Predicting AGI: What can we say when we know so little?

Fallenstein, Benja Mennen, Alex

December 2, 2013

(Working Paper)

1 Time to taxi

Our situation now looks fairly similar to our situation 20 years ago with respect to AGI. Right now, there's been a lot of exciting recent progress in AI, suggesting to some that strong AI seems reachable sometime soon. 20 years ago, there was a lot of exciting recent progress in AI, suggesting to some that strong AI seems reachable sometime soon. Roughly speaking, we have been in that same epistemic situation since the Dartmouth Conference in 1956.

This analysis does not attempt to predict when AGI will actually be achieved, but instead, to predict when this epistemic state with respect to AGI will change, such that we will have a clear idea of how much further progress is needed before we reach AGI. Metaphorically speaking, instead of predicting when AI takes off, we predict when it will start taxiing to the runway. This taxiing could consist of anything from "AGI produced with no prior warning" to "AGI proven impossible" (though the latter seems unlikely), with intermediate possibilities such as "research demonstrates that AGI can be produced with only a few straightforward advances, which are similar to problems that have historically taken about 10 years to solve".

Taxiing does not mean achieving some particular narrow AI ability. There have been many predictions by AI experts that certain narrow AI abilities would be quickly followed by Strong AI, but in many of those cases, the narrow AI ability has since been achieved, and we don't appear to be on our way to Strong AI. For example, I.J. Good predicted that when an AI could play chess at a Grandmaster level, we would be on the verge of Strong AI. AIs have been able to beat Grandmasters at chess for 15 years now. We have also seen many advancements that look like they could

plausibly have enabled AGI, such as back-propagation, deep learning algorithms, and the realization that it would help for AIs to be able to use probabilistic reasoning. Instead of referring to a specific milestone, taxiing means that the AI researchers at the time can see a clear path forward that will lead to AGI.

2 What is a Pareto distribution?

The Pareto distribution has 2 free variables: x_m, and α. If a random variable is distributed according to a Pareto distribution, the probability that the variable takes a value less than x_m is 0, and for $x > x_m$, the probability that the value is less than x is $1 - \left(\frac{x_m}{x}\right)^{\alpha}$. For our purposes, the random variable being described is the time that we start taxiing to AGI. 0 represents the Dartmouth Conference, x_m represents right now, and the exponent α depends on how quickly we expect AGI to arrive.

One might quite reasonably ask what is special about the Pareto distribution. To answer that, we will have to introduce the concept of a hazard rate, which represents how likely the event is to occur imminently at any point in time. The hazard rate of a distribution at a point x is the probability density at x conditional on the event not occurring before x (more formally: $\lambda(x) = \lim_{\varepsilon \to 0^+} \frac{P(x < X < x + \varepsilon)}{\varepsilon \cdot P(X > x)}$, where X is the random variable describing the distribution). If you assume that the hazard rate is constant, then you get an exponential distribution, where the cumulative probability function is $1 - e^{-\lambda x}$. If you assume that the hazard rate is inversely proportional to the time elapsed (i.e. $\lambda(ax) = \frac{\lambda(x)}{a}$), then you get a Pareto distribution, where α is the proportionality constant ($\alpha = x\lambda(x)$). α can be thought of as a "relative hazard rate", and this concept can be generalized to arbitrary distributions as $\alpha(x) = x\lambda(x)$.

Here are some tables showing the probabilities predicted by Pareto distributions with various α that we will taxi within certain time ranges, with current time since Dartmouth rounded to 60 years.

Years from Dartmouth	≤61	≤65	≤70	≤75	≤80	≤90	≤100	≤120
Years from now	≤1	≤5	≤10	≤15	≤20	≤30	≤40	≤60
$Pr(\alpha = .5)$	0.82%	3.9%	7.4%	11%	13%	18%	23%	29%
$Pr(\alpha = .67)$	1.1%	5.2%	9.8%	14%	17%	24%	29%	37%
$Pr(\alpha = .75)$	1.2%	5.8%	11%	15%	19%	26%	32%	41%
$Pr(\alpha = 1)$	1.6%	7.7%	14%	20%	25%	33%	40%	50%
$Pr(\alpha = 1.5)$	2.4%	11%	21%	28%	35%	46%	54%	65%
$Pr(\alpha = 2)$	3.3%	15%	27%	36%	44%	56%	64%	75%
$Pr(\alpha = 3)$	4.8%	21%	37%	49%	58%	70%	78%	88%

Years from Dartmouth	≥120	≥150	≥180	≥240	≥600	≥1200	≥6000
Years from now	≥60	≥90	≥120	≥180	≥540	≥1140	≥5940
$Pr(\alpha = .5)$	71%	63%	58%	50%	32%	22%	10%
$Pr(\alpha = .67)$	63%	54%	48%	40%	22%	14%	4.6%
$Pr(\alpha = .75)$	59%	50%	44%	35%	18%	11%	3.2%
$Pr(\alpha = 1)$	50%	40%	33%	25%	10%	5.0%	1.0%
$Pr(\alpha = 1.5)$	35%	25%	19%	12%	3.2%	1.1%	0.10%
$Pr(\alpha = 2)$	25%	16%	11%	6.2%	1.0%	0.25%	<.1%
$Pr(\alpha = 3)$	12%	6.4%	3.7%	1.6%	0.10%	<.1%	<.1%

3 Models that result in a Pareto distribution

Why should we use a Pareto distribution to model our uncertainty about time to taxi? We have no single very compelling answer for that, but here are several weak arguments.

3.1 Dimensional analysis

There is a fairly simple reason to expect a constant relative hazard rate. It's been about 60 years since Dartmouth, so let's round x_m to 60. What is the probability that we will taxi within the next 30 years? That depends on the value of α, of course. Conditional on not taxiing within the next 20 years, what is the probability that we will taxi within 40 years after that? That also depends on α, but it turns out that no matter what α is, these two probabilities are exactly the same. Similarly, if we did these calculations 20 years ago, with $x_m = 40$, the probability of taxiing within 20 years from then would have been the same as the probability, conditional on not taxiing until now, of taxiing within 30 years from now. More generally, $P(X < cx_{m_1} | X > x_{m_1}) = P(X < cx_{m_2} | X > x_{m_2})$, where X is the time since Dartmouth until taxiing occurs, x_{m_1} and x_{m_2} are some arbitrary times since Dartmouth, and c is an arbitrary contant greater than 1. In the above example, $c = \frac{3}{2}$ and the values of x_m considered are 40, 60, and 80.

This should be intuitive, because our current state of knowledge can be summed up as "There's been exciting AI research going on for the past 60 years, much of which seems like it could contribute to AGI, but there's been no clear progress on AGI in particular." 20 years ago, our state of knowledge could be summed up as "There's been exciting AI research going on for the past 40 years, much of which seems like it could contribute to AGI, but there's been no clear progress on AGI in particular." These are identical, except that the time mentioned has been increased by a factor

of $\frac{3}{2}$ in the current one. Thus it seems reasonable that we should assign the same probability to the statement that we will start taxiing to AGI within the next 30 years as we would have assigned 20 years ago to the statement that we will start taxiing to AGI within 20 years. The probability distribution over time to taxi conditional on not taxiing within n years should look the same as the current distribution, but stretched out in time. That is, $P\left(cx_m < X < dx_m | X > x_m\right)$ should not depend on x_m. Pareto distributions are the only distributions that obey this condition.

3.2 Stable distributions

The dimensional analysis argument stops looking like it makes any sense if you scale it back far enough. Surely, a week into the Dartmouth Conference, no one thought they'd be on the verge of Strong AI in another week, even though 60 years later, it looks entirely plausible that we will be within the next 60 years. This is a general problem with Pareto distributions with small x_m, but the Pareto distribution is considered a useful model anyway because many more reasonable probability distributions have tails that approximate a Pareto distribution.

When many different independent factors are involved, we often find the data to follow a stable distribution; that is, a probability distribution in which the sum of random variables from two copies of the distribution can be described by the same distribution if shifted and scaled properly. A generalization of the central limit theorem says that the sum of a large number of independent and identically distributed variables from a distribution with Pareto-like tails must be described by a stable distribution. The most well-known stable distribution is the normal distribution. However, normal distributions have tails that fall off super-exponentially. That is, even if you are already well past most of the probability mass for time to taxi in a normal distribution, then each day that we don't taxi, you will get more and more confident that we will taxi the next day. But intuitively, the reverse should be true; you would think eventually we would notice that AGI just doesn't happen as quickly as we initially expected it to. Thus a normal distribution doesn't look adequate to model our uncertainty about what time we would expect to taxi.

The normal distribution is the best-known stable distribution because it is the only one that has a finite variance. But there are many other stable distributions that do not have a well-defined variance, and all of these other stable distributions have Pareto-like tails. Early AI researchers expected that they would get Strong AI fairly quickly, and this did not happen, suggesting that we are already in the late tail of the distribution that they expected then. Thus the probability distribution conditional on not taxiing by now should look like a Pareto distribution. If you disagree that we are already most of the way through the a priori probability distribution on time to

taxi, then a Pareto distribution might not seem like such a reasonable model.

3.3 Exponential distribution with unknown hazard rate

Consider the model in which there is some underlying exponential distribution that should predict when we will taxi, but we don't know which exponential distribution it is. This could be the case if there were one crucial insight that would lead to AGI, and prior work doesn't change the probability of someone achieving that insight soon very much, so the hazard rate would be constant. But of course we wouldn't know in advance what the hazard rate is. (Of course, I've just been trying to convince you that we should expect Pareto tails, but note that all the other arguments refer to our state of uncertainty, not to an underlying process. In fact, if there is such an underlying unknown distribution, you might expect it to be slightly superexponential since the last insight required to taxi probably will build on previous work, and thus become more likely over time.)

According to a continuous generalization of Laplace's rule of succession, if there's an underlying exponential distribution with unknown hazard rate λ, then after updating on the fact that we aren't taxiing yet, our posterior distribution over when we will taxi should be a Pareto distribution with $\alpha = 1$. This corresponds to using an improper prior in which the probability density function for λ is constant. Another way to describe this prior is to break up time into small discrete chunks (e.g. years), and use the standard Laplace's rule of succession to update your uncertainty over how likely we are to taxi in a random year after updating on the fact that we haven't taxied in any of the previous 60 years. It turns out that this assumption describes approximately the same probability distribution, and if you change the size of the discrete chunks instead (e.g. use months instead of years), it does not make much difference, provided both sizes are sufficiently small relative to the timescale being considered.

3.4 Intuitions

Despite the fact that expert intuition is fairly bad at predicting things like AI, intuitions aren't completely useless. So it's worth pointing out that many people find that a Pareto distribution for some α (usually somewhat large) decently approximates their intuitions about when we are likely to taxi to AGI.

4 Potential objections

4.1 Oversimplification

"This model assumes that one day, we'll suddenly realize that we're on our way to Strong AI, and the previous day, our epistemic state will not be much different than it is right now with regards to time to AGI. That doesn't seem very realistic. Whether or not we are taxiing to AGI does not have a well-defined binary answer."

That is true. We made the taxiing assumption not because it is entirely accurate, but because it makes the situation easier to model, and is not too horribly far off. There could certainly be edge cases, but it seems that roughly speaking it should be possible to divide most cases into taxiing and not taxiing, and right now we are not taxiing.

4.2 Tail

"This model assumes that we have already passed most of the probability mass for taxiing to AGI. But I wouldn't have predicted ahead of time that we'd get AGI by now, so I don't think it is accurate to assume that we are already in the tail of the distribution."

In the early days of AI, most AI researchers expected that we would get AGI fairly quickly, so it seems reasonable to assume that we are in the tail of the a priori distribution now. But if you think you have a good a priori reason to believe we probably would not have gotten AGI before now, then a model that assumes we are in the tail should not look compelling.

4.3 Time versus work

"The time elapsed since we started working seriously on AI is less relevant than the total amount of effort that has been put into working on AI. But this model only pays attention to the amount of time since we started working on AI."

Yes, the total amount of effort and resources invested in AI (which could be measured in person-hours, for instance) is more directly relevant to getting AGI than elapsed time is, but it is also more difficult to model. Explicitly translating a model of work to taxi into a model of time to taxi would require estimating how much work has already been invested in AI, and generating a probabilistic model for how much work we will invest in AI in the future, which is hard. However, if we assume the rate at which work is invested in AI does not vary too wildly, then starting from a Pareto distribution for work to taxi, we should get a Pareto-like distribution for time to taxi.

On a related note, if we assume that we have a Pareto-tailed (with relative hazard rate α) distribution D over the intrinsic difficulty of AGI, and a distribution E involving effort invested and random noise and so on, such that $D \cdot E$ represents our overall distribution over time to taxi, then $D \cdot E$ is also Pareto-tailed, with the same relative hazard rate α, provided that E^α has a finite mean (informally: that just puts a limit on how wide E can be, but it's a fairly weak condition).

5 Policy implications

Our predictions for when AGI is likely to arrive affect what strategies we should be implementing now to ensure that it is Friendly. If we think that AGI will probably arrive quickly, then we should focus on strategies that we expect to pay off quickly. If we think there is little risk of AGI arriving quickly, then we should focus on strategies that take longer but have a higher chance of working.

Let's assume a Pareto distribution with $\alpha = 1$. The median time to taxi is 60 years, a fairly generous amount of time. This might make it seem like interventions on about that scale are very useful; if a research program takes 40 years before it delivers useful results, then it still has a chance to affect the majority of possible AGI outcomes. However, the distribution assigns significant probability that time to taxi will be quite a lot more than 60 years. There is only a 20% chance that taxiing will occur between 40 and 90 years from now, and after 90 years, it is likely that our efforts won't have much of an effect. 20% is still significant, so we should hardly ignore interventions on the 40 to 90 year timescale, but it no longer looks like a good idea to concentrate the bulk of our efforts there. In comparison, there is a $\frac{1}{3}$ chance that taxiing will occur in the next 30 years, so it could be quite valuable to use interventions that will pay off within 30 years, even though the median time to taxi is much longer than that. Different values of α give similar results; in general, a Pareto distribution suggests that we should put a much greater emphasis on short-term strategies than a less skewed distribution (e.g. a normal distribution) with the same median would.