climate change as a potential problem in the early 1970s. However, there was ambiguity as to whether human activity was systematically causing warming (because of carbon emissions) or cooling (because of smog particles).

- Some scientists thought that there was systematic anthropogenic warming in the late 1970s, but they had relatively little visibility.

- The first IPCC was published in 1990, and stated that there's substantial anthropogenic global warming coming from greenhouse gases.

- In the coming years there was a lot of research, and by 2001 there was a very strong consensus.

One would have to deep dive

(i) How strong the consensus was as a function of time during the 1990s.
(ii) The history of views on the size and sign of the humanitarian impact of climate change.

to develop a clear sense for how quick society has been to address expected damage from climate change.

For the purposes of the present project, I think that what would be most interesting is to investigate the history of discovery and response to the *tail risk*. My intuition is that rational historical estimates for median case negative humanitarian value are too low for median case global warming risk to be in the same reference class about AI risk, but that the negative expected value coming from the tail might be large enough to put it in the same reference class.

So to start, I'm going to look at Posner's book (which has other relevant information). After doing so, I might return to the expected value other than that coming from the extreme tail.

From: **Luke Muehlhauser** <lukeprog@gmail.com>
Date: Mon, Jul 22, 2013 at 3:29 PM
To: Jonah Sinick <jsinick@gmail.com>

Ok. Please also send your thoughts on the negative EV of geomagnetic storms to some of the people who think it's a really big deal, to see whether they have a rebuttal.

From: **Jonah Sinick** <jsinick@gmail.com>
Date: Tue, Jul 23, 2013 at 12:29 PM
To: Luke Muehlhauser <lukeprog@gmail.com>

Responding on the point of cyberwarfare:

- Last year, the Department of Defense published a report titled Resilient Military Systems and the Advanced Cyber Threat which says

  The Task Force believes that the integrated impact of a cyber attack has the potential of existential consequence. While the manifestation of a nuclear and cyber attack are very different, in the end, the existential impact to the United States is the same.

- In 2011, Admiral Michael McMullen (who was the highest ranking US military officer at the time) said:

  The single biggest existential threat that's out there, I think, is cyber…Cyber actually, more than theoretically, can attack our infrastructure, our financial systems…There are countries who are very good at it. More than anything else that is the long-term threat that really keeps me awake.

- In Chapter 4 of Cyber War: The Next Threat to National Security and What to Do About It, former National Coordinator for Security, Infrastructure Protection, and Counter-terrorism Richard Clarke wrote:

  Why had Clinton, Bush, and then Obama failed to deal successfully with the problem posed by America's private-sector vulnerability to cyber war?

  apparently suggesting that the problem is neglected.

- Some people have suggested that the military and its contractors are motivated to overstate the risks from cyberwarfare in order to justify large cybersecurity budgets, for example, here.

- Jason Healey (who's the director of an initiative at a national security think tank) wrote an article in US News responding to the Department of Defense report and McMullen, saying that the risk has been overstated, and is far lower than that of nuclear war, but that he anticipate that America's electric grid will be integrated with the internet in the future, and that this could make the risk of cyber attacks much worse.

- The Organisation for Economic Co-operation and Development (OECD) published a report saying

  The authors have concluded that very few single cyber-related events have the capacity to cause a global shock

  Catastrophic single cyber-related events could include: successful attack on one of the underlying technical protocols upon which the Internet depends, such as the Border Gateway Protocol which determines routing between Internet Service Providers and a very large-scale solar flare which physically destroys key communications components such as satellites, cellular base stations and switches.

  For the remainder of likely breaches of cybsersecurity such as malware, distributed denial of service, espionage, and the actions of criminals, recreational hackers and hacktivists, most events will be both relatively localised and short-term in impact.

- The OECD report also briefly mentions **electromagnetic pulses** as a cybersecurity risk:

An electro-magnetic pulse (EMP) is a burst of high-energy radiation sufficiently strong to create a powerful voltage surge that would destroy significant number of computer chips, rendering the machines dependent on them useless. It is one of the few forms of remote cyber attack that causes direct permanent damage. The best-known trigger for EMP is with a high-latitude nuclear explosion and was first noticed in detail in 1962 during the Starfish Prime nuclear tests in the Pacific. Studies have investigated the possible effects on the United States power grid. (Oak Ridge National Laboratory, 2010).

I've come across the claim that a nuclear weapon detonated at high altitude over America could create an EMP.

I'll investigate the issue of nuclear weapons causing EMPs separately.

I'm not sure where to proceed with the investigation of cyberwarfare: different people have very different accounts of how big a threat it is. Maybe I should try contacting people who have subject matter knowledge?

Best,
Jonah

From: **Luke Muehlhauser** <lukeprog@gmail.com>
Date: Tue, Jul 23, 2013 at 1:01 PM
To: Jonah Sinick <jsinick@gmail.com>

Contacting people re: cyber-risks sounds good.

From: **Jonah Sinick** <jsinick@gmail.com>
Date: Thu, Jul 25, 2013 at 3:28 PM
To: Luke Muehlhauser <lukeprog@gmail.com>

Hi Luke,

Responding on the point of climate change:

Broadly, I feel comfortable with characterizing climate change mitigation as disanalogous to the AI risk mitigation situation (with respect to how well elites will address / are addressing the risk), such that shortcomings in society's response to climate change should be given very little weight in thinking about whether elites will successfully address AI risk.

Some points about climate change:

- The harms are expected to fall disproportionately on people in poor countries. So developed world elites have lower personal incentives to solve the climate change problem than they do to solve the AI risk problem.

- As I wrote about in Potential Impacts of Climate Change, the median case humanitarian loss seems to be in the neighborhood of 20% GDP per year for the poorest people. In light of anticipated substantial economic development and marginal diminishing utility, this is a *much* smaller negative humanitarian impact than AI risk (even ignoring future generations).

Economist Indur Goklany, wrote a paper titled Is Climate Change the Number One Threat to Humanity? in which he said

Through 2085, only 13% of mortality from hunger, malaria and extreme weather events (including coastal flooding from sea level rise) should be from warming.

- Even if one looks at the *mean* negative impact of climate change rather than the *median* negative impact of climate change (i.e. taking into account tail risk aside from the really extreme negative tail), the situation doesn't look *that* bad.

  Evan Soltas of the Washington Post and Bloomberg View wrote an article arguing that the bulk of the negative expected value of climate change comes from the tail. He considers several models for the probability distribution of temperature rise, as well as several models for how the cost of temperature rise varies as a function of temperature, and uses these to compute "risk-weighted cost functions." He concludes that the mean negative impact is in the vicinity of the negative impact of a 7 degree C temperature increase. (He doesn't specify the time scale.)

  For context, the current global temperature is 15 degrees C (~59 degrees F), so that a 7 degree C increase would move it up to 22 degrees C (~ 71.5 degrees F). This might give rise to a large loss of biodiversity, but would presumably leave a great deal of inhabitable land: again, the problem looks to be much smaller than the problem posed by AI risk.

  In a famous 2009 paper, Harvard economist Martin Weitzman assigns a 5% chance of 10 degree C warming by 2200, and a 1% chance of 20 degree C warming by 2200. The latter increase could conceivably catastrophe of size similar to AI risk (if one ignores future generations). But:

  (a) The probability is small
  (b) The temporal distance is a lot larger than the temporal distance between now and when AI will probably be created, and even if elites don't act to mitigate events that are 190 years away, they make act to mitigate events that are 40 years away.
  (c) By the time 190 years have lapsed, all sorts of things will have changed (e.g. AI will likely have been created), and the case for planning for events so far out in the future is weak.

  It would be nice to know the 5% and 1% level temperature rises that Weitzman would give for 2050 and 2100, but I wasn't able to easily find information about this.

- The risk of human extinction seems very low.

  According to the 5th IPCC:

  Some thresholds that all would consider dangerous have no support in the literature as having a non-negligible chance of occurring. For instance, a “runaway greenhouse effect”—analogous to Venus-- appears to have virtually no chance of being induced by anthropogenic activities.

  One person who differs is Columbia professor James Hansen. In a slideshow titled "Threat to the Planet Implications for Energy Policy" he wrote

  The Venus syndrome is the greatest threat to the planet, to humanity’s continued existence. [...] In my opinion, if we burn all the coal, there is a good chance that we will initiate the

> *runaway greenhouse effect. If we also burn the tar sands and tar shale (a.k.a. oil shale), I think it is a dead certainty.*
>
> But he's the only credentialed scientist who I was able to find who makes this claim. I also searched for climate change and existential threat, and didn't come across any other claims that there's a possibility of human extinction.
>
> So that's another way in which the situation is disanalogous (although I recognize that people may not be sensitive to the difference between 50% of the population killed and 100% of the population killed).

Putting this all together, I think that expected negative humanitarian impact of climate change (especially for decision makers) is sufficiently small so that one shouldn't give large weight to whether or not elites have successfully addressed climate change in determining whether or not elites will successfully address AI risk.

From: **Luke Muehlhauser** <lukeprog@gmail.com>
Date: Fri, Jul 26, 2013 at 9:57 PM
To: Jonah Sinick <jsinick@gmail.com>

Jonah,

Thanks for your climate change analysis. I'd like to spend a bit more time on it. In particular, I'd like to look a bit harder for people who might disagree with your analysis.

Bjørn Lomborg has a big-picture view about climate change that is similar to yours, though the details are different. He once summed up his view this way: "Global warming is real – it is man-made and it is an important problem. But it is not the end of the world." But I know many climate scientists have criticized Lomborg extensively — perhaps you can find some who have criticized his big picture conclusion, and not just particular details of his work?

You could also send your analysis to 15+ of leading researchers on the big picture of climate change and see if you get a response from any of them.

Luke

From: **Jonah Sinick** <jsinick@gmail.com>
Date: Wed, Jul 31, 2013 at 6:34 PM
To: Luke Muehlhauser <lukeprog@gmail.com>

Hi Luke,

I did some more research on climate change.

Critics of Lomborg

You asked whether the climate scientists who criticized Lomborg have criticized his big picture conclusion. There are climate scientists and economists who have criticized his prescriptions for economic policy, but I wasn't able to find claims in the direction of "climate change is the end of the world" among the criticisms that I read.

(A) In Jerry Mahlman's 2001 article Global Warming: Misuse of Data and Ignorance of Science: A review of the “global warming” chapter of Bjørn Lomborg’s The Skeptical Environmentalist: Measuring the Real State of the World, Mahlman wrote:

> I found Lomborg to be reasonable in his noting of tendencies by some to exaggerate both the likelihood and dire consequences of an increased frequency of storms, hurricanes, and other extreme weather events due to global warming. Currently, we have little evidence for that. However, we cannot rule out such. [...] I was uncomfortable with Lomborg’s views on the economics of future climate change costs and how these costs are to be paid. I was also uncomfortable with his assertion that “the cost of global warming is $5 trillion.” Clearly, that number should, by his own arguments, be presented with a broad range of uncertainty and with considerable humility. Again, where is the statistician?

This remark seems to be of the type "One shouldn't restrict attention to the median case: one also needs to consider the tail risk." Mahlman has passed away since writing his review, so I can't contact him to confirm.

(B) In his 2002 article Global Warming: Neglecting the Complexities, climatologist Stephen Schneider wrote:

> Then again, Lomborg cites only one value for climate damages—$5 trillion—even though the same economics papers he refers to for costs of climate change policy generally acknowledge that climate damages can vary from benefits up to catastrophic losses. It is precisely because the responsible scientific cannot rule out such catastrophic outcomes at a high level of confidence that climate change mitigation policies are seriously proposed.

This remark seems to be of the same type as Mahlman's. As with Mahlman, Schneider has passed away, and so I can't contact him to confirm.

(D) In Hot, it’s not: Reflections on Cool It, by Bjorn Lomborg, environmental economist Frank Ackerman wrote

> Climate outcomes are uncertain in several respects, ranging from short-term variations in weather to the long-term sensitivity of the climate to greenhouse gas concentrations, to the probability of irreversible catastrophes. Lomborg’s embrace of the A1B scenario as “the standard” suggests an all too common strategy for economic analysis: adopt a best guess or expected value as the point estimate, and ignore the question of uncertainty. The Stern Review, in contrast, highlights the role of uncertainty, applying a Monte Carlo analysis in which dozens of parameters are allowed to vary around the current best estimates. This expanded treatment of uncertainty is a principal reason why Stern’s estimates of the social cost of carbon are higher than those of many other economists. Since people are risk-averse, including both better-than-average and worse-than-average possibilities in the analysis makes climate change look more threatening: the better possibilities have limited effect, while the worse ones loom large.

and goes on to write about Weitzman's 2007 paper, which argues that the tail risk is worse than the Monte Carlo analysis suggests. This remark seems to be of the same type as Mahlman's and Schneider's remarks. If you'd like, I can write to Ackerman for clarification/confirmation.

(D) In his 2008 review of Lomborg's book Cool It: Cool It: The Skeptical Environmentalist’s Guide to Global Warming, IPCC lead author Brian O'Neill wrote

> It is undoubtedly true that in some cases the climate change issue is being painted in the starkest of doomsday colors, at odds with the scientific literature. Lomborg is right to call people to task when they speak of impending catastrophe or of the very existence of civilization being at risk. You won’t find support for such assertions in the reports of the Intergovernmental Panel on Climate Change (IPCC), the international scientific body that produces authoritative assessments of the science every five years or so. I, like many others working in this field, am increasingly uncomfortable with the growing use in public discourse of catastrophic language and of phrases like “tipping points” to describe possible changes in the climate system. They may be technically correct descriptions, in an academic sense, of possible nonlinear or irreversible changes, but they are too imprecise for the public conversation.

Thus, Brian O'Neill appears to agree with Lomborg's general view on the significance of climate change relative to other problems. I wrote to Brian O'Neill asking for more detail.

Papers on worst case scenarios

I spent time looking for papers about climate change impacts which are most pessimistic.

(A)

In his 2009 article The worst-case scenario Stephen Schneider wrote

> An atmosphere in 2100 with 1,000 parts per million of carbon-dioxide equivalent would be catastrophic. To understand the effect of this, we need to peer into what Harvard University economist Marty Weitzman calls the ‘fat tail’ of the probability distribution for climate damage. Although the likelihoodis uncertain — and probably low — we should give these events more attention because not doing so could be potentially disastrous.
>
> [...]
>
> The IPCC estimates about 2.5 °C to 6.4 °C as the “likely” range for warming by 2100 under A1FI, so there is a 5–17% chance that temperatures will go up by more than 6.4 °C by 2100.
>
> [...]
>
> In a 1,000 p.p.m. scenario, many unique or rare systems would probably be lost, including Arctic sea ice, mountain-top glaciers, most threatened and endangered species, coral-reef communities, and many high-latitude and high-altitude indigenous human cultures. People would be vulnerable in other ways too: Asian mega-delta cities would face rising sea levels and rapidly intensifying tropical cyclones, creating hundreds of millions of refugees; valuable infrastructure such as the London or New York underground systems could be damaged or lost; the elderly would be at risk from unprecedented heat waves; and children, who are especially vulnerable to malnutrition in poor areas, would face food shortages.
>
> [...]
>
> The economic outlook is no better. With warming of just 1–3 °C, projections show a mixture of benefit and loss. More than a few degrees of warming, however, and aggregate monetary

> impacts become negative virtually everywhere; and in a 1,000 p.p.m. scenario current literature suggests the outcomes would be almost universally negative and could amount to a substantial loss of gross domestic product. Millions of people at risk from flooding and water supply problems would provide further economic challenges
>
> The number and intensity of abrupt events and the possibility of irreversible damages goes up non-linearly with warming. If CO2 levels were to reach 1,000 p.p.m., a rise in sea levels of up to 10 meters after many centuries from the melting of the Greenland and West Antarctic ice sheets would be more likely . So would damage to coral and oceanic phytoplankton, as their calcium carbonate skeletons could dissolve in acidified oceans. Tropical rainforests would become more vulnerable to wildfire, and in some models such forests would switch from CO2 sinks to sources, adding yet more emissions. Extinction of some half of known plant and animal species would become much more likely, particularly if climate sensitivity is in the middle-to-upper part of the bell curve.

The impacts that Schneider describes seem qualitatively similar to median case impacts, differing in degree rather than in kind. The phrase "a substantial loss of gross domestic product" doesn't sound *that* bad – if he had said "a great majority of gross domestic product" things would be different.

(B)

Recall that Weitzman gave a 5% chance of 10+ Celsius increase by 2200 and a 1% chance of a 20+ degree Celsius increase by 2200. (I haven't checked to see whether he's assuming mitigation efforts.) In his 2011 paper Fat-Tailed Uncertainty in the Economics of Catastrophic Climate Change he wrote

> For me, 10 degree C offers both a vivid image and a reference point, especially in light of a recent study, which estimated that global average temperature increases of about 11–12 degrees C (with, importantly, accompanying humidity in the form of a high wet-bulb temperature) would exceed an absolute thermodynamic limit to metabolic heat dissipation (Sherwood and Huber 2010). Beyond this threshold, represented by a wet-bulb temperature of 35 C, more than half of today's human population would be living in places where, at least once a year, there would be periods when death from heat stress would ensue after about six hours of exposure. (By contrast, the highest wet-bulb temperature anywhere on Earth today is about 30 C). Sherwood and Huber (2010) further emphasize: ‘‘This likely overestimates what could practically be tolerated: Our [absolute thermodynamic] limit applies to a person out of the sun, in a gale-force wind, doused with water, wearing no clothing and not working.’’ Even at wet-bulb temperatures, much lower than 35 C, human life would become debilitating and physical labor would be unthinkable. The massive unrest and uncontainable pressures this might bring to bear on the world's human population are almost unimaginable. The Earth's ecology, whose valuation is another big uncertainty, would be upended. Thus, a temperature change of 10 C would appear to represent an extreme threat to human civilization and global ecology as we now know it, even if it might not necessarily mean the end of Homo sapiens as a species.
>
> It must be emphasized strongly that very high atmospheric temperature changes such as T ~ 10 C would likely take several centuries to attain. The higher the limiting temperature, the longer it takes to achieve equilibrium because the oceans will first have to absorb the enormous amounts of heat being generated. Alas, if the oceans are building up enormous amounts of heat it could set in motion irreversible long-term methane clathrate releases from the continental shelves along with some other nasty surprises. Thus, overall damages generated by equilibrium T ~ 10 C are best conceptualized as associated with being on the trajectory whose asymptotic limiting atmospheric temperature change is T ~ 10 C.

This makes a 10 degree C increase sound really bad, but it's significant that he says that it would likely take several centuries to attain.

(C) I looked at the 2010 paper by Sherwood and Huber that Weitzman cites above. They write

> Despite the uncertainty in future climate-change impacts, it is often assumed that humans would be able to adapt to any possible warming. Here we argue that heat stress imposes a robust upper limit to such adaptation. Peak heat stress, quantified by the wet-bulb temperature TW, is surprisingly similar across diverse climates today. TW never exceeds 31 °C. Any exceedence of 35 °C for extended periods should induce hyperthermia in humans and other mammals, as dissipation of metabolic heat becomes impossible. While this never happens now, it would begin to occur with global-mean warming of about 7 °C, calling the habitability of some regions into question. With 11–12 °C warming, such regions would spread to encompass the majority of the human population as currently distributed. Eventual warmings of 12 °C are possible from fossil fuel burning.
>
> Recent studies have highlighted the possibility of large global warmings in the absence of strong mitigation measures, for example the possibility of over 7 °C of warming this century alone (1).

In line with Weitzman's paper, this makes 10+ degree Celsius paper sound really bad. But the authors say that the peak heat stress will only *begin* to call into question habitability of some regions at 7 degrees Celsius, *and* only raise the possibility of *7+ degrees* Celsius by 2100 (as opposed to a higher increase). So the scenario that Sherwood and Huber discuss seems really far out in the future (if it will happen at all).

(D) I looked at Probabilistic Forecast for 21st Century Climate Based on Uncertainties in Emissions (without Policy) and Climate Parameters that Sherwood and Huber cite concerning the possibility of 7+ degree increase. On page 35 of the paper, the authors refer to one model that gives an 16.7% chance of a 6.42+ degree C warming by 2090-2100 and a 5% chance of 7.37+ degree C warming. Here too, I haven't checked to see what authors are assuming about mitigation efforts.

(E) Page 8 of the 2008 paper Did the Stern Review underestimate U.S. and global climate damages? estimates that assuming no adaptation, there's a 16.7% chance that GDP drop from climate change in 2100 will be 2.6% in the US, and 13.5% outside of OECD countries.

Some things that I haven't looked into

- Criticism of the claims of the authors above.
- How much smaller the problem would have been had mitigation efforts been taken decades ago.
- What the social cost of mitigation efforts would have been / would be.

Relevance to whether elites will successfully mitigate AI risk

We could dig deeper into the climate change issue, but here's where I stand at the moment:

- I think that the projections beyond 2100 have low relevance to whether elites will successfully mitigate AI risk:

  (a) It may be rational not to act based on predictions 90+ years out (because so many things will have changed by then). So if elites aren't taking into account what will happen after 2100, that could be a positive sign rather than a negative sign.

  (b) We plausibly have a better idea of how to reduce global warming 90+ years out than we do how to reduce AI risk conditional on AI risk being 90+ years out. It's plausible that it's substantially more likely that people in a position to influence AI risk will have more "skin in the game" (in the sense of AI having the potential to kill them personally), than do people in a position to influence climate change.

  Similar but weaker remarks apply to projections to the late 21st century.

- Restricting consideration years <= 2100, assume that there's no adaptation, no mitigation, no economic growth, and that the temperature increase is at the 95th percentile of the probability distribution. Under these (very strong) assumptions, examining the excerpts that I cited above, it's plausible to me that GDP wouldn't drop more than 50%. This would correspond to a drop in quality of life from present day standards to ~ 1970 standards. Though this would be bad, AI killing everyone seems at least 5x more bad (and maybe a lot more), even by near-term standards.

- It could be that there's a 1% chance of an 11-12 degree C temperature increase by 2100 in absence of mitigation efforts, rendering 50% of inhabitable land uninhabitable by present day standards, that we won't develop technology to make the land inhabitable, and that geoengineering doesn't work. This could result in a 2x cut in utils (corresponding to there being room for only 50% of people who would otherwise have room). Maybe the probability of the conjunction of these things is on the order of 0.1%. So maybe current policy makers aren't sufficiently capable so as to mitigate risks of 50% cuts in utility 90 years out. But this probability is a lot smaller than the probability of AI x-risk that most MIRI staff and supporters assign.

Jonah

From: **Jonah Sinick** <jsinick@gmail.com>
Date: Wed, Aug 14, 2013 at 2:11 PM
To: Luke Muehlhauser <lukeprog@gmail.com>

Hi Luke,

Responding concerning the 2008 financial crisis:

I read After the Music Stopped: The Financial Crisis, the Response, and the Work Ahead by Alan S. Blinder, who was the vice chairman of the federal reserve for 1.5 years during the Clinton administration.

Based on my reading, I think the conglomerate of poor decisions surrounding the 2008 financial crisis constitute a small but significant challenge to the view that the world's elites will successfully address AI risk.

My notes on some highlights of the book are below. My reading focused on Part II, which is about what caused the financial crisis, and about the government's immediate response. I also spent some time on Parts III and IV, which are about the government's recovery efforts and reform efforts.

The author's list of seven main factors that lead to the recession

These are (pg. 27)

1. Inflated asset prices, especially of houses (the housing bubble) but also of certain securities (the bond bubble)

2. Excessive leverage (heavy borrowing) throughout the financial system and the economy;

3. Lax financial regulation, both in terms of what the law left unregulated and how poorly the various regulators performed their duties;

4. Disgraceful banking practices in subprime and other mortgage lending;

5. The crazy-quilt of unregulated securities and derivatives that were built on these bad mortgages;

6. The abysmal performance of the statistical rating agencies, which helped the crazy-quilt get stitched together;

7. The perverse compensation systems in many financial institutions that created powerful incentives to go for broke.

- **Re: #1:**

  In 2001-2002, housing prices had gone up by 30% (adjusted for inflation) over a period of 5 years. At this time, some people raised the possibility that there was a housing bubble, but reasonable people can disagree, even with the benefit of hindsight. During 2004-2005, it gradually became more clear that there was a bubble and by 2006-2007 there were many indicators. (pgs 33-35)

  The bubble came from

  (a) Homebuyers buying houses that they couldn't afford with the expectation that they'd appreciate substantially. (pg. 35)

  (b) Capital migrating from the tech sector to the ostensibly safer housing sector after the 2000 tech stock bubble (pg. 37)

  (c) Bank and government policy encouraging people to get mortgages (pg. 38).

  There was also a *bond bubble*. The book's author predicted this during 2004-2006, but people were unreceptive to his warnings. The bubble started as a result of the Federal Reserve's monetary policy (lowering its interest rate down to 1954 levels). (pg. 45)

  These bubbles wouldn't have caused nearly as much damage as they did without leverage.

Re: #4

Subprime mortgages (mortgages to people whose credit ratings are too low for them to qualify for an ordinary mortgage) increased from $35 billion (< 5% of total mortgages) to $625 billion (~20% of total mortgages) from 1994-2005. Almost 1/3 of these were granted to borrowers with little or no documentation. (pg. 69)

Re: #5

The financial products based on mortgages that were designed and sold were very complex. This included the creation of ostensibly very safe bets that were in fact risky, which were then bet on with leverage. The complexity of these made it hard for buyers to assess what they were getting. (pg. 74-75)

The people who created these financial products were motivated to create them by the opportunity to take advantage of less savvy investors. (pg. 77)

Re: #6

The mortgage related financial products were often granted AAA ratings (corresponding to them being ostensibly very safe bets). These ratings were historically granted to only a small handful of assets. One reason why these products were granted such high ratings is overconfidence / neglect of tail risk. Another reason is that credit rating agencies are motivated to give good credit ratings to assets, because they're paid for rating the assets by the assets' owners. (pg. 79-80)

Re: #2:

Firms such as Bear Stearns, Lehman Brothers, Merrill Lynch, Morgan Stanley and Goldman Sachs were investing with 30:1 or 40:1 leverage (corresponding to having capital to back at most a ~3% drop in assets). This put them in a position where they could end up owing investors substantially more than they could afford to pay back. (pg. 52).

A large fraction of the leveraged investment took the form of bets on assets that neither party owned (pg. 67)

Re: #7

People in finance often invest with other people's money rather than their own. Making risky bets carries large upside for them (large bonuses) but little downside (being fired). They don't bear the cost of their bad choices. (pg. 81-82)

Re: #3

Government organizations such as the Federal Reserve, the Office of the Comptroller of Currency (OCC), the Office of Thrift Supervision (OTC) and the FDIC didn't keep track of and regulate the explosive growth of novel subprime mortgage lending. This is despite the fact that journalists were writing about these risky lending practices by 2004, and Edward Gramlich of the Federal Reserve and treasury official Sheila Bair warning that this was problematic. The government organizations repeatedly stated that they were going to regulate subprime mortgage lending, but never did. They may have been influenced by political pressures to increase home ownership. (pg. 58-59)

> There was also a "shadow banking system" that was heavily involved in borrowing and lending, and which was nearly entirely unregulated by the law. This system was probably significantly larger than the conventional banking system. (pg. 59)
>
> The derivatives market was largely unregulated. This is striking in light of the fact that Long-Term Capital Management's (LTCM) use of derivatives had contributed to a financial crisis in 1998. Even before the LTCM incident, Brooksley Born, the head of the CFTC suggested that derivatives be regulated by the CFTC. She was criticized publicly by the leadership of federal financial organizations (Alan Greenspan, Robert Rubin, Lawrence Summers, and Arthur Levitt). After the LTCM incident, Treasury secretary Lawrence Summers helped push through the Commodity Futures Modernization Act, which **explicitly barred** the possibility of CFTC regulation of derivatives. (pg. 62-63)
>
> Insurance company AIG had bet $500 billion against mortgages defaulting, in the form of credit default swaps. This was possible because AIG had a AAA credit rating, and because insurance regulators didn't regular AIG's credit default swaps (as they weren't officially classified as insurance). (pgs. 130-132)

What happened when the bubbles broke

The bubbles broke on August 9, 2007 (pg. 89)

The Federal Open Market Committee convened on August 16, 2007. The author of the book claims that they should have cut the federal funds rate immediately and failed to do so. The Federal Reserve didn't cut the funds rate until September 18, 2007. (pg. 92). It didn't act until October 31, when it cut the funds rate further by a small amount, and then again in December 11 (pg. 94). On January 22, 2008, FOMC took more drastic measures (pg. 97)

In March 2008, the investment bank Bear Stearns collapsed, and the Federal Reserve lent the firm $30 billion so that it could merge with JP Morgan (pg.100). This decision was met with a great deal of criticism, because it was seen to create a moral hazard (signaling that the government would be willing to bail out banks that made irresponsibly risky investments). (pg. 109-110). The decision was justified on the grounds that Bear Stearns was too interconnected for it to be a good idea to allow it to fail. It's unclear whether this is true (pg. 112)

The author claims that the bankruptcy of Lehman Brothers (on September 15, 2008) is nearly universally regarded as the watershed event of the financial crisis (pg. 127) There were various attempts to save Lehman Brothers. For example, it was proposed that Bank of America buy Lehman Brothers. Bank of America was willing to do so only if the government were to take over the risk connected with $40 billion dodgy assets. Treasury Secretary Henry Paulson refused. The US government had received negative publicity in connection with having bailed out Bear Stearns, and had also made large commitments to Fannie Mae and Freddie Mac, and there was fear of moral hazard, so Paulson decided against providing public money to bail Lehman Brothers out. (pg. 122) Technically speaking, bailing out Lehman Brothers was illegal, and this was seen as another reason not to bail out Lehman Brothers. (pg. 126) The impending bankruptcy of Lehman Brothers perceived to be as acceptable, on the grounds that the markets had had ample time (6 months) to prepare for the event (pg. 125).

When Lehman Brothers went bankrupt, the Reserve Primary Fund (the world's oldest money market) received a flood of redemption requests. It turned out that this fund had invested 1.2% of its capital in Lehman Brothers. Between these two things, Reserve Primary Fund was only able to return $0.97 for each dollar of investment. This was the first time that a money market hadn't returned at least $1 per dollar of investment. (pg. 143). This precipitated panic that money market funds might not be safe investments, and investors withdrew $350 billion from money market funds within a week. Many of America's biggest companies rely on loans from money market funds to cover short term expenses. So there was risk that companies wouldn't be able to meet payroll. Had this happened, there might have been a much larger financial crisis than the one that occurred. The Federal Reserve and the Treasury reacted quickly, and averted this outcome. (pg. 144).

Paulson approved $50 billion to back an insurance fund for money market balances. (pg. 145) This created perverse incentives (incentivizing people to pull money out of ordinary banks to put them in money market funds) and could have disrupted the entire banking system (pg. 146) The Federal Reserve also established a fund to give banks very favorable loans to give loans to money market funds. (pg. 147) Both policies were poorly crafted, and could have caused serious problems if they had been maintained, but they were ended before serious problems came up. The money markets stabilized and started functioning (pg. 148)

After Lehman Brothers went bankrupt, investors pulled capital out of the biggest five investment banks: Merrill Lynch, Washington Mutual, Wachovia, Morgan Stanley and Goldman Sachs (pg. 152) and experienced large losses as a result of having bet heavily on mortgage backed securities. The last three failed and were bought by Bank of America, JP Morgan, and Wells Fargo (respectively), with some government assistance. Citigroup suffered huge losses, and was rescued by the US government.

Government recovery programs

Congress enacted a program called Troubled Assets Relief Program (TARP). This was drafted as a contingency plan in case things got worse in March 2008. (pg. 177). The author of the book characterizes it as potentially one of the most successful economic policy innovations (pg. 177), and characterizes it as a smashing success (pg. 205) on account of having a net cost to taxpayers of only $32 billion.

A large part of TARP was the treasury forcing large banks to accept $125 billion in government loans under favorable terms (pg. 200). The author considers this to have been a bad economic policy, because the banks didn't need these loans, granting them was unlikely to increase lending, and the loans were given under unnecessarily favorable terms. (pg. 202)

The Federal Reserve created Term Asset-Backed Securities Loan Facility (pg. 206), which fulfilled TARPs initial purpose of buying some troubled assets.

There was also the American Reinvestment and Recovery Act of 2009. The author of the book says "when it came to targeting jobs, the 2009 stimulus package earned a B at best." (pg. 229)

Interest rate spreads

There was a problem of interest rate spreads. The author rates the US government's efforts to address this with an A (pg. 320)

Reforms

The US government created and passed a 2319 page reform bill known as Dodd-Frank to reduce the severity of future financial crises. The author says that there are good reasons to believe that the bill will reduce the severity and costs of future financial excesses (pg. 318) The author rates it with a B (pg. 320).

Mortgage foreclosure problem

The author rates the US government's efforts to address the problem of people's mortgages foreclosing with a C- or D (pg. 320). Fixing the foreclosure problem would have required $200 billion in government loans. This may have been paid back in full, but the Treasury decided that it wasn't worth it (pg. 322). One reason was legal problems associated with property rights (pg. 322). Another reason was that there were political pressures not to give loans to those whose homes had foreclosed, because they were seen to be irresponsible (pg. 323). The author says that failure to solve the foreclosure problem is a major reason for recovery from the 2008 recession being week for so long, and that it likely could have been addressed with more work (pg. 341 - 342).

Public relations problems

The US government did a poor job of choosing policies in response to the financial crisis that were good for public relations, and communicating what the policies were and what the rationale was:

- It's widely thought that the government gave banks money to keep, as opposed to loans. In 2012, only 15% of Americans who were surveyed correctly recognized that the banks had returned all or most of their loans to the US government (pg. 178).

- The US government gave the banks loans on very favorable terms, and this upset people because they felt that the government was using tax dollars to help special interest groups with no benefit to the public. (pg. 356)

- When the US government bailed out the major insurance company AIG by giving it $182 billion in loans, AIGs executives and traders received large bonuses, which angered the public (pgs. 136-137)

- Americans confused the troubled asset relief program with the stimulus package, because each involved about $700 billion, and so believed that the stimulus package consisted of bailing out banks rather than helping ordinary Americans.

- The government didn't clearly communicate that the financial crisis could have been much worse, and that the government's efforts should be judged based on how much damage was prevented rather than based on how much had occurred (pg. 354)

There was heavy backlash against the President and Congress (pg. 345) and the Federal Reserve (pg. 348)

From: **Jonah Sinick** <jsinick@gmail.com>
Date: Wed, Aug 14, 2013 at 3:18 PM
To: Luke Muehlhauser <lukeprog@gmail.com>

Hi Luke,

I haven't investigated any of

- The Iraq War
- Leaders being deposed or assassinated
- Recombinant DNA
- Nanotech
- Near earth objects
- Swine flu

in sufficient detail to have something substantive to say about them.

I haven't investigated research on cyberterrorism beyond what I wrote about in an earlier email in this thread: in particular, I didn't contact experts.

I did some reading on the Cuban Missile Crisis, and will send you thoughts on that. I haven't otherwise investigated nuclear proliferation.

I'm in the midst of reading about the history of chlorofluorocarbons, and will send you an email about what I learn.

Jonah

From: **Jonah Sinick** <jsinick@gmail.com>
Date: Fri, Aug 23, 2013 at 10:38 AM
To: Luke Muehlhauser <lukeprog@gmail.com>

Hi Luke,

Following up on

Please also send your thoughts on the negative EV of geomagnetic storms to some of the people who think it's a really big deal, to see whether they have a rebuttal.

I sent my analysis to William Radasky, a scientist who prepared a statement for the US House Homeland Security Subcommittee on Emerging Threats, Cybersecurity, and Science and Technology about the threat of geomagnetic storms.

He said that the relevant quantity is *not* the magnetic field (which is what I had been quoting from the OECD report — in my earlier email to you I wrote nT/min, but the numbers that I had been quoting were actually nT), but rather, the *rate of change* of magnetic field with time (nT/min).

The OECD estimated the frequency of storms with *magnetic field* the size of the 1859 storm (Carrington event) to be 7.41 per million years. That's the number that I was using. But If the relevant quantity is nT/min rather than nT, then this number is not very informative.

Radasky also said that there are studies that have indicated that storms at the Carrington event level (as measured in nT/min) could occur more often than each 100 years. But I wasn't able to find references for this.

If a Carrington event level storm occurred once every 100 years, then the negative expected value coming from such storms would be on the order of $30 billion per year rather than $30 million per year. This amounts to $1.5 trillion every 50 years. Compare with the negative expected value coming from the 2008 financial crisis, which is perhaps a 1 in 50 year event, and the cost of which has been estimated at $12.8 trillion (for Americans alone), and so an order of magnitude higher.

From: **Jonah Sinick** <jsinick@gmail.com>
Date: Fri, Aug 23, 2013 at 12:37 PM
To: Luke Muehlhauser <lukeprog@gmail.com>

Hi Luke,

Responding regarding the Cuban missile crisis:

I read the Wikipedia article, which is quite detailed.

At a high level, my main thought is that there's a lot of ambiguity as to how grave a threat the crisis posed (specifically, ambiguity as to the probability of escalation), but that even if things wouldn't have been bad in the median case, there was very large negative expected value coming from tail risk.

Martin Hellman did a Fermi calculation in which he estimated the probability of a nuclear exchange per year as 0.2%-1%. The assumptions that he uses to arrive at this number seems reasonable, but *if the article accurately represents his position*, he concludes that "the risk of a person not living out his or her natural life because of nuclear war is at least 10 percent," which seems like a *huge* leap — he seems to have assumed that [the probability that a nuclear weapon being used leads to a war that kills most people] is very close to 1 (!!!).

There are hints that the Americans and Russians were willing to take refrain from retaliating against perceived affronts in order to prevent escalation. For example, the article quotes Robert McNamara as saying

We had to send a U-2 over to gain reconnaissance information on whether the Soviet missiles were becoming operational. We believed that if the U-2 was shot down that—the Cubans didn't have capabilities to shoot it down, the Soviets did—we believed if it was shot down, it would be shot down by a Soviet surface-to-air-missile unit, and that it would represent a decision by the Soviets to escalate the conflict. And therefore, before we sent the U-2 out, we agreed that if it was shot down we wouldn't meet, we'd simply attack. It was shot down on Friday. ... Fortunately, we changed our mind, we thought "Well, it might have been an accident, we won't attack." Later we learned that Khrushchev had reasoned just as we did: we send over the U-2, if it was shot down, he reasoned we would believe it was an intentional escalation. And therefore, he issued orders to Pliyev, the Soviet commander in Cuba, to instruct all of his batteries not to shoot down the U-2.

One could imagine this sort of reasoning taking place with reasonably high probability at every stage of hypothetical escalation, driving the probability of an all-out nuclear way down.

Still, even if the probability of the Cuban missile crisis leading to an all out nuclear war was only 1% or so, the risk was still sufficiently great so that the way in which the actors handled the situation is evidence against elites handling the creation of AI well. (This contrasts with the situation with climate change, in that elites had strong personal incentives to avert an all-out nuclear war.) Here is a list of some apparently poor decisions that were made, with the qualifier that my knowledge here is very shallow:

- Kennedy launched the Bay of Pigs invasion of Cuba, which was botched, signaling to the USSR that the United States was militarily weak and that aggression on the USSR's part would be successful.

- The United States alienated Cuba by trying to invade it, signaling that it might invade Cuba again.

- The US and USSR didn't successfully negotiate to nullify the threats that the USSR perceived to East Germany from West Germany, and also from missiles that the US had recently installed in Turkey.

- Khrushchev was actively and intentionally deceptive, ensuring Kennedy that the USSR wouldn't bring missiles to Cuba, rather than bargaining (e.g. saying "We feel threatened and so will be bringing missiles to Cuba unless you remove your missiles from Turkey") and presumably this eroded trust.

- In response to learning that missiles had been deployed in Cuba, the US Joint Chiefs of Staff (consisting of high ranking US military officials) unanimously agreed that the United States should invade Cuba and overthrow Castro. This could have precipitated escalation. But Kennedy decided not to follow their recommendation.

- Khrushchev was not at all sympathetic to Kennedy in regards to Kennedy ordering a blockade on Cuba (characterizing it as an aggressive act rather than a defensive act).

- The United States dropped depth charges on a Soviet submarine that had a nuclear torpedo, not considering the possibility that it might have a nuclear torpedo. Two of the three Soviet officers on the submarine wanted to use the torpedo, and the outcome was only averted because one of the officers dissented.

From: **Jonah Sinick** <jsinick@gmail.com>
Date: Fri, Aug 30, 2013 at 12:05 PM
To: Luke Muehlhauser <lukeprog@gmail.com>

Hi Luke,

I read Ozone Crisis: The 15-Year Evolution of a Sudden Global Emergency by Sharon Roan, and took notes on the first 100 pages, which are below. There's interesting stuff in the remainder of the book, which would be worth getting notes on.

Jonah

- The fact that CFCs deplete the ozone layer was discovered through ordinary basic science (pgs. 3-6)

- When CFCs enter the stratosphere, they're exposed to ultraviolet light, and free chlorine atoms split off. Chlorine atoms bond to ozone, which in turn becomes oxygen and chlorine monoxide, but chlorine monoxide is unstable, and bonds with free oxygen to create a free chlorine atom and oxygen. The cycle then repeats. The free oxygen atom would otherwise have combined with oxygen to produce ozone. So ozone is continually converted into oxygen. (pg. 8-9).

  Others had made a similar discovery for nitrogen oxides 1971, and raised the possibility that supersonic jets might deplete the ozone layer by releasing these into the atmosphere (pgs. 12-14). This led to the senate voting against a measure to fund supersonic jets (pg. 15)

- When Rowland publicized his discovery, the chemical company Du Pont (which was the main manufacturer of CFCs) started to take measures to minimize the issue (pgs. 20 onward).

- Rowland and Molina put together a 150 page paper in preparation for the 1974 September meeting of the American Chemical Society (ACS). ACS news manager Dorothy Smith read the abstract of their paper and sent information about this to the media. A Du Pont official contacted her to say that it was an insignificant story that was not good chemistry. But she consulted independent scientists who said that the work was good, so she proceeded. (pgs 25-26)

  Rowland and Molina estimated that if CFC production were to increase by 10% a year until 1990 and then remain steady, 5%-7% of the ozone would be lost by 1995, and 30%-50% would be lost by 2050, and highlighted the dangers of skin cancer and possible changes in climatic patterns. (pg 27)

- Rowland and Molina were initially unskilled with public relations issues, and initially floundered in their efforts to communicate their discovery and its significance to the press. However, they worked hard to communicate their ideas, and environmentalists and scientific journalists picked up on them. Some Harvard atmospheric scientists also helped with publicity (pgs. 27-28)

- In October 1974, the National Academy of Sciences (NAS) created a panel to determine how serious the CFC-ozone problem was. (pg. 29) In November 1974, The Natural Resources Defense Council became involved (pg. 31).

- By Spring 1975, 11 states had legislative proposals regarding CFC regulations. There was more pushback from Du Pont. (pg. 47)
  Consumers started to oppose CFCs on out of fear of increases in skin cancer (pgs. 58-59)
  Companies started marketing products without CFCs at consumers (pg. 60)

- In 1975, the US government policy makers were unclear on what to do, because there was uncertainty as to whether the CFC-ozone problem was real, and because banning CFCs would have negative economic consequences (pg. 39) The federal government created the "Committee on the Inadvertent Modification of the Stratosphere," which determined that if the NAS found that CFCs were hazardous, the the federal government should act to restrict CFC use. The NAS named a 12 person panel to investigate. (pgs. 41-42).

- During 1975, industry representatives highlighted six scientific uncertainties calling into question the theory that CFCs were depleting the ozone layer. Industry supported researchers worked to disprove the theory, and scientists (pgs. 65-66). Other scientists worked to prove the hypothesis (pg.64 & pgs. 68-70)

  Some scientists questioned whether the consequences of CFC emissions would be severe. (pg. 67)

  There was a scientific controversy involving chlorine nitrate (pg. 70-79) that delayed the publication of the of the NAS report until September 1976. The NAS report concluded that CFCs were damaging the atmosphere and should be restricted, and agreed with the theory that reduced ozone would lead to more ultraviolet light reaching the earth, increasing skin cancer. (pg. 81)

- In May 1977, the US Environmental Protection Agency, Food and Drug Administration, and Consumer Product Safety Commission announced a time table to phase out nonessential use of CFCs the end of 1978. The aerosol industry adapted. (pgs 84-86)

- Chemists started researching the impact of other chemicals on the ozone layer (pgs. 90-93) There was more and more research having to do with ozone depletion, with some evidence supporting the CFC-ozone connection and other opposing it (pgs. 94-100, pg. 112) Banning CFCs was postponed very substantially.

- Under the Reagan administration, restriction of CFC use was stalled on the grounds that there wasn't enough evidence that CFCs deplete the ozone layer, and that restricting their use could cause economic hardships (pgs. 104-105)

- There wasn't very much research into substitutes for CFCs, and the US government not providing funding for such research is considered by some to have been a significant failure in how the CFC-ozone situation was handled (pg. 100).

From: **Luke Muehlhauser** <luke@intelligence.org>
Date: Thu, Sep 5, 2013 at 3:08 PM
To: Jonah Sinick <jsinick@gmail.com>

Hi Jonah,

Back to the 2008 financial crisis: Can you say more about why you think that "the conglomerate of poor decisions surrounding the 2008 financial crisis constitute a small but significant challenge to the view that the world's elites will successfully address AI risk"?

It'd be nice to see you outline how the 2008 crisis matches our criteria (like you did with smallpox eradication), and then explain why you think the kinds of failures you learned about when reading *After the Music Stopped* should give us some concern that policy makers will not handle AGI wisely.

Luke

From: **Jonah Sinick** <jsinick@gmail.com>
Date: Wed, Sep 11, 2013 at 2:45 PM
To: Luke Muehlhauser <luke@intelligence.org>

Hi Luke,

Responding again on the 2008 financial crisis:

We had highlighted 6 dimensions along which a historical case might be analogous to the AI risk situation. Assessing the 2008 financial crisis along these dimensions:

#1 AI may become a major threat in a somewhat unpredictable time.

1. People have said that the Great Recession that followed the 2008 financial crisis was the worst since the 1933 Great Depression. This places the frequency of such events at ~1 every 75 years. One can't expect to predict such events with precision.

2. Almost tautologically, a financial crisis is unexpected.

3. Though I don't have page numbers handy, the impression that I got from reading *After The Music Stopped* was that even though people recognized that there was a housing bubble, almost nobody anticipated the severity of the ensuing crisis.

So I think that the situation is analogous to the situation with AI along this dimension.

#2 AI may become a threat when the world has very limited experience with it.

There's a long historical precedent of financial crises, both domestic and international. Wikipedia lists 22 financial crises from 1900 through 2001. Arguably, the world had little experience with financial crises of the *magnitude* of the 2008 financial crisis, but the general impression that I got from *After The Music Stopped* was that the 2008 crisis differed from previous crises in *degree* rather than *kind*, and there were known countermeasures.

#3 A good outcome with AI may require solving a difficult global coordination problem.

While the 2008 financial crisis seems to have been largely US specific (while having broader ramifications), there's a sense in which preventing it would have required solving a difficult coordination problem. The causes of the crisis are diffuse, and responsibility falls on many distinct classes of actors. It's hard to create an exhaustive list, but some classes of actors involved are

- Actors who established bank and government policy that encouraged people to buy homes and get mortgages.

- Actors who bought houses with the expectation that they'd appreciate in value substantially, even though the housing market was overpriced.

- The Federal Reserve, which interest rates substantially in the early 2000's, giving rise to a bond bubble.

- Actors who bet on bonds even though the the bond market was overpriced.

- Government organizations such as the Federal Reserve, the OCC, the OTC, and the FDIC, which didn't keep track of the explosive growth of the subprime mortgage lending

- Actors who designed complex, highly nontransparent financial products that were mortgage related.

- Rating agencies that gave mortgage related financial products AAA ratings.

- Alan Greenspan, Robert Rubin, Lawrence Summers, and Arthur Levitt, who refused regulate leveraged investments.

- Actors in finance who made leveraged investments that they didn't have the money to cover if needed.

One gets a strong sense that it was a "tragedy of the commons" sort of situation.

#4 Preparing for the AI threat adequately may require lots of careful work in advance.

My understanding of the situation isn't strong enough for me to assess whether the 2008 financial crisis is analogous in this respect.

#5 Elites have strong personal incentives to solve the AI problem.

- The incentives weren't as strong as they would be in the case of AI risk (it wasn't a life or death matter for almost any of the actors involved)

- The people in finance who were responsible didn't have incentive to prevent the financial crisis — *When The Music Stopped* repeatedly emphasize that they were in a "heads I win, tails you lose" situation, where they would reap the benefits if risky investments playing out well, but wouldn't bear the losses if they played out poorly. Also, as a factual matter, some of the people in finance who were responsible got big bonuses after the crisis precipitated.

- I would guess that the government decision makers had incentive to be remembered favorably in retrospect rather than negatively.

- Many interest groups were adversely affected by the financial crisis, and one might expect that they'd be motivated to lobby for government policy that would have prevented it.

#6 A bad outcome with AI would be a global disaster, a good outcome with AI would have global humanitarian benefit.

Here too, the magnitude of the disaster isn't nearly as large as the magnitude AI risk. But the cost of the financial crisis has been estimated at $14 trillion, which is equal to 1 year of US GDP.

Compare with, e.g. Apple's market cap of ~$500 billion (though Apple probably has unusually high positive externalities).

To quantify the comparison with AI risk a bit: if we assume that the financial crisis made the average American's life 10% worse for 3 years, and assume that with discounting, people assess the disvalue of AI killing everyone at 30 DALYs/American, then the financial crisis is something like 1% as bad as

AI killing everyone. Scope insensitivity could make the decision relevant percentage higher. I think that this is is high enough for the financial crisis to be relevant on this dimension.

With all of that as background, elaborating on my claim "the conglomerate of poor decisions surrounding the 2008 financial crisis constitute a small but significant challenge to the view that the world's elites will successfully address AI risk":

I should clarify that the decisions that I see as most relevant are those **leading up to** the 2008 financial crisis — the decision making **afterward** doesn't seem so bad.

A few things to highlight here (partially summarizing key points from what I wrote above):

1. As above, I think that the magnitude of the problem is nontrivial (even if small) compared with the magnitude of the AI risk problem.

2. The financial crisis adversely affected a very broad range of people, apparently including a large fraction of those people in positions of power. A recession is bad for most businesses and for most workers. Yet these actors weren't able to recognize the problem, coordinate, and prevent it.

3. The reasons that they weren't able to recognize the problem, coordinate, and prevent it seem directly related to reasons why people might not recognize AI risk as a problem, coordinate, and prevent it:

   (a) A number of key actors involved seem to have exhibited conspicuous overconfidence and neglect of tail risk (e.g. Summers, etc. ignoring Brooksley Born's warnings about excessive leverage). If true, this shows that people in positions of power are notably susceptible to overconfidence and neglect of tail risk. Avoiding overconfidence and giving sufficient weight to tail risk may be crucial in mitigating AI risk.

   (b) One gets a sense that bystander effect and tragedy of the commons played a large role in the case of the financial crisis. There are risks that weren't adequately addressed because doing so didn't fall under the purview of any of the existing government agencies. This may have corresponded to a mentality of the type "that's not my job — somebody else can take care of it." If people think that AI risk is large, then they might think "if nobody's going to take care of it then I will, because otherwise I'm going to die." But if people think that AI risk is small, they might think "This probably won't be really bad for me, and even though someone should take care of it, it's not going to be me."