SSC Journal Club: AI Timelines

Posted on June 8, 2017 by Scott Alexander

I

A few years ago, Muller and Bostrom et al surveyed AI researchers to assess their opinion on AI progress and superintelligence. Since then, deep learning took off, AlphaGo beat human Go champions, and the field has generally progressed. I've been waiting for a new survey for a while, and now we have one.

Grace et al (*New Scientist* article, paper, see also the post on the author's blog AI Impacts) surveyed 1634 experts at major AI conferences and received 352 responses. Unlike Bostrom's survey, this didn't oversample experts at weird futurist conferences and seems to be a pretty good cross-section of mainstream opinion in the field. What did they think?

Well, a lot of different things.

The headline result: the researchers asked experts for their probabilities that we would get AI that was "able to accomplish every task better and more cheaply than human workers". The experts

thought on average there was a 50% chance of this happening by 2062 – and a 10% chance of it happening by 2026!

But on its own this is a bit misleading. They also asked by what year "for any occupation, machines could be built to carry out the task better and more cheaply than human workers". The experts thought on average that there was a 50% chance of this happening by 2139, and a 20% chance of it happening by 2037.

As the authors point out, these two questions are basically the same – they were put in just to test if there was any framing effect. The framing effect was apparently strong enough to shift the median date of strong human-level AI from 2062 to 2139. This makes it hard to argue AI experts actually have a strong opinion on this.

Also, these averages are deceptive. Several experts thought there was basically a 100% chance of strong AI by 2035; others thought there was only a 20% chance or less by 2100. This is less "AI experts have spoken and it will happen in 2062" and more "AI experts have spoken, and everything they say contradicts each other and quite often themselves".

This *does* convey more than zero information. It conveys the information that AI researchers are *really unsure*. I can't tell you how many people I've heard say "there's no serious AI researcher who thinks there's any chance of human-level intelligence before 2050". Well actually, there are a few dozen conference-paper-presenting experts who think there's a *one hundred* percent chance of human-level AI before that year. I don't know what drugs they're on,

but they exist. The moral of the story is: be less certain about this kind of thing.

II

The next thing we can take from this paper is a timeline of what will happen when. The authors give a bunch of different tasks, jobs, and milestones, and ask the researchers when AI will be able to complete them. Average answers range from nearly fifty years off (for machines being able to do original high-level mathematical research) to only three years away (for machines achieving the venerable accomplishment of being able to outperform humans at *Angry Birds*). Along the way they'll beat humans at poker (four years), writing high school essays (ten years), be able to outrun humans in a 5K foot race (12 years), and write a New York Times bestseller (26 years). What do these AI researchers think is the hardest and most quintessentially human of the tasks listed, the one robots will have the most trouble doing because of its Olympian intellectual requirements? That's right – AI research (80 years).

I make fun of this, but it's actually interesting to think about. Might the AI researchers have put their own job last not because of an inflated sense of their own importance, but because they engage with it every day in Near Mode? That is, because they imagine writing a New York Times bestseller as "something something pen paper be good with words okay done" whereas they understand the complexity of AI research and how excruciatingly hard it would be to automate away every piece of what they do?

Also, since they rated AI research (80 years) as the hardest of all occupations, what do they mean when they say that "full automation of all human jobs" is 125 years away? Some other job not on the list that will take 40 years longer than AI research? Or just a combination of framing effects and not understanding the question?

(it's also unclear to what extent they believe that automating AI research will lead to a feedback loop and subsequent hard takeoff to superintelligence. This kind of theory would fit with it being the last job to be automated, but not with it taking another forty years before an unspecified age of full automation.)

III

The last part is the most interesting for me: what do AI researchers believe about risk from superintelligence?

This is very different from the earlier questions about timelines. It's possible to believe that AI will come very soon but be perfectly safe. And it's possible to believe that AI is a long time away but we really need to start preparing now, or else. A lot of popular accounts collapse these two things together, "oh, you're worried about AI, but that's dumb because there's no way it's going to happen anytime soon", but past research has shown that short timelines and high risk assessment are only modestly correlated. This survey asked about both separately.

There were a couple of different questions trying to get at this, but it looks like the most direct one was "does Stuart Russell's argument for why highly advanced AI might pose a risk, point at an important problem?". You can see the exact version of his argument quoted in the survey on the AI Impacts page, but it's basically the standard Bostrom/Yudkowsky argument for why AIs may end up with extreme values contrary to our own, framed in a very normal-sounding and non-threatening way. According to the experts, this was:

No, not a real problem	11%
No, not an important problem	19%
Yes, a moderately important problem	31%
Yes, an important problem	34%
Yes, among the most important problems in the field	5%

70% of AI experts agree with the basic argument that there's a risk from poorly-goal-aligned AI. But very few believe it's among "the most important problems in the field". This is pretty surprising; if there's a good chance AI could be hostile to humans, shouldn't that automatically be pretty high on the priority list?

The next question might help explain this: "Value of working on this problem now, compared to other problems in the field?"

Much less valuable	22%

Less valuable	41%
As valuable as other problems	28%
More valuable	7%
Much more valuable	1.4%

So charitably, the answer to this question was coloring the answer to the previous one: AI researchers believe it's plausible that there could be major problems with machine goal alignment, they just don't think that there's too much point in working on it now.

One more question here: "Chance intelligence explosion argument is broadly correct?"

Quite likely (81-100% chance)	12%
Likely (61-80% chance)	17%
About even (41-60% chance)	21%
Unlikely (21-40% chance)	24%
Quite unlikely (0-20% chance)	26%

Splitting the 41-60% bin in two, we might estimate that about 40% of AI researchers think the hypothesis is more likely than not.

Take the big picture here, and I worry there's sort of a discrepancy.

50% of experts think there's at least a ten percent chance of above-human-level AI coming within the next ten years.

And 40% of experts think that there's a better-than-even chance that, once we get above-human level AI, it will "explode" to suddenly become vastly more intelligent than humans.

And 70% of experts think that Stuart Russell makes a pretty good point when he says that without a lot of research into AI goal alignment, AIs will probably have their goals so misaligned with humans that they could become dangerous and hostile.

I don't have the raw individual-level data, so I can't prove that these aren't all anti-correlated in some perverse way that's the opposite of the direction I would expect. But if we assume they're not, and just naively multiply the probabilities together for a rough estimate, that suggests that about 14% of experts believe that all three of these things: that AI might be soon, superintelligent, and hostile.

Yet only a third of these – 5% – think this is "among the most important problems in the field". Only a tenth – 1.4% – think it's "much more valuable" than other things they could be working on.

IV

How have things changed since Muller and Bostrom's survey in 2012?

The short answer is "confusingly". Since almost everyone agrees that AI progress in the past five years has been much faster than

expected, we would expect experts to have faster timelines – ie expect AI to be closer now than they did then. But Bostrom's sample predicted human-level AI in 2040 (median) or 2081 (mean). Grace et al don't give clear means or medians, preferring some complicated statistical construct which isn't exactly similar to either of these. But their dates – 2062 by one framing, 2139 by another – at least seem potentially a little bit later.

Some of this may have to do with a subtle difference in how they asked their question:

Bostrom: "Define a high-level machine intelligence as one that can carry out most human professions as well as a typical human…"

Grace: "High-level machine intelligence is achieved when unaided machines can accomplish every task better and more cheaply than human workers."

Bostrom wanted it equal to humans; Grace wants it better. Bostrom wanted "most professions", Grace wants "every task". It makes sense that experts would predict longer timescales for meeting Grace's standards.

But as we saw before, expecting AI experts to make sense might be giving them too much credit. A more likely possibility: Bostrom's sample included people from wackier subbranches of AI research, like a conference on Philosophy of AI and one on Artificial General Intelligence; Grace's sample was more mainstream. The most

mainstream part of Bostrom's sample, a list of top 100 AI researchers, had an estimate a bit closer to Grace's (2050).

We can also compare the two samples on belief in an intelligence explosion. Bostrom asked how likely it was that AI went from human-level to "greatly surpassing" human level within two years. The median was 10%; the mean was 19%. The median of top AI researchers not involved in wacky conferences was 5%.

Grace asked the same question, with much the same results: a median 10% probability. I have no idea why this question – which details what an "intelligence explosion" would entail – was so much less popular than the one that used the words "intelligence explosion" (remember, 40% of experts agreed that "the intelligence explosion argument is broadly correct"). Maybe researchers believe it's a logically sound argument and worth considering but in the end it's not going to happen – or maybe they don't actually know what "intelligence explosion" means.

Finally, Bostrom and Grace both asked experts' predictions for whether the final impact of AI would be good or bad. Bostrom's full sample (top 100 subgroup in parentheses) was:

Extremely good	24%	(20)
On balance good	28%	(40)
More or less neutral	17%	(19)
On balance bad	3%	(13)

Extremely bad – existential catastrophe	18%	(8)

Grace's results for the same question:

Extremely good	20%
On balance good	25%
More or less neutral	40%
On balance bad	10%
Extremely bad – human extinction	5%

Grace's data looks pretty much the same as the TOP100 subset of Bostrom's data, which makes sense since both are prestigious non-wacky AI researchers.

V

A final question: "How much should society prioritize AI safety research"?

Much less	5%
Less	6%
About the same	41%
More	35%
Much more	12%

People who say that real AI researchers don't believe in safety research are now just empirically wrong. I can't yet say that most of them want more such research – it's only 47% on this survey. But next survey AI will be a little bit more advanced, people will have thought it over a little bit more, and maybe we'll break the 50% mark.

But we're not there yet.

I think a good summary of this paper would be that large-minorities-to-small-majorities of AI experts agree with the arguments around AI risk and think they're worth investigating further. But only a very small minority of experts consider it an emergency or think it's really important right now.

You could tell an optimistic story here – "experts agree that things will probably be okay, everyone can calm down".

You can also tell a more pessimistic story. Experts agree with a lot of the claims and arguments that suggest reason for concern. It's just that, having granted them, they're not *actually* concerned.

This seems like a pretty common problem in philosophy. "Do you believe it's more important that poor people have basic necessities of life than that you have lots of luxury goods?" "Yeah" "And do you believe that the money you're currently spending on luxury goods right now could instead be spent on charity that would help poor people get life necessities?" "Yeah." "Then shouldn't you stop buying luxury goods and instead give all your extra money be-

yond what you need to live to charity?" "Hey, what? Nobody does that! That would be a lot of work and make me look really weird!"

How many of the experts in this survey are victims of the same problem? "Do you believe powerful AI is coming soon?" "Yeah." "Do you believe it could be really dangerous?" "Yeah." "Then shouldn't you worry about this?" "Hey, what? Nobody does that! That would be a lot of work and make me look really weird!"

I don't know. But I'm encouraged to see people are even taking the arguments seriously. And I'm encouraged that researchers are finally giving us good data on this. Thanks to the authors of this study for being so diligent, helpful, intelligent, wonderful, and (of course) sexy.

(I might have forgotten to mention that the lead author is my girlfriend. But that's not biasing my praise above in any way.)