Subject: **Can we know what to do about AI?**

From: **Luke Muehlhauser** <luke@intelligence.org>
Date: Sat, Jun 15, 2013 at 9:12 PM
To: Jonah Sinick <jsinick@gmail.com>

Hi Jonah,

This is the email thread — with the subject line "Can we know what to do about 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 write a reply to Scott Aaronson's question "Given our current ignorance about AI, how on earth should we act on [the conclusion that AIs might not be human-friendly]?"

Part of the project involves you and I working to steel-man Scott's objection as much as we can, given resource limitations. Then I'll try to develop whatever response I think is reasonable for that steel-manned objection, whatever it turns out to be.

However, we learned in our recent conversation with Scott that his objection about not knowing how to reduce AI risks just comes back to his prediction that AI is probably several centuries away — a point I already replied to with my article When Will AI Be Created? When I asked Scott if he would retain his objection (re: "Can we know what to do about AI?") if he had roughly the same probability distribution over year of AI creation as I gave in 'When Will AI Be Created', he said "No, if my distribution assigned any significant weight to AI in (say) a few decades, then my views about the most pressing tasks today would almost certainly be different."

Still, the objection I want to develop is "Even if AI is somewhat likely to arrive during the latter half of this century, how on earth can we know what to do about it *now*, so far in advance?" — even though that's not exactly *Scott's* objection, since his objection comes from a belief that AI is many centuries away.

Let me say a bit more about what I have in mind for this project.

I think there are plausibly many weak arguments and historical examples suggesting that *P*: "it's very hard to nudge specific distant events in a positive direction through highly targeted actions or policies undertaken today." Targeted actions might have no lasting effect, or they might completely miss their mark, or they might backfire.

If *P* is true, this would weigh against the view that a highly targeted intervention today (e.g. Yudkowsky's Friendly AI math research) is likely to positively affect the future creation of AI, and might instead weigh in favor of the view that all we can do about AGI from this distance is to engage in broad interventions likely to improve our odds of wisely handling future crises in general — e.g. improving decision-making institutions, spreading rationality, etc.

I'm interested in abstract arguments for *P*, but I'm even more interested in historical data. What can we learn from seemingly analogous cases, and are those cases analogous in the relevant ways? What sorts of counterfactual history can we do to clarify our picture?

One case worth looking into is Club of Rome's *The Limits to Growth*. It's often depicted as a failed doomsday prophecy by a well-regarded think tank, but I know some in the academic community say its failures are overblown. I want to know: were Club of Rome's projections roughly right for the past 30 years? Were the policy recommendations they took from their projects executed? If so, what effects do they seem to have had? If not, can we make a good guess at what their effects would have been, or is it impossible to say?

An early draft of Intelligence Explosion: Evidence and Import collected additional examples of people trying to usefully influence things from several decades away with targeted interventions:

...mere mortals have at times managed to reason usefully and somewhat accurately about the future, even with little data. When Leo Szilard conceived of the nuclear chain reaction, he realized its destructive potential and filed his patent in a way that kept it secret from the Nazis (Rhodes 1995, 224–225). Svante Arrhenius' (1896) models of climate change lacked modern climate theory and data but, by making reasonable extrapolations from what was known of physics, still managed to predict (within 2°C) how much warming would result from a doubling of CO2 in the atmosphere (Crawford 1997). Norman Rasmussen's (1975) analysis of the safety of nuclear power plants, written before any nuclear accidents had occurred, correctly predicted several details of the Three Mile Island incident that previous experts had not (McGrayne 2011, 180).

But Anna and I didn't have time to check whether these were reasonable interpretations of the events, so the paragraph was cut. You could investigate those cases more closely

I'll share other ideas later, but this is a start. Of course, I hope you have your own ideas about how to proceed, too.

Cheers!

Luke Muehlhauser
Executive Director

From: **Jonah Sinick** <jsinick@gmail.com>
Date: Mon, Jun 17, 2013 at 1:55 PM
To: Luke Muehlhauser <luke@intelligence.org>

Hi Luke,

I just spent some time looking into *The Limits of Growth*. My first (superficial) impressions:

- It seems like it should have been *a priori* clear that the more extreme claims made in the book shouldn't be given much weight. There may be an element of hindsight bias here, but:

  (a) The team of people behind the book don't seem to have had a strong track record of tangible high quality work before having written the book. (I haven't carefully vetted this claim — it's just a first impression). Assuming that my impression is right, I would not give much weight to a book written by team with similar credentials on any nontechnical subject.

  (b) The fundamental premise behind the model seems to be that the need for resources increases exponentially, while the supply of resources due to technology increases linearly. The assumption of linear increase relies on the assumption that the resources that are depleted can't be substituted with other resources, and this is a very strong assumption. The point about substitution was noted by Julian Simon in 1980, and presumably was noted by others much earlier on.

Of course, it's *a priori* unclear that substitution would suffice to avert major problems coming from resource shortage, and indeed, unclear that it will suffice in the future. But the argument seems very noncontingent: it seems likely to be an argument that could have been made at many points in history, without having turned out to be true.

- In view of the above point, it seems to me as though the book is insufficiently credible to warrant giving its track record nontrivial weight in determining whether best possible predictions about the future will turn out to be accurate.

- With the above point in mind, I'm uninclined to look into the question of what the book's influence was, and whether its influence was positive or negative. But let me know if you'd like me to do so, and if so, I will.

- Julian Simon seems like a stronger candidate for being a sound predictor of the future than the team behind *The Limits of Growth* is, and it may be interesting to look into why he thought what he thought, and whether his reasons turned out to be right.

- According to Wikipedia, there are some people who still defend the book's conclusions. In particular, there's a 2008 paper by Graham Turner arguing that the book's predictions have been accurate to date, and implicitly suggesting that because this is the case, one should give weight to the book's future predictions as well.

  Vipul Naik pointed out that a model may make asymmetrically bold claims about the near/medium future vs. the far future (making bolder claims about the latter than the former), and that for this reason, one should be wary about using the historical success of a model as evidence for the reliability of the model's future predictions. He also pointed out that models will (by design) generally make good predictions about the near term, and that if a model's predictions become *better* over time, that's stronger evidence for the reliability of the model than a model's initial predictions being good.

- Vipul recommended

  (a) *The Ultimate Resource* by Julian Simon, which argues that as resources get scarcer, the prices go up, which incentivizes people to find substitutes, which mitigates the resource shortage problem. The book's central thesis seems pretty obvious, but the book may have interesting case studies that are worth looking at.

  (b) The Skeptical Environmentalist by Bjørn Lomborg, which argues that claims of problematic resource shortage are unfounded. I think I should table this one unless you'd like me to dig further into *The Limits of Growth*.

  (c) The Rational Optimist by Matt Ridley, which gives historical perspective related to environmental issues and war. This book is also notable for being reviewed by Bill Gates.

  (d) Nate Silver's book The Signal and the Noise about the success and failure of predictions about the future. I know you're probably well-acquainted with this book. It might make sense for me to read it as background for this project.

- You had recommended that I read Philip Tetlock's book Expert Political Judgment: How Good Is It? How Can We Know?

As a next step, I'm inclined to read Tetlock's book and Silver's book, but I can dig deeper into the resource shortage predictions first, if you'd prefer that I do so.

Best,
Jonah

From: **Luke Muehlhauser** <luke@intelligence.org>
Date: Tue, Jun 18, 2013 at 2:26 PM
To: Jonah Sinick <jsinick@gmail.com>

Hi Jonah,

See below for my comments.

> (b) The fundamental premise behind the model seems to be that the need for resources increases exponentially, while the supply of resources due to technology increases linearly.

Or, more generally, resource depletion will occur if the growth rate of the need is a faster exponential than the growth rate of the supply. If need doubles every 20 years, and supply doubles every 40 years, the resource will be depleted. (This doesn't take into account other factors, e.g. the substitution point, which I agree seems pretty obvious.)

> Of course, it's a priori unclear that substitution would suffice to avert major problems coming from resource shortage, and indeed, unclear that it will suffice in the future. But the argument seems very noncontingent it seems likely to be an argument that could have been made at many points in history, without having turned out to be true.

The claim that "We need major technological advances to keep up the economic trends of recent decades" wasn't true for most of history — that is, until the Industrial Revolution. Even since then, I've heard that trends in resource prices (Google "real oil prices" or "real commodities prices") have sometimes favored Simon's view, and sometimes not. But I haven't checked those data myself.

> Julian Simon seems like a stronger candidate for being a sound predictor of the future than the team behind The Limits of Growth is, and it may be interesting to look into why he thought what he thought, and whether his reasons turned out to be right.

I've heard that Simon basically just did naive trend extrapolation on the actual variable of interest (resource prices). I've also heard he made some pretty weird claims about unlimited growth and relative magnitudes, but I haven't checked.

> I'm uninclined to look into the question of what [*The Limits to Growth*'s influence] was, and whether its influence was positive or negative

I kind-of agree. At least, let's see whether we can find historical examples that "a priori" seem more likely (than *Limits to Growth*) to have been a successful attempt at figuring out how to use targeted interventions to influence a decades-away event. But, please *at least* check in more detail where you were right about your speculation that "The team of people behind the book don't seem to have had a strong track record of tangible high quality work before having written the book."

> Vipul Naik pointed out that a model may make asymmetrically bold claims about the near/medium future vs. the far future (making bolder claims about the latter than the former), and that for this reason, one should be wary about using the historical success of a model as evidence for the reliability of the model's future predictions. He also pointed out that models will (by design) generally make good predictions about the near term, and that if a model's predictions become better over time, that's stronger evidence for the reliability of the model than a model's initial predictions being good.

Yup!

> (a) *The Ultimate Resource* by Julian Simon

Sure, you could glance through this to see if it looks promising for our project.

> (b) *The Skeptical Environmentalist* by Bjørn Lomborg

Yes, let's skip for now.

> (c) *The Rational Optimist* by Matt Ridley

What's the plausible relevance of this book to our current project?

Yes, reading *Signal and the Noise* and *Expert Political Judgment* will be good for this project in general. Please also read Tetlock (2010), which contains some important qualifications of *Expert Political Judgment*, and Tetlock's two chapters in Tetlock et al. (2006), which contain important points on the difficulty of doing counterfactual history, which will be necessary for this project.

As you're reading them, please take notes on whether any parts of their content may be particularly relevant to our current project. I can't remember whether they contain good leads for our current project.

Please also look into the examples given in the paragraph I sent you from an early draft of *Intelligence Explosion: Evidence and Import*, to see whether any of those examples look like promising case studies for this project.

Luke

From: **Jonah Sinick** <jsinick@gmail.com>
Date: Sat, Jun 22, 2013 at 11:55 AM
To: Luke Muehlhauser <luke@intelligence.org>

Hi Luke,

For now, just responding on the point about the track record of the team behind *Limits to Growth*.

One very salient feature of the team is that they were young at the time when they published the book:

- Donella Meadows was born in 1941, and so was ~31 years old

- Dennis Meadows was born in 1942, and so was ~30 years old.

- Jørgen Randers was born in 1945, and so was only ~27 years old. He finished his PhD in 1973, 1 year after the publication of the book.

- I can't find William Behrens's age, but he finished his PhD in 1975, 3 years after the publication of the book.

I would not expect 4 people under 31 to do a good job predicting the future, even if they were the smartest people who I know. Nate Silver was already 30 years old when he predicted the presidential election results in 49 of 50 states.

Donella Meadows seems to have had the strongest credentials of the collection of authors. She got a PhD in biophysics from Harvard in 1968, and published about 9 papers in biophysics before writing *The Limits to Growth*. Based on the reference class "prestige of institution, credibility of field of research, and number of publications," her accomplishments are similar to those of the most accomplished people who we know in our age group. Her thesis has been cited by ~200 publications, which is respectable, though over a span of 40+ years.

It's hard to assess Dennis Meadow's credentials. His work was in the field of System Dynamics. I can't tell whether the field has born any fruit, or whether it's just something trendy that caught on.

As indicated above, the other two authors hadn't yet finished their PhDs when they coauthored the book.

To summarize, based on my cursory impressions it appears that there are a number of reasons for thinking that the project was *a priori* doomed to failure:

1. The team either missed or didn't give sufficient weight to the apparently obvious point about substitutability.

2. There wasn't a historical precedent for work of the type that they were doing succeeding.

3. The team was very young and inexperienced.

4. Only one member of the team had a strong track record of achievement.

I recognize that I haven't investigated #2 and #4 sufficiently carefully to have very high confidence in them, but I think that the broad picture is such that what the team did should be thought of as having *a priori* very low probability of success.

I think that the main updates against MIRI being able to do valuable work in AI forecasting come from

- The team being young and inexperienced (MIRI's team is similarly young and inexperienced).
- The fact that somebody as accomplished as Donella Meadows seems to have missed the mark by a large margin, despite having made the issue her life focus.

I feel comfortable considering the investigation of *The Limits to Growth* to be complete.

Best,
Jonah

From: **Jonah Sinick** <jsinick@gmail.com>
Date: Sat, Jun 22, 2013 at 12:37 PM
To: Luke Muehlhauser <luke@intelligence.org>

Hi Luke,

Responding on the point of Svante Arrhenius and climate change:

- Arrhenius's chemistry was sound: the equation for how the Earth's temperature varies as a function of concentration of carbon dioxide is the same equation used today.
- For the most part, Arrhenius didn't model how of increased carbon concentrations would impact other factors that influence the Earth's temperature. I don't know if this is because he wasn't aware of these, because he thought that they were sufficiently small to ignore, or because he didn't try to.
- Knut Ångström criticized Arrhenius's claim on scientific grounds, giving a different model that predicted no climate change from increased carbon concentrations. My surface impression is that Arrhenius was a much more accomplished scientist than Ångström was. To the extent that this is true, I think that Ångström's view should be heavily discounted, but I haven't investigated further.
- While Arrhenius recognized that the use of fossil fuels could increase atmospheric concentrations, he failed to predict how fast carbon emissions would increase (by a huge margin) because he didn't recognize how widespread fossil fuel use would become.
- People later thought that Arrhenius's prediction that atmospheric carbon would increase was wrong, because they thought that oceans would serve as great carbon sinks. It would be interesting to look into whether they had good reasons for thinking this at the time.

- Arrhenius predicted that global warming would have positive humanitarian impacts, for reasons that are retrospectively wrong, and instead global warming appears to have negative humanitarian impacts.

Taking this all together, based on my surface impressions, I think that this case study should cause one to update away from predicting the far future being useful. Some points:

- To the extent that Arrhenius was right, he was largely ignored.

- Arrhenius could have been wrong (the countervailing theories could have been right), but this warrants further investigation.

- Based on our present understanding, if, in the subsequent years, people had started burning fossil fuels more with a view toward giving rise to positive humanitarian impacts, they would have done more harm than good.

Best,
Jonah

From: **Jonah Sinick** <jsinick@gmail.com>
Date: Sat, Jun 22, 2013 at 1:35 PM
To: Luke Muehlhauser <luke@intelligence.org>

Hi Luke,

Responding on the point of the report on nuclear power plant risk by Rasmussen:

- The report predicted one core damage accident per 20,000 years of nuclear power plant operation.

  As a point of comparison, at the time of the Three Mile Island incident, only 500 years of nuclear power plant operation had occurred: far fewer than 20,000 years. This could be a fluke. The Chernobyl accident occurred only 6 years later. If the cause was the same, this strongly suggests that the Three Mile Island incident was not a fluke.

  However, the report was based on reactor designs that didn't include the Three Mile Island type.

- The report did discuss tidal waves as a potential cause for nuclear disaster, anticipating the recent disaster in Japan. But the report is 21 volumes long, and so this (weak) prediction could have been cherry picked retrospectively.

- The report is considered to be obsolete, and its main lasting value seems to have been pioneering use of probabilistic risk assessment.

- The report was controversial, and other scientists criticized the report as understating the risks of nuclear power plants.

It appears that the situation can be summarized as:

"People were concerned about nuclear power plants being dangerous. The Nuclear Regulatory Commission created a report analyzing the risks. The report was really long, and didn't address key relevant factors. Other scientists thought that the report understated the risks. The report was quickly recognized to have major flaws and became obsolete."

This doesn't seem to have much relevance to MIRI's work on AI forecasting. Some differences:

- The issue was one that a lot of people were already concerned about.
- The issue was highly domain specific
- Rather than there being a few salient predictions, there were a huge number of small predictions.

One way in which the situation is relevant is that risk was ignored. If we want to make a list of situations in which risk was ignored, this might be a good example. But I would guess that there's no paucity of such examples.

Best,
Jonah

From: **Jonah Sinick** <jsinick@gmail.com>
Date: Sat, Jun 22, 2013 at 2:12 PM
To: Luke Muehlhauser <luke@intelligence.org>

Hi Luke,

Responding on the point of Leo Szilard recognizing the destructive potential of a nuclear chain reaction and with this in mind, patenting his discovery in secret so that the Nazis wouldn't find out:

I think that this isn't a good example of a nontrivial future prediction. The destructive potential seems pretty obvious – anything that produces a huge amount of concentrated energy can be used in a destructive way. As for the Nazis, Szilard was himself Jewish and fled from the Nazis, and it seems pretty obvious that one wouldn't want a dangerous regime to acquire knowledge that has destructive potential.

It would be more impressive if the early developers of quantum mechanics had kept their research secret on account of dimly being aware of the possibility of destructive potential, or if Szilard had filed his patent secretly in a hypothetical world in which the Nazi regime was years away.

I've now addressed the three historical examples that you and Anna mentioned in the early draft of *Intelligence Explosion: Evidence and Import*.

I think that the Arrhenius climate change thing is worth looking at more carefully. I'm curious about how much confidence Arrhenius could justifiably have had given the information available at the time. As I said, I don't think that Arrhenius making his prediction had high expected value (and indeed, the sign of the expected value is ambiguous), but it's still a striking example of somebody having correctly made a controversial prediction about something many decades away and turning out to be right.

What do you think?

Best,
Jonah

From: **Luke Muehlhauser** <lukeprog@gmail.com>
Date: Sat, Jun 22, 2013 at 3:32 PM
To: Jonah Sinick <jsinick@gmail.com>

Re: *Limits to Growth*.

> I think that there are a number of reasons for thinking that the project was a priori doomed to failure

But *did* the analysis in *Limits to Growth* seriously miss the mark? In what ways did it succeed, and in what ways did it fail? Did the researchers actually miss the substitution point, or did they have reasons to think it wouldn't matter? Without knowing this we can't update; we're just left with *a priori* guesses about the relevance of age and prestige to predictive powers. Please do spend some time investigating this further; thanks!

Thanks for the details on Arrhenius' predictions. I agree that this case warrants further investigation.

Thanks also for the details on the Rasmussen report and the Szilard case. I don't think we need to dig deeper for those cases.

Another interesting case study you could look into is Drexler on nanotechnology. My rough impression of what happened is this: Drexler launched the field with a 1981 paper and especially his 1986 book *Engines of Creation*. The book mentioned the possibility of the "grey goo" scenario, which caught the public's attention despite it being a relatively predictable and avoidable scenario. Public concern about the dangerous effects of nanotechnology made it harder for some researchers to get funding for their harmless nanomaterials research, and they managed to exclude Drexler from having much of an influence over the National Nanotechnology Initiative (despite him being an obvious choice to have the *most* influence over it, especially after the impressive *Nanosystems*). I'm also under the impression that while Drexler's expected timelines for atomically precise manufacturing didn't hold up, his physical analyses in *Nanosystems* still seem remarkably prescient. *Radical Abundance* is a good source for Drexler's side of this story. Wikipedia and other sources, including a few pop-sci books whose names I can't remember at the moment, could be good sources for other perspectives on these events.

Luke

From: **Jonah Sinick** <jsinick@gmail.com>
Date: Mon, Jun 24, 2013 at 2:43 PM
To: Luke Muehlhauser <lukeprog@gmail.com>

Hi Luke,

I just browsed through the synopsis of Limits to Growth: The 30-year update.

I'll investigate further, but my first reaction is that the support for the authors' views is very weak. I didn't read every word of the synopsis, but it appears that they don't address substitution of resources at all. The evidence that they cite in favor of their position appears to be cherry picked.

One thing that I'll investigate further is soil depletion. I can imagine substituting agricultural land being very difficult.

Best,
Jonah

From: **Jonah Sinick** <jsinick@gmail.com>
Date: Tue, Jun 25, 2013 at 1:43 PM
To: Luke Muehlhauser <lukeprog@gmail.com>

Hi Luke,

I'm tabling *Limits to Growth* for now, and going to deep dive the Arrhenius climate change material first, because amount of information gained per unit time spent on *Limits to Growth* seems lower at the margin. I'll probably cycle back to *Limits to Growth* — just triaging for now. Let me know if you'd like me to finish *Limits to Growth* first.

Best,
Jonah

From: **Jonah Sinick** <jsinick@gmail.com>
Date: Tue, Jun 25, 2013 at 5:26 PM
To: Luke Muehlhauser <lukeprog@gmail.com>

Hi Luke,

I just read Chapters 4 and 5 of Nate Silver's book. A couple of thoughts:

- The discussion of the success of the Monte Carlo method for predicting weather caused me to update in favor the feasibility of predicting the future several decades down the road. *A priori*, I wouldn't have expected the sort of analysis that meteorologists do nowadays to have predictive power. On the other hand, they have very good feedback loops

- The history of efforts to predict earthquakes with better accuracy than the basic statistical law turning out to have worse predictive power than the basic statistical law is a cautionary tale.

More soon.

Jonah

From: **Luke Muehlhauser** <luke@intelligence.org>

Date: Fri, Jun 28, 2013 at 7:33 PM
To: Jonah Sinick <jsinick@gmail.com>

Jonah,

In the meantime, here are some additional shallow investigations you could make, to see whether any of them are worth deeper investigations:

- Asilomar Conference
- NEO deflection plans
- Cold War efforts to win decades later, e.g. math education for Russian children
- One child policy in China
- Ethically concerned scientists
- Skim the history of some transformative sci-tech breakthroughs to find instances of people trying to push on events 30 years away for social value (e.g. transistor, humane genome)
- Green revolution
- Cryptography
- Ask Nick Beckstead if he has promising cases worth looking into
- Ford Foundation set up policy foundation in India that helped India recover unusually quickly after 1991 crisis? (see relevant GiveWell conversation)
- Early climate change mitigation efforts

Luke

From: **Jonah Sinick** <jsinick@gmail.com>
Date: Mon, Jul 1, 2013 at 1:44 PM
To: Luke Muehlhauser <luke@intelligence.org>

Hi Luke,

I finished Nate Silver's book. Below I've summarized the points that I found most relevant. The first section gives high-level points, and the subsequent sections give points from individual chapters of the book.

Overview:

- The deluge of data available in the modern world has exacerbated the problem of people perceiving patterns where none exist, and overfitting predictive models to past data.

- Because of the risk of overfitting a model to past data, using a simple model can give more accurate results than using a refined model does.

- A major reason that predictions fail is model uncertainty into account. Looking at a situation from multiple different angles can be a guard against failure to give adequate weight to model uncertainty.

- Average different perspectives often yields better predictive results than using a single perspective.

- Humans have a very strong tendency toward being overconfident when making predictions.

- People make better predictions in domains where they have tight feedback loops to use to test their hypotheses.

- Sometimes people's failure to make good predictions is the result of perverse incentives.

Introduction:

Increased access to information can do more harm than good. This is because the more information is available, the easier it is for people to cherry-pick information that supports their pre-existing positions, or to perceive patterns where there are none.

The invention of the printing press may have given rise to religious wars on account of facilitating the development of ideological agendas.

Chapter 1: The failure to predict the 2008 housing bubble and recession

(a) There was an issue of people failing to take into account model uncertainty. In particular, people shouldn't have taken the forecasted 0.12% default rate of mortgage securities at face value. This rate corresponded to the rating agencies giving mortgage securities AAA ratings, which are usually reserved only the world's most solvent governments and best-run businesses.

(b) Some of the actors involved failed to look at the situation from many different angles. For example, the fact that the increase in housing prices wasn't driven by a change in fundamentals seems to have been overlooked by some people.

(c) Each individual factor that contributed to the housing bubble, and to the recession, seems like a common occurrence (e.g. perverse incentives, inadequate regulation, ignoring of tail risk, and irrational behavior coming from consumers). The severity of the situation seems to have come from the factors all being present simultaneously (by chance). Any individual factor would ordinarily be offset by other safeguards built into our social institutions.

Chapter 2: Political Predictions

(a) Political pundits and political experts usually don't do much better than chance when forecasting political events, and usually do worse than crude statistical models.

(b) Averaging individual experts' forecasts gives better forecasts than the forecasts of the average individual, with the effect size being about 15-20%.

(c) There are some experts who do make predictions that are substantially more accurate than chance.

(d) The experts who do better tend to be multidisciplinary, pursue multiple approaches to forecasting at the same time, be willing to change their minds, offer *probabilistic* predictions, and rely more on observation than on theory.

(e) Making definitive predictions that fall into a pre-existing narrative is associated with political partisanship. It's negatively correlated with making accurate predictions, but positively correlated with getting media attention. So the most visible people may make systematically worse predictions than less visible people.

(f) The failure to predict the fall of the Soviet Union seems to have arisen from a failure to integrate multiple perspectives. There were some people who were aware of Gorbachev's progressiveness and other people who recognized the dysfunctionality of the Soviet Union's economy, but these groups were largely nonoverlapping.

(g) Nate Silver integrates poll data, historical track record of poll data, information about the economy and information about the demographics of states, in order to make predictions about political elections.

(h) There's an organization called the Cook Political Report that has a very impressive track record of making accurate predictions about how political elections will go.

Chapter 3: Baseball predictions

(a) Baseball statistics constitute a very rich collection of data, and people who aspire to predict how well players will play in the future have rapid feedback loops that allow them to repeatedly test the validity of their hypotheses.

(b) A simple model of how the performance of a baseball player varies with age outperformed a much more complicated model that attempted to form a more nuanced picture of how performance varies with age by dividing players into different classes. This may have been because the latter model was over-fitted to the existing data.

Chapter 4: Weather Predictions

(a) Weather forecasters have access to a large amount of data, which offers them rapid feedback loops that allow them to repeatedly test their hypotheses.

(b) The method of predicting what would happen under certain initial conditions for many different examples of initial conditions and then averaging over the results is tantalizing. It suggests the possibility of reducing uncertainty in situations that seem hopelessly complicated to analyze, by averaging over the predictions made under different assumptions.

(c) It's impressive that the weather experts are well calibrated.

(d) Local news networks sacrifice accuracy and honesty to optimize for viewer satisfaction.

(e) The integrated use of computer models and human judgment calls does notably better than computer models alone.

(f) The human input is getting better over time.

(g) Hurricane Katrina wasn't appropriately addressed because the local government didn't listen to the weather forecasters early enough, and the local people didn't take the hurricane warning sufficiently seriously.

Chapter 5: Earthquake predictions:

(a) The Gutenberg-Richter law predicts the frequency of earthquakes of a given magnitude in a given location. One can use the frequency of earthquakes of a given magnitude to predict the frequency of earthquakes of a higher magnitude (even without having many data points).

(b) Efforts to build models that offer more precise predictions than the Gutenberg-Richter law does have been unsuccessful, apparently owing to overfitting existing data, and have generally done worse than the Gutenberg-Richter law.

Chapter 6:

(a) Communicating a prediction of the median case without giving a confidence interval can be very pernicious, because outcomes can be highly sensitive to error.

(b) Economists have a poor track record of predicting GDP growth. There's so much data pertaining to factors that might drive GDP growth that it's easy to perceive patterns that aren't real.

(c) The economy is always changing, and often past patterns don't predict the future patterns.

(d) Prediction markets for GDP growth might yield better predictions than economists' forecasts do. But existing prediction markets aren't very good.

Chapter 7: Disease Outbreaks

(a) Predictions can be self-fulfilling (e.g. in election primaries races) or self-canceling (e.g. when disease outbreaks are predicted, measures can be taken to prevent them, which can nullify the prediction).

Chapter 8: Bayes' Theorem

(a) When gauging the strength of a prediction, it's important to view the inside view in the context of the outside view. For example, most medical studies that claim 95% confidence aren't replicable, so one shouldn't take the 95% confidence figures at face value.

Chapter 9: Chess computers

(a) Our use of prediction heuristics makes us vulnerable to opponents who are aware of the heuristics that we're using and who can therefore act in unexpected ways that we're not prepared for.

Chapter 10: Poker

(a) Elite poker players use Bayesian reasoning to estimate the probability of a hand based on the cards on the table, contingent on opponents' behavior.

(b) Elite poker players also additional information, such as the fact that women end to play more conservatively than men do, in order to refine their predictions about what cards the opponent has

(c) Often the 80%/20% rule applies to getting good at predictions relative to what's in principle possible. A relatively small amount of effort can result in large improvements. In competitive contexts such as poker, serious players have all already put this amount of effort in, so beating them requires further effort. But in arenas such as election results predictions, where not many people are trying hard, it's possible to do a lot better than most people do with relatively little effort.

Chapter 11: The stock market

(a) It's difficult to distinguish signal from noise when attempting to predict the stock market.

(b) There are some systematic patterns in the stock market. For example, between 1965 and 1975, rises in stock prices one day were correlated with rises in stock prices the next day. But such patterns are rapidly exploited once people recognize them, and disappear.

(c) It's not so hard to predict a stock market bubble. One can look at the average price to earnings ratio across all stocks, and when it's sufficiently high, that's a signal that there's a bubble.

(d) It's hard to predict when a bubble is about to pop.

(e) Most investors are relatively shortsighted. This is especially the case because most investors are investing other people's money rather than their own.

(f) There are incentives not to short-sell stocks too much, both cultural and legal. This may give rise to a market inefficiency.

(g) An 1970 investment of $10k in S&P 500 would have yielded $63k in profit in 2009, but if one adopted the strategy of pulling money out when the market dropped by 25% and putting it back in when it had recovered to 90% of its earlier price, the profit would only be $18k. Many investors behave in the latter fashion.

Chapter 12: Climate change

(a) There's a lot of uncertainty around climate change predictions: there's uncertainty about the climate models, uncertainty about the initial conditions, and uncertainty about society's ability to adapt.

(b) There may be global *cooling* coming from sulfer emissions

(c) The amount of uncertainty can easily justify focus on mitigating climate change, because of the risk of the problem being worse than expected entailing more potential negative consequences than the consequences in the median case.

(d) A simple regression analysis looking at the correlation between CO2 levels and temperature may give a better predictive model than more sophisticated climate models.

Chapter 13: Terrorism

(a) Governments often prepare for terrorist attacks, but often prepare for the wrong kinds of terrorist attacks, unaware of bigger threats.

(b) The September 11th secnario hadn't been considered and rejected, but rather, hadn't been considered at all.

(c) If one looks at number of terrorist attacks as a function of their magnitude, they seem to obey a power law.

(d) There are some reasons to be concerned about the possibility of a nuclear weapon terrorist attack, or bioterrorism, in the United States. Such an attack could kill over a million people

----
From: **Jonah Sinick** <jsinick@gmail.com>
Date: Mon, Jul 1, 2013 at 3:52 PM
To: Luke Muehlhauser <luke@intelligence.org>

Hi Luke,

Regarding China's One-Child Policy:

It seems difficult to estimate the sign of the impact on China's welfare, and (to a lesser degree) the size of the impact on population. Some points:

- The policy was adopted in response to a sharp drop-off in child mortality rates.

- Wikipedia says that the original motivation for the policy was concern about natural resource shortage. I don't know whether or not this concern was well grounded, or whether this was the real reason for the policy.

- Fertility dropped from 5.9 children in 1970 to 2.63 children in 1980 (a big drop!). Since then, fertility has dropped to 1.63 children. This may not be so much larger than the counterfactual drop. The expected impact of the policy on population is reduced by parents having taken various measures to get around the policy.

- The effects of the policy are numerous, and the expected value of the policy unclear. Some relevant points:

  • In the counterfactual scenario, the country may not have had the capacity to adapt to a very rapid influx of children.
  • When people have fewer children, they have more money to invest, and investing contributes to economic growth. They also have more time and money to invest in their own children.
  • It's been claimed that increased population spurs innovation, even in poor societies
  • There were human rights violations, which gave rise to political instability
  • Human capital was wasted on efforts to circumvent the policy and efforts to enforce the policy.
  • The policy resulted in selective breeding for males, reducing males' future prospects for finding wives.

There doesn't seem to be a decisive case for the policy having been good, or having been bad. I would guess that it was bad on net, but that's based on very superficial considerations.

One could investigate further, but it seems likely that the situation would remain murky regardless. For now, I'll just note the example as a targeted intervention that had expected value of ambiguous sign.

Best,
Jonah

From: **Jonah Sinick** <jsinick@gmail.com>
Date: Mon, Jul 1, 2013 at 4:08 PM
To: Luke Muehlhauser <luke@intelligence.org>

Responding on

Cold War efforts to win decades later, e.g. math education for Russian children

I took a look at this, and I think that it's too hard to tie these efforts to war outcomes for this to be fruitful to investigate further.

I would guess that education and research & development efforts coming out of the Cold War had high humanitarian value (as they can be thought of as constituting fungible investment in the future). But

(a) Probably a lot of the benefits weren't mediated through impact on the Cold War
(b) Probably it's hard to tie any specific effort to impact on the Cold War (at least in expectation)

From: **Jonah Sinick** <jsinick@gmail.com>
Date: Mon, Jul 1, 2013 at 4:49 PM
To: Luke Muehlhauser <luke@intelligence.org>

Regarding Kaj Sotala's post on ethically concerned scientists:

- One class of examples that he discusses is that of scientists concealing their discoveries out of concern that they would be used for military purposes. These don't seem relevant to the question of present day attempts to predict and alter things ~30 years from now. It could be that Napier and Boyle postponed the development of weapons by 30+ years, but if this is the case, it seems very much contingent on them having been alive at a time when there were very few scientific researchers. Later examples are of scientists who lived in times with much more efficient scientific markets, where concealing one's discoveries wouldn't have a long-term impact relative to the counterfactual.

- Another class of examples that he discusses is advocacy efforts to reduce the use of dangerous weapons. These don't seem to constitute examples of trying to influence the future decades out: there was eminent danger of these weapons being used at the time of advocacy.

- I'll investigate the case of Recombinant DNA.

- The case of Weiner writing Some Moral and Technical Consequences of Automation might constitute an example. Should I investigate further?

From: **Luke Muehlhauser** <lukeprog@gmail.com>
Date: Mon, Jul 1, 2013 at 4:57 PM
To: Jonah Sinick <jsinick@gmail.com>

Yes, investigating Weiner more closely sounds good. Off the top of my head, I don't know whether he actually *did* anything specific with his concerns about the future of AI, but he was clearly one of the first people writing about the problem. One of Weiner's statements would have made Yudkowsky proud: "The machines will do what we ask them to do and not what we ought to ask them to do."

Luke

From: **Jonah Sinick** <jsinick@gmail.com>
Date: Sun, Jul 7, 2013 at 10:14 PM
To: Luke Muehlhauser <lukeprog@gmail.com>

Hi Luke,

I read most of *The Limits to Growth*.

Having read it, my impression of the book is very different from what it was before I read it. Both critics and advocates seem to have significantly misrepresented it. So I'm very glad that I read it.

Some points:

- The authors make very few, if any, strong claims. Their core claim *if* exponential growth of resource usage continues, *then* there will likely be a societal collapse by 2100. They're very explicit about not making specific predictions, and their remarks carry a strong connotation that the scenarios that they work out are intended as thought experiments.

- Overall, I'm very impressed by the epistemic standard that the book meets. The authors are generally careful to qualify their claims as appropriate. The book doesn't look naive even in retrospect, which is impressive given that it was written 40 years ago.

- My sense is that the book's reputation was tarnished because other people writing in the same space weren't nearly as careful as the authors.

- The authors do discuss substitutability at length in Chapter 4. They argue that:

  (a) Even if we had unlimited resources, exponential growth of resource use would still result in exponential growth of pollution, unless pollution is curbed

(b) The marginal cost of reducing pollution grows astronomically as the fraction that you want to reduce it by tends toward zero, so that it might not be possible to quell the increase in pollution coming from an exponential increase in usage.

(c) A sufficiently large increase in pollution could lead to societal collapse.

- The authors discuss potential ways to mitigate the problems that they identify, but at a theoretical level. They recognize the complexity of the issues, and don't make policy recommendations.

- There may have been examples of people trying to implement policies based on the book and surrounding literature. I'll spend a little time investigating this and let you know what I find.

- The book's discussion of the general issue of society needing to adapt to large-scale emergent problems has some relevance to MIRI's mission beyond the question of predicting the future in general. I can compile relevant excerpts.

Best,
Jonah

From: **Jonah Sinick** <jsinick@gmail.com>
Date: Mon, Jul 8, 2013 at 12:15 PM
To: Luke Muehlhauser <lukeprog@gmail.com>

In line with my previous email, from an article in *The Nation*:

Lomborg and others have accused Limits of hysteria. The book is not hysterical. It was, however, taken up by a social milieu heavily populated by hysterics like the peak oilers and the (sometimes crypto-racist) population bugs. Paul Ehrlich embodies this most clearly: for almost fifty years, he's been pushing a simple-minded Malthusian condemnation of population growth. In the late 1960s, with world population at 3.5 billion, Ehrlich said in a TV interview: "Sometime in the next fifteen years, the end will come. And by 'the end' I mean an utter breakdown of the capacity of the planet to support humanity." Limits, meanwhile, never made such an outlandish claim.

From: **Jonah Sinick** <jsinick@gmail.com>
Date: Mon, Jul 22, 2013 at 1:12 PM
To: Luke Muehlhauser <lukeprog@gmail.com>

Hi Luke,

Following up on Norbert Wiener's predictions about dangers of automation:

I read Wiener's 1960 paper in which he discusses the danger of people programming machines in ways that result in unintended destructive behaviors that are executed too quickly for humans to figure out what's going on and stop them. He doesn't make sufficiently specific predictions for it to be possible to assess their veracity. I wasn't able to find subsequent writings by him on this subject.

Perhaps more interesting is Wiener's predictions of automation displacing low skilled workers, rendering them unemployable, and his advocacy attempts to stop this. I found a paper titled Some Notes on Wiener’s Concerns about the Social Impact of Cybernetics, the Effects of Automation on Labor, and “the Human Use of Human Beings”, which summarizes this. Some points:

- Wiener believed that unless countermeasures were taken, automation would render low skilled workers unemployable. He believed that this would precipitate an economic crisis far worse than that of the Great Depression. He viewed this potential outcome as highly undesirable.

  He considered giving top professional priority to disseminating his views on this point.

- He attempted to network with labor unions to lobby against automation replacing human labor. Union leaders were unresponsive during the mid 40's, but in 1949 he communicated with Walter Reuther, the president of the United Automobile Workers, which was America's largest and most powerful union, and Reuther was more responsive.

  In the beginning of the 1960's, many factory workers were laid off because automatic factories could replace them. These workers may have been reemployed in a different capacity in short order.

- Wiener believed that society was increasingly adopting a focus on economic value to the exclusion of other considerations, and he believed that this focus was detrimental to humanitarian values.

- Based on the summary, Wiener appears to have been naive in that he doesn't seem to have had an appreciation of equilibrating influences in markets. I'm reminded of Robin Hanson's post Eventual Futures where Robin writes

  [People] project forward a current trend to an extreme, while assuming other things don’t change much, and then recommend an action which might make sense if this extreme change were to happen all at once soon. This is usually a mistake. The trend may not continue indefinitely. Or, by the time a projected extreme is reached, other changes may have changed the appropriate response. Or, the best response may be to do nothing for a long time, until closer to big consequences. Or, the best response may be to do nothing, ever – not all negative changes can be profitably resisted.

  Wiener also seems to have underestimated the value to the poor coming from economic growth and automation.

- While it's possible that a closer look at Wiener's writings would reveal more nuanced and sophisticated views, I'm happy with closing the investigation of this subject and summarizing by saying that based on what I've read:

  (i) Wiener had very strong views on the subject.
  (ii) Wiener doesn't seem to have updated very much in response to incoming evidence.
  (iii) It's unclear that he made falsifiable predictions, and in particular, whether he would believe

> that they had been born out to date, or whether he would believe that they would be born out in the future.
> (iv) At least with the benefit of hindsight, his views seem naive.
> (v) It appears that he was using one relatively strong argument style (which I described to be characteristic of mathematicians) and wasn't viewing the situation from many different angles, apparently being a "hedgehog" in this domain (in Tetlock's taxonomy).
> (vi) The case study lends support to Tetlock's idea that "hedgehogs" are bad at predicting the future.

From: **Jonah Sinick** <jsinick@gmail.com>
Date: Wed, Aug 14, 2013 at 2:36 PM
To: Luke Muehlhauser <lukeprog@gmail.com>

Hi Luke,

Responding on the point of the Ford Foundation's think tank role in averting India's 1991 financial crisis:

I wrote to Lant Pritchett and asked him if there are papers that discuss this, and he said that he doesn't know of any, and that his own knowledge of it comes from personal contact with some of the people involved. He referred me to one researcher who might know of papers discussing it, and I wrote to her, but didn't hear back.

Best,
Jonah

From: **Jonah Sinick** <jsinick@gmail.com>
Date: Wed, Aug 14, 2013 at 2:46 PM
To: Luke Muehlhauser <lukeprog@gmail.com>

Hi Luke,

I didn't investigate the situation with the Asilomar Conference on Recombinant DNA enough to have something substantive to say about it. The extent of my reading was the Wikipedia article and the references cited therein.

Jonah

From: **Jonah Sinick** <jsinick@gmail.com>
Date: Wed, Aug 14, 2013 at 3:04 PM
To: Luke Muehlhauser <lukeprog@gmail.com>

Hi Luke,

On the point of early climate change mitigation efforts:

The most relevant reference that I found (in my brief search) is an article titled Impacts of Climate Change from a website that's a companion to the book The Discovery of Global Warming.

Best,
Jonah

From: **Jonah Sinick** <jsinick@gmail.com>
Date: Wed, Aug 14, 2013 at 3:10 PM
To: Luke Muehlhauser <lukeprog@gmail.com>

Hi Luke,

I didn't investigate any of

- Asteroid strike deflection efforts
- Possible deliberate long term efforts to produce revolutionary scientific technologies.
- Long term computer security research

or read *Expert Political Judgment: How Good Is It? How Can We Know?* by Philip Tetlock.

I've now addressed all of the potential examples that we generated and that I listed in my LW discussion thread post titled "Can we know what to do about AI?": An Introduction."

Jonah