Don’t rule out AGI in the first 100 years of trying

An explicit aim of the field of AI R&D since its inception, has been to create AGI.[1] While some AI researchers view AGI as a pipe dream, others believe it’s likely to be achieved within a few decades[2] and forty organisations are actively researching it.[3] For example, developing AGI is a central objective of DeepMind, a prominent and well-funded AI research lab.[4]

So AGI is in the reference class of “ambitious but feasible technology that a serious STEM field is explicitly trying to build”. By ‘feasible’ I mean that i) a large proportion of experts claim it will be eventually possible, ii) a reasonable proportion of experts claim it may be possible within a number of decades.[5]

Ideally, this section would gather data on the success rate of STEM fields with ambitious goals and use this to determine the *first-trial probability*. Instead, I simply motivate the claim that such fields have often been very successful and use this as an intuition pump to place a lower bound on the *first-trial probability* for AGI.

The intuition pump

Human scientific and technological R&D has recently had a very good track record. Without any pretence at completeness, here are some examples:

- In 1895 we didn’t know energy was divided into quanta or about the existence of electrons, the nucleus, the weak force or the strong force. But by 1970 the standard model had been finalised, which gives a unified and empirically adequate description of all the fundamental forces except gravity

[1] The proposal for the Dartmouth conference, widely seen as the birth of AI R&D, states that ‘*The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves.*’

Stuart Russell, professor of Computer Science and author of a best selling textbook in AI, says that “The [AI] field’s goal had always been to create human-level or superhuman AI” (*Human Compatible*, pp. 1-2).

[2] Grace, Katja (2017) ‘*When Will AI Exceed Human Performance? Evidence from AI Experts*’

[3] Baum, Seth (2017), "*A Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy*"

[4] DeepMind, having been acquired by Google in 2014 for $500 million, spends hundreds of millions of dollars each year on salaries alone. Forbes writes that ‘*In 2018, DeepMind spent $483 million (£398 million) on around 700 employees, up from $243 million (£200 million) in 2017. Other significant costs included technical infrastructure and operating costs. In addition, DeepMind spent $17 million (£14 million) on academic donations and sponsorships.*’

[5] This is evidenced by expert surveys, see Katja (2017) and Gruetzemacher (2020), and the existence of many firms specifically trying to build AGI (see introduction).

- In 1930 we didn't know about the strong nuclear force, but by 1945 we'd created the atom bomb and by 1951 the first nuclear power plant (source).
- In 1789 leading chemists were sceptical of Antoine Lavoisier's *Elementary Treatise of Chemistry,* which defined an element and listed some examples. But by 1871, Mendeleev published his periodic table, which predicted the discovery of further elements and has not been changed significantly to this day (source).
- In 1927 we hadn't identified the components parts of DNA and didn't know that it carried genetic information; but by 1957, the relationship between DNA, RNA and proteins (the "central dogma") was understood (source). In 2020, we can rapidly sequence DNA, synthesize DNA, and cut-and-paste DNA in living cells (with CRISPR) (source).
- In 1900, treatments for infections were mostly based on medicinal folklore, but by 1962 most of the antibiotic classes we use today had been discovered and introduced to the market (source).
- Before 1900 we hadn't managed sustained and controlled heavier-than-air flight, but by 1969 we'd landed a man on the moon and flown a jumbo jet that could carry 366 passengers.
- The first general-purpose digital computer was built in 1945 and by 2015 more than 3 billion people were using the internet
- In 1881, the first (highly inefficient) solar panel was created. By 2017, 2.6% of the world's electrical power was from solar panels.
- The process of electrification in the US and Britain happened between 1880 and 1960.
- The first commercial steam engine was introduced in 1712. By 1860, steam engines generated 80% of total power in the US.

These examples are clearly very far from complete or comprehensive. However, they show that in a few decades human R&D can make very significant advances in diverse and central areas of human understanding and technology: physics, chemistry, biology, medicine, transportation, communication, information, and energy.

Given these achievements, it would be unreasonable to assign very low probability to the serious STEM field of AI R&D achieving one of its central aims even after 100 years of sustained effort.[6] A lot of progress can happen in 100 years and no systematic attempt had been made to develop AGI before 1956.

If we had reason to think AGI was impossible to create, things would be different. But, to the contrary, the human brain is an existence proof of a generally intelligent system[7] and most experts believe that AGI is feasible.

[6] Remember, the *trial definition* I'm currently considering is '1 calendar year of sustained R&D effort'. This means that if the research community gave up, no further trials would occur. Given this trial definition, it is reasonable to interpret the occurrence of a trial as a positive sign that some progress has been made, or at least that researchers believe progress will be made in the future.

[7] This evidence does of course leave open the possibility that non-biological generally intelligence is impossible.

A lower bound for the *first-trial probability*

The following table shows the probability of success in the first *N* years of trying, for different values of *first-trial probability* and *N*.

first-trial probability	**pr(AGI in the first 100 years of effort)**	**pr(AGI in the first 50 years of effort)**	**pr(AGI in the first 30 years of effort)**
1/10	92%	85%	77%
1/50	67%	51%	38%
1/100	50%	33%	23%
1/300	25%	14%	9%
1/1000	9%	5%	3%
1/3000	3.2%	1.6%	1.0%
1/10,000	1.0%	0.5%	0.3%

To avoid overconfidence, I claim you should assign >3% chance to AGI being created within 100 years of sustained effort and >1%. So I recommend bounding *first-trial probability* **above 1/3000**, and probably above 1/1000.

Estimates of the *first-trial probability*

These central estimates are extremely rough. But, as discussed in the main text, but to have a rough estimate of the *first-trial probability* than none at all. I'll give an optimistic estimate and a conservative one.

The time taken to achieve the R&D milestones discussed above varies significantly - this underscores the deep uncertainty about how long it might take to develop AGI - and the mean time was 78 years (see calculation). A naive conclusion would be that AGI might take a comparable amount of time, suggesting a *first-trial probability* in the range 1/100 to 1/50. If you thought AGI was more relevantly similar to the things on the list that took less long to develop, you might optimistically set *first-trial probability* = **1/50**.

A more conservative estimate would adjust downwards based on multiple considerations:

- Building AGI may be an especially ambitious task, even compared to those listed above.

- AGI may come in the form of thousands of distinct AI systems, each specialised for a different narrow task.[8] From this perspective, it seems plausible that the process of "expanding our AI tool kit to cover all human tasks" won't finish within 100 years. In addition, this view could mean that AGI requires multiple distinct breakthroughs.
- Perhaps explicit goals ahead of time are less likely to succeed than flexible exploration into a new area. This might suggest that while AI R&D might produce something transformative, it needn't be in the form of AGI. That said, our definition of AGI was purposefully broad: it includes any system, or collection of systems, with the requisite cognitive abilities.
- There is a selection bias in the examples above. They were selected because they turned out to be successes *ex-post*. But there are many failed R&D attempts that I did not include in the list. The attempt to build AGI could (from the perspective of setting the *first-trial probability* in 1956) be more like the failures than those included on the list.
    - This is, to a small extent, addressed by considerations already raised in this section, which give reason to think AGI will be like the items on the list.
        - Firstly, the occurrence of a trial requires a 'sustained effort to develop AGI by a serious research field for a calendar year'. If AI was a non-starter we wouldn't expect a sustained effort to continue for long. So if a large number of trials *do* occur, this suggests that developing AGI, or something in its vicinity, is a promising avenue of research, like the examples on the list above.
        - Secondly, the expert view that AGI is feasible suggests AGI is likely to be similar to the items on the above list.
    - Overall, this point makes me reduce the *first-trial probability* for the conservative estimate by a factor of 3.

Based on these considerations, my conservative point estimate is **1/300**. This implies 25% of success in the first 100 years, which feels roughly right to me as a conservative but plausible view.

[8] This is the 'AI services' vision described in Drexler (2019).