Analysis of Müller and Bostrom 2016

Allan Dafoe

September 26, 2016

Analysis for Allan Dafoe and Stuart Russell "Yes. A response to Etzioni's article, 'Are the Experts Worried About the Existential Risk of Artificial Intelligence?' "

All files related to this analysis can be found on dataverse here:
http://dx.doi.org/10.7910/DVN/JHR1GX

Müller and Bostrom paper is here: www.nickbostrom.com/papers/survey.pdf

Müller and Bostrom data is here: http://www.pt-ai.org/polls/experts

Research Assistance by Baobao Zhang

Timelines

Etzioni's Results

Etzioni asked his respondents the following:
"In his book, Nick Bostrom has defined Superintelligence as 'an intellect that is much smarter than the best human brains in practically every field, including scientific creativity, general wisdom and social skills.' When do you think we will achieve Superintelligence?"

Etzioni then provided the experts with four answer options: "In the next 10 years" (0% selected this), "In the next 10-25 years" (7.5%), "In more than 25 years" (67.5%), and "Never" (25%).

Commentary on Etzioni's Results

From this Etzioni draws his main inference that "In essence, according to 92.5 percent of the respondents, superintelligence is beyond the foreseeable horizon."

As we discuss, we think there are problems with this interpretation. First, and most importantly, it declares >25 years beyond the foreseeable horizon and thus as not worth worrying about. By that logic we should not worry about climate change, retirement, etc...

Second, we see that 7.5% of these experts actually thought that *superintelligence* would be achieved within 25 years. Given that superintelligence would be a world changing achievement, that 7.5% think it will arrive so soon is noteworthy.

Third, though Etzioni's question seems intuitively straightforward, it is actually deeply ambiguous. Should the reader interpret this as asking them by when they believe there is a greater than 50% chance? A greater than 90% chance? A greater than 99% chance? Or a 100% chance? As we know from theory and empirics, these different questions ask about very different points on a person's probability distribution over the future, and lead to answers that differ on the order of decades.

Timeline for Müller and Bostrom's Top 100 AI researchers

Müller and Bostrom (2016) surveyed the top 100 cited AI researchers; 29 responded.

They ask the Top100 the following:

"Define a 'high–level machine intelligence' (HLMI) as one that can carry out most human professions at least as well as a typical human."

"For the purposes of this question, assume that human scientific activity continues without major negative disruption. By what year would you see a (10% / 50% / 90%) probability for such HLMI to exist?"

The descriptive statistics for their answers are as follows, written in order of 10%, 50%, and 90%:

```
summary(d$Year.10.[r])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2012    2020    2025    2050    2030    2508
```

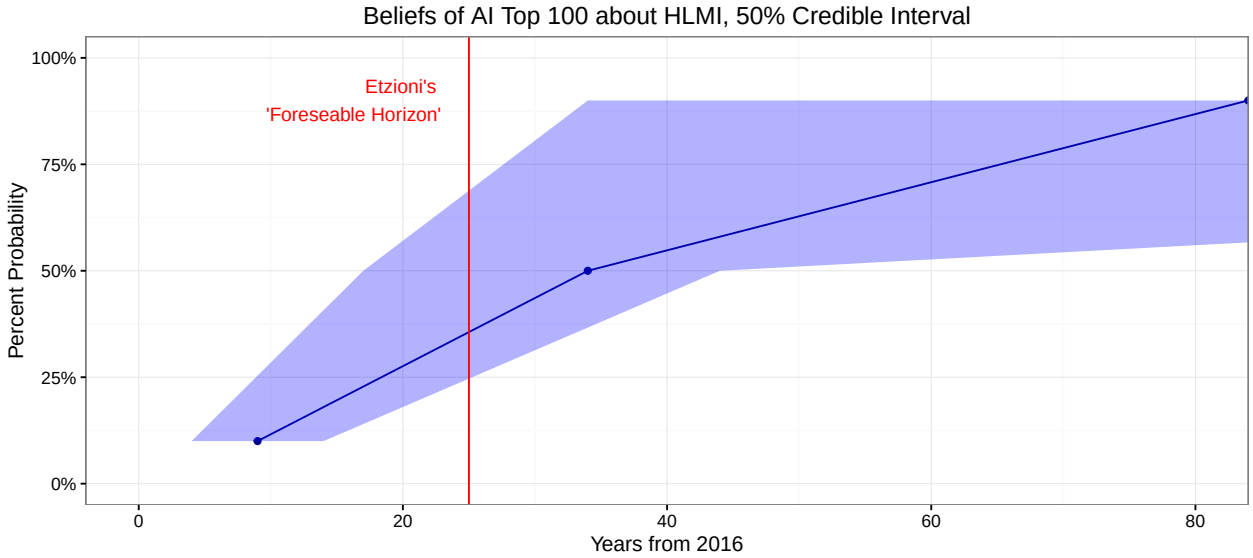
```
summary(d$Year.50.[r])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2017    2033    2050    2104    2060    3000
```

```
summary(d$Year.90.[r])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2025    2050    2100    2754    2300    5000
```

We can see these responses using a plot of the cumulative distribution function for the beliefs of the Top 100 about High Level Machine Intelligence existing, over time.



The blue line depicts the median belief. Left edge of the "credible interval" depicts the beliefs of the 25th percentile, right edge the 75th percentile. If each expert is as likely to be correct as another, we can interpret this figure as a 50% credible interval for the arrival time of HLMI. The red line depicts Etzioni's "Foreseeable Horizon", beyond which he states we need not worry.

Comparison Between Etzioni's, and Müller and Bostrom's Results

Etzioni found that 92.5% of respondents selected >25 years for the arrival of superintelligence. As is clear from the above figure, the Top 100 also put a lot of probability mass on >25 years. Specifically, we know

that 55% of the sample was at least 50% confident that HLMI would not arrive within 25 years, and 86% of the sample was at least 10% confident that HLMI would not arrive within 25 years. Given that HLMI is a much lower technical bar than superintelligence, these numbers are entirely consistent with Etzioni's results. (This is the basis for our claim that "more than half of Bostrom's respondents gave dates beyond 25 years for a mere 50% probability of achieving mere human-level intelligence.")

Etzioni reported that 25% of his sample selected "never" for when superintelligence will be achieved. Similarly, 24% of Müller and Bostrom Top 100 thought there was at least a 10% chance that HLMI would not have been achieved within 300 years. If respondent's interpreted Etzioni's question asking about when **we will achieve** superintelligence as asking for at least 90% confidence, then we see that Müller and Bostrom's results are entirely consistent with Etzioni's.

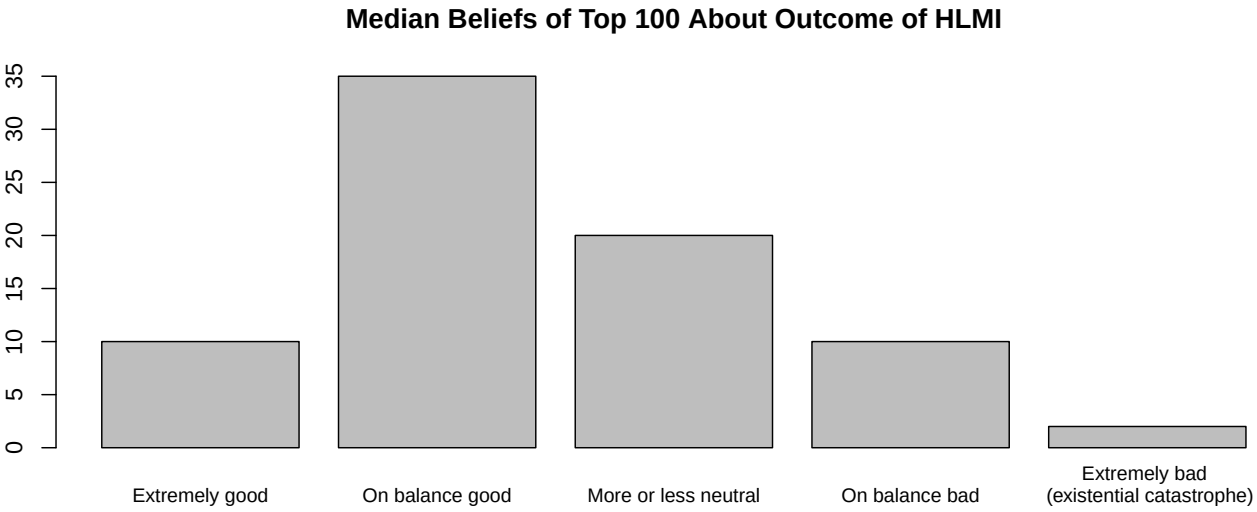
Is Superintelligence a Threat to Humanity?

Etzioni concluded that "few believe AI is a threat to humanity", though he did not ask any questions about this.

To provide data on this question, we turn to the following question asked by Müller and Bostrom:

"Assume for the purpose of this question that such HLMI will at some point exist. How positive or negative would be overall impact on humanity, in the long run? Please indicate a probability for each option. (The sum should be equal to 100%.)"

The five options were: "Extremely good – On balance good – More or less neutral – On balance bad – Extremely bad (existential catastrophe)"



Thus, most of the Top 100 experts thought that high level machine intelligence is likely to be good or very good for humanity, as do we, moreso to the extent that these issues are taken seriously. At least half of the Top 100 thought there was at least a 15% chance of an "on balance bad" or "extremely bad (existential catastrophe)" outcome. A 2016 (unpublished) survey of recently published AI researchers has found similar results (with a 5% median prediction on existential catastrophe).

Misc

We should also clear up one apparent misunderstanding regarding the survey populations. Etzioni says that Bostrom conducted "four different surveys of groups such as participants in a conference called 'Philosophy and Theory of AI' ... and members of the Greek Association for Artificial Intelligence." Etzioni's own survey,

in contrast, was of Fellows of AAAI, described as "leading researchers in the field ... who are recognized as having made significant, sustained contributions." Inexplicably, he neglects to mention that Bostrom's fourth survey was of the top 100 most cited living AI researchers, or that the results of all four surveys were largely similar.