Part 2. Part 3. Appendices.

1 Introduction

1.1 Executive summary

The goal of this report is to reason about the likely timing of the development of artificial general intelligence (AGI). By AGI, we mean computer program(s) that can perform virtually any cognitive task as well as any human,[1] for no more money than it would cost for a human to do it. The field of AI is largely held to have begun in Dartmouth in 1956, and since its inception one of its central aims has been to develop AGI.[2]

We forecast when AGI might be developed using a simple Bayesian framework, and choose the inputs to this framework using commonsense intuitions and reference classes from historical technological developments. The probabilities in the report represent reasonable degrees of belief, not objective chances.

One rough-and-ready way to frame our question is this:

> *Suppose you had gone into isolation in 1956 and only received annual updates about the inputs to AI R&D (e.g. # of researcher-years, amount of compute[3] used in AI R&D) and the binary fact that we have not yet built AGI? What would be a reasonable pr(AGI by year X) for you to have in 2021?*

There are many ways one could go about trying to determine pr(AGI by year X). Some are very judgment-driven and involve taking stances on difficult questions like "since AI research began in 1956, what percentage of the way are we to developing AGI?" or “what steps are needed to build AGI?”. As our framing suggests, this report looks at what it would be reasonable to believe *before* taking evidence bearing on these questions into account. In the terminology of Daniel Kahneman’s *Thinking Fast and Slow*, it takes an “outside view” approach to forecasting, taking

---

[1] Notice that this definition applies equally whether it is a single artificial agent that can perform all these tasks, or a collection of narrower systems working together. The ‘single agent’ perspective is the focus of Bostrom’s *Superintelligence*, while Drexler (2019) argues that general AI intellectual capabilities will first come in the form of many diverse AI systems. There are various ways to make this definition more precise - see Muehlhauser (2013).

[2] The proposal for the Dartmouth conference 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). Well-funded research labs are actively researching AGI, including DeepMind and Open AI. Baum (2017) identifies many other active AGI R&D projects.

[3] ‘Compute’ means computation. In this report we operationalise this as the number of floating point operations (FLOP).

into account relevant reference classes but not specific plans for how we might proceed.[4] The report outputs a pr(AGI by year X) that can potentially be updated by additional evidence.[5]

Our framework only conditions on the *inputs* to AI R&D - in particular the *time* spent trying to develop AGI, the *number of AI researchers*, and the *amount of compute* used - and the fact that we haven't built AGI as of the end of 2020 despite a sustained effort.[6] We place subjective probability distributions ("beta-geometric distributions") over the amount of each input required to develop AGI, and choose the parameters of these distributions by appealing to analogous reference classes and common sense. Our most sophisticated analysis places a hyperprior over different probability distributions constructed in this way, and updates the weight on each distribution based on the observed failure to develop AGI to date.

For concreteness and historical reasons,[7] we focus throughout on what degree of belief we should have that AGI is developed by the end of 2036: pr(AGI by 2036).[8] Our central estimate is about 8%, but other parameter choices we find plausible yield results anywhere from 1% to 18%. Choosing relevant reference classes and relating them to AGI requires ***highly subjective judgments***, hence the large confidence interval. Different people using this framework would arrive at different results.

To explain our methodology in some more detail, one can think of inputs to AI R&D - time, researcher-years, and compute - as "trials" that might have yielded AGI, and the fact that AGI has not been developed as a series of "failures".[9] Our starting point is Laplace's rule of succession, sometimes used to estimate the probability that a Bernoulli "trial" of some kind will "succeed" if *n* trials have taken place so far and *f* have been "failures". Laplace's rule places an *uninformative* prior over the unknown probability that each trial will succeed, to express a maximal amount of uncertainty about the subject matter. This prior is updated after observing the result of each trial. We can use Laplace's rule to calculate the probability that AGI will be developed in the next "trial", and so calculate pr(AGI by 2036).[10]

---

[4] See this appendix for more detailed discussion of the evidence that this report does and doesn't take into account, and this appendix for discussion of how we might update the prior in response to additional evidence.

[5] Of course, technically speaking the distribution is also a posterior as it updates on the failure to develop AGI by the end of 2020.

[6] The analysis can easily be extended to cover beliefs like, 'We can see that AGI will not be developed in the next 6 years, but after that we don't know'.

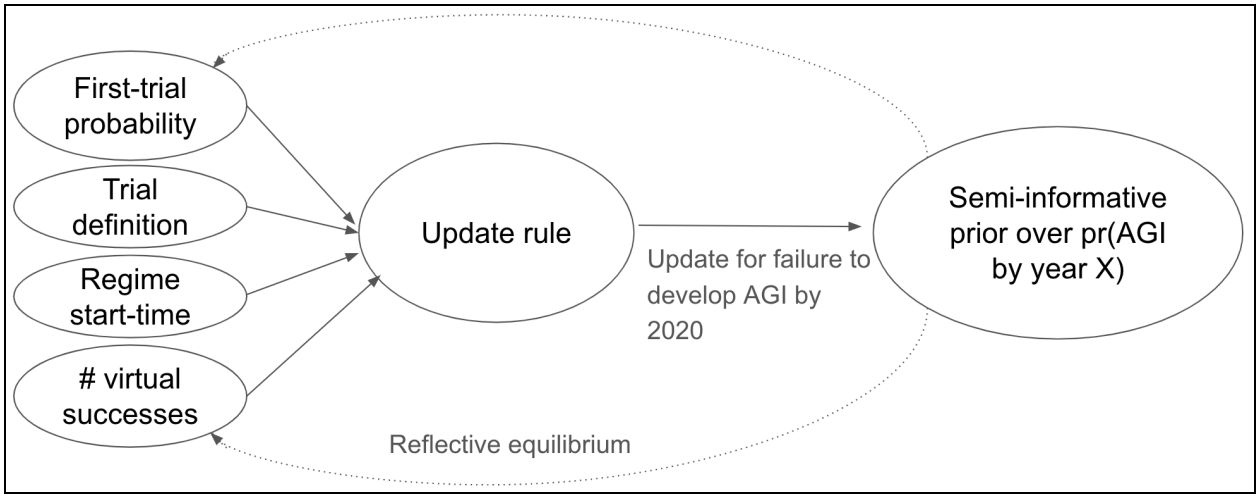
[7] In a 2016 blog post, Open Philanthropy CEO Holden Karnofsky stated that he would personally estimate a >10% chance of transformative artificial intelligence being developed within 20 years. Using 2036 for this report allows its bottom line to be more easily compared with Holden's statement. My colleague Ajeya Cotra briefly discusses the relation of AGI to 'transformative artificial intelligence' in this section of her draft report.

[8] The analysis is easily extended to give the probability of AGI in any period.

[9] Though this assumption is not literally true, we find it gives rise to a fruitful framework that can approximate other reasonable distributions that we might have used, and that our results are not driven by the framework but by our inputs to it. We defend this claim at length in appendix 12.

[10] Strictly speaking, we should write this as pr(AGI by 2036 | no AGI by 2020), but we shorten this to pr(AGI by 2036) throughout.

We identify severe problems with this calculation. In response, we introduce a family of *update rules,* of which the application of Laplace's rule is a special case.[11] Each update rule can be updated on the failure to develop AGI by 2020 to give pr(AGI by year X) in later years. When a preferred update rule is picked out using common sense and relevant reference classes, we call the resultant pr(AGI by year X) a 'semi-informative prior'. We sometimes use judgements about what is a reasonable pr(AGI by year X) to constrain the inputs, trying to achieve reflective equilibrium between the inputs and pr(AGI by year X).



A specific update rule from the family is specified by four inputs: a *first-trial probability* (ftp), a number of *virtual successes*, a *regime start-time,* and a *trial definition*.

- The *first-trial probability* gives your odds of success on the first trial. Roughly speaking, it corresponds to how easy you thought AGI would be to develop *before* updating on the observed failure to date.
    - The main problem with Laplace's rule is that it uses a first-trial probability of 50%, which is implausibly high and results in inflated estimates of pr(AGI by 2036).
- The *number of virtual successes* influences how quickly one updates away from the first-trial probability as more evidence comes in (etymology explained in the report).
- The *regime start-time* determines when we start counting successes and failures, and we think of it in terms of when serious AI R&D efforts first began.
- The *trial definition* specifies the increase of an R&D input corresponding to a "trial" - e.g. 'a calendar year of time' or 'a doubling of the compute used to develop AI systems'.

We focus primarily on a *regime start-time* of 1956, but also do sensitivity analysis comparing other plausible options. We argue that a *number of virtual successes* outside of a small range has intuitively odd consequences, and that answers within this range don't change our results

[11] The family of update rules generalises Laplace's uniform distribution to a beta distribution. Other popular uninformative priors like the Jeffreys prior or the Haldane prior are also beta distributions, so the framework can express these variants on Laplace's rule.

much. Within this range, our favoured *number of virtual successes* has the following implication: if the *first-trial probability* is 1/X, then pr(AGI in the first X years) ≃ 50%.

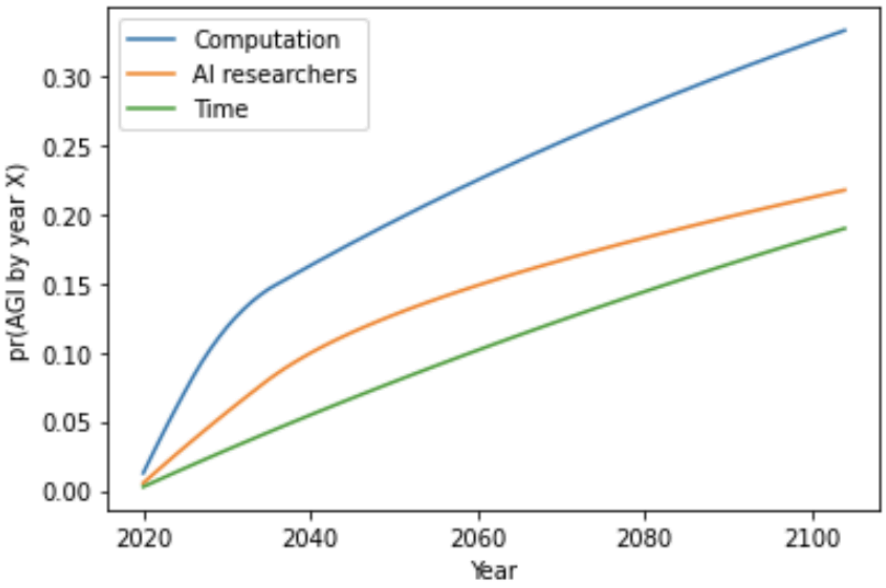
The *first-trial probability* is much harder to constrain, and plausible variations drive more significant differences in the bottom line than any other input. Taking a trial to be a 'a calendar year of time', we try to constrain the *first-trial probability* by considering multiple reference classes for AGI - for example "ambitious but feasible technology that a serious STEM field is explicitly trying to build" and "technological development that has a transformative effect on the nature of work and society" - and thinking about what *first-trial probability* we'd choose for those classes in general. On this basis, we favor a *first-trial probability* in the range [1/1000, 1/100], and feel that it would be difficult to justify a *first-trial probability* below 1/3000 or above 1/50. A *first-trial probability* of 1/300 combined with a 1956 regime start-time and 1 virtual success yields pr(AGI by 2036) = 4%.

We consider variations on the above analysis with trials defined in terms of researcher-years and compute used to develop AI, rather than time.[12] We find that these variations can increase the estimate of pr(AGI by 2036) by a factor of 2 - 4. We also find that the combination of a high *first-trial probability* and a late *regime start-time* can lead to much higher estimates of pr(AGI by 2036).

Trial definition	*Low ftp*	*Central ftp*	*High ftp*	*High ftp and late start-time: 2000*
Calendar year	1%	4%	9%	12%
Researcher-year	2%	8%	15%	25%
Compute[13]	2%	15%	22%	28%

Here are our central estimates for pr(AGI by year X) out to 2100, which rely on crude empirical forecasts past 2036.[14]

[12] We do account for diminishing returns to R&D effort. In particular, we define trials as percentage increases in i) the total number of AI researcher-years, and ii) the compute used to develop the largest AI systems. We discuss our choice of trial definition at greater length in section 6 and in this appendix.
[13] The 2nd and 3rd columns of the compute row assign 50% weight to a start-time of 1956 and 50% to a late start-time - the regime started when the amount of computation needed to run a human brain first became affordable. This is in contrast to the other rows, which assign 100% to start-time = 1956 in the first three columns, and 100% to start-time = 2000 in the fourth column.
[14] The graph assumes that the number of AI researchers will grow at 11% until 2036 (based on recent data), and then grow at 4% (the US R&D average growth over the last 80 years). It also assumes spending on computation will rise to $1 billion by 2036, and then stay constant, while the cost of computation will fall by 100X by 2036, and then halve every 2.5 years. These compute assumptions are based on the tentative forecasts of my colleague Ajeya Cotra in her draft report.



;

To form an all-things-considered judgment, we place a hyperprior over different update rules (each update rule is determined by the 4 inputs). The hyper-prior assigns an initial weight to each update rule and then updates these weights based on the fact that AGI has not yet been developed.[15]

The inputs leading to low-end, central, and high-end estimates are summarized in this table (outputs in bold, inputs in standard font).

	Low-end	Central	High-end
first-trial probability (trial = 1 calendar year)	1/1000	1/300	1/100
Regime start-time	1956	85% on 1956 15% on 2000	20% on 1956 80% on 2000
Initial weight on *time* update rule	50%	30%	10%
Initial weight on *researcher-year*	30%	30%	40%

[15] This hyper prior update is a standard application of Bayesian updating. Suppose you have two rules, r and s. Suppose the likelihoods of the evidence (AGI not being developed by 2020) for both rules is as follows: pr(e|r) = 50%, pr(e|s) = 25%. If your initial weights in r and s are in the ratio 1:1, the ratio after updating will be 2:1, as s is twice as surprised by the evidence. So if you initially place 50% weight on each, then after updating you'll place 67% on r and 33% on s.

update rule			
Initial weight on *compute* update rule	0%	30%	40%
Initial weight on AGI being impossible	20%	10%	10%
pr(AGI by 2036)	**1%**	**8%**	**18%**
pr(AGI by 2100)	**5%**	**20%**	**35%**

The 4 rows of weights are set using intuition, again highlighting the highly subjective nature of many inputs to the framework. We encourage readers to use this tool to see the results of their preferred inputs.

Much of the report investigates and confirms the robustness of these conclusions to a variety of plausible variations on the analysis and anticipated objections. For example, we consider models where developing AGI is seen as a conjunction of independent processes or a sequence of accomplishments; some probability is reserved for AGI being impossible; different empirical assumptions are used to fix the first-trial probability for various trial definitions. We also consider whether using this approach would produce absurd consequences in other contexts (e.g. what does analogous reasoning imply about other technologies?), whether it matters that the framework is discrete (dividing up continuous R&D inputs into arbitrarily sized chunks), and whether it's a problem that the framework models AI R&D as a series of Bernoulli trials. On this last point, we argue in appendix 12 that using a different framework would not significantly change the results because our bottom line is driven by our choices of inputs to the framework rather than our choice of distribution.

One final upshot of interest from the report is that the failure to develop AGI to date is not strong evidence for low pr(AGI by 2036). In this framework, pr(AGI by 2036) lower than ~5% would primarily be a function of one's *first-trial probability*. In other words, a pr(AGI by 2036) lower than this would have to be driven by an expectation — before AI research began at all — that AGI would probably take hundreds or thousands of years to develop.[16]

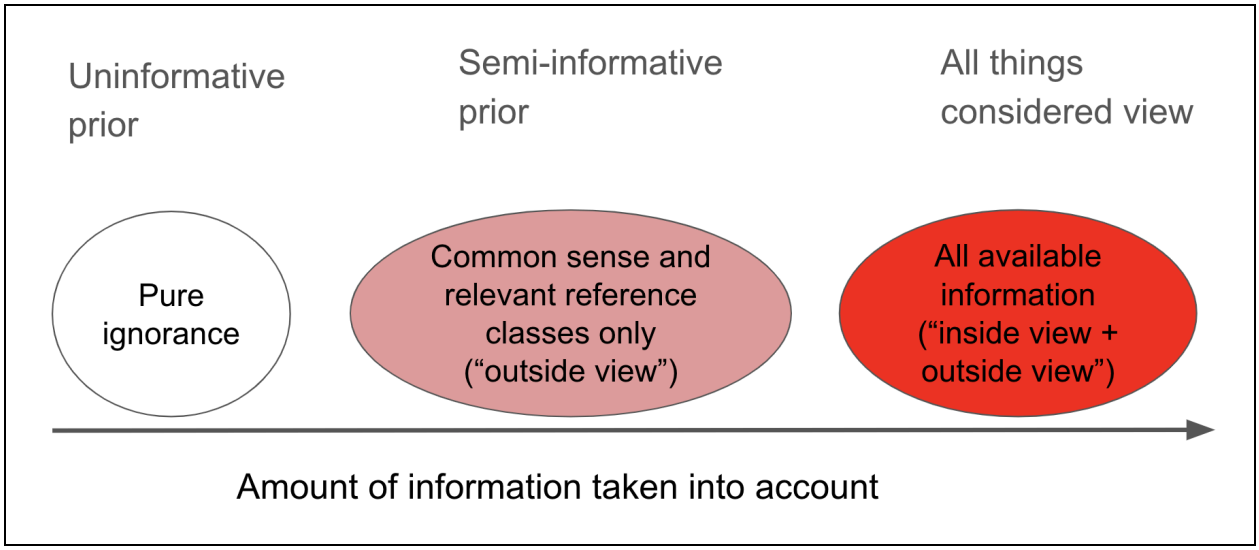
Acknowledgements: TODO

[16] A low pr(AGI by 2036) could potentially also be driven by evidence that this report does not take into account. For example, if you believe we can measure how far we are away from AGI, and how fast we are progressing towards it, then this might allow you to argue that pr(AGI by 2036) is very low.

1.2 Structure of report

Section 2 applies Laplace's rule of succession to calculate pr(AGI by year X). We call the result an 'uninformative prior over AGI timelines', because of the rule's use of an uninformative prior. This approach yields pr(AGI by 2036) of 20%.

Section 3 identifies a family of update rules of which the previous application of Laplace's rule is a special case, highlighting some arbitrary assumptions made in section 2. When a preferred update rule from the family is picked out using common sense and relevant reference classes, we call the resultant pr(AGI by year X) a 'semi-informative prior over AGI timelines'.



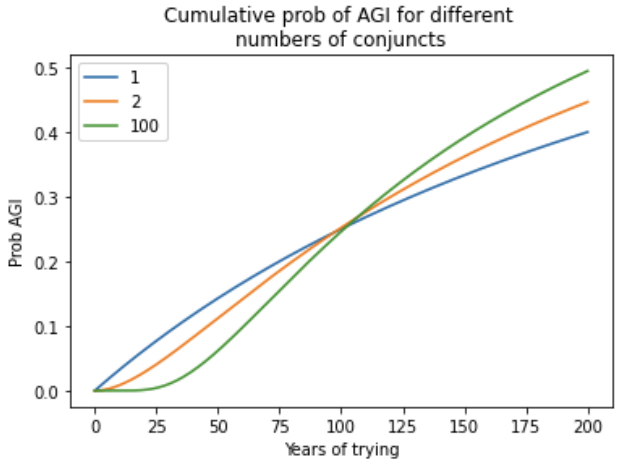
Section 3 also identifies severe problems with the application of Laplace's rule to AGI timelines, but suggests that these do not arise in context of the broader family of update rules. Lastly, it conducts a sensitivity analysis which highlights that one input is particularly important to pr(AGI by 2036) - the *first-trial probability*.

Section 4 describes what I think is the correct methodology for constraining the *first-trial probability* in principle, and discusses a number of considerations that might help the reader constrain their own *first-trial probability* in practice. I then explain the range of values for this input that I currently favour. Much more empirical work could be done to inform this section; the considerations I discuss are merely suggestive. This is somewhat unfortunate as the *first-trial probability* is the single most important determinant of your bottom line pr(AGI by 2036) in this framework.

Section 5 analyses how much the *number of virtual successes* and *regime start-time* affect the bottom line, once you've decided your *first-trial probability*. Its key conclusion is that they don't matter very much.

Section 6 considers definitions of a 'trial' researcher-years and compute. (Up until this point a 'trial' was defined as a year of calendar time.) More specifically, it defines trials as percentage increases in i) the total number of AI researcher-years, and ii) the compute used to develop the largest AI systems.[17] I find that each successive trial definition increases the bottom line, relative to those before it. This is because the relevant quantities are all expected to change rapidly over the next decade, matching recent trends[18], and so an outsized number of 'trials' will occur.

Section 7 extends the model in three ways, and evaluates the consequences for the bottom line. First, it explicitly models AGI as conjunctive. In this simple extension, multiple goals must be achieved to develop AGI and each goal has its own semi-informative prior. I also consider models where these goals must be completed sequentially. The main consequence is to dampen the probability of developing AGI in the initial decades of development. These models output similar values for pr(AGI by 2036), as they make no assumption about how many conjuncts are completed as of 2020.

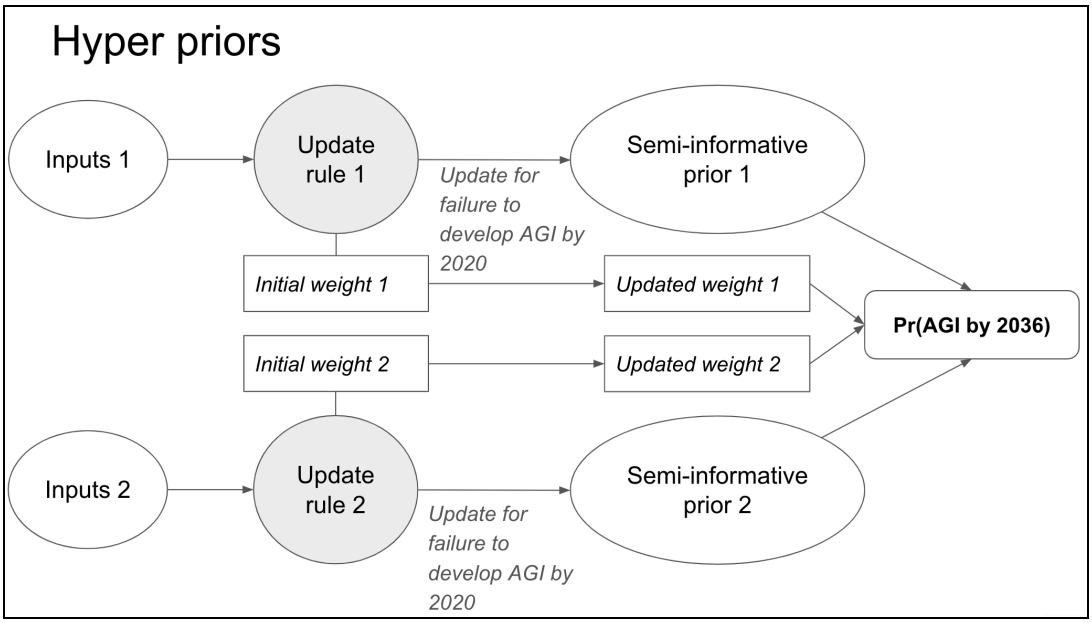


Second, section 7 places a *hyperprior* over multiple semi-informative priors. The hyperprior assigns initial weights to the semi-informative priors and updates these weights based on how surprised each prior is by the failure to develop AGI to date. The semi-informative priors may differ in their *first-trial probability*, their *trial definition*, or in other ways. Thirdly, it explicitly models the possibility that AGI will never be developed, which slightly decreases pr(AGI by 2036).

---

[17] Why use percentage increases in inputs, rather than absolute increases? Essentially because only the former reflects our deep uncertainty about the *order of magnitude* of effort that will be required to develop AGI. It turns out that in our framework the latter (absolute increases) implies that the unconditional probability of developing AGI is concentrated almost entirely in 1 - 2 orders of magnitude of effort. We don't find this plausible. Further, when inputs are increasing exponentially (as they are currently) the latter choice concentrates the probability of developing AGI in a ~20 year period. Again, we don't find that this adequately reflects our uncertainty about when AGI will be developed. We explain our choice of trial definition at greater length in section 6 and in this appendix.

[18] The compute used to develop the largest AI systems has increased significantly over the last 60 years due to Moore's Law, and has recently been increasing even faster due to increased $ spending. The number of researchers has recently also been growing quickly (see these sources).



Section 8 concludes, summing up the main factors that influence the bottom line. My own weighted average over semi-informative priors implies that pr(AGI by 2036) is about **8%**. Readers are strongly encouraged to enter their own inputs using this tool.

The appendices cover a number of further topics, including:

- In what circumstances does it make sense to use the semi-informative priors framework (here)?
- Is it a problem that the framework unrealistically assumes that AI R&D is a series of Bernoulli trials (here)?
- Is it a problem that the framework treats inputs to AI R&D as discrete, when in fact they are continuous (here)?
- Does this framework assign too much probability to crazy events happening (here)?
- Is the framework sufficiently sensitive to changing the details of the AI milestone being forecast? I.e. would we make similar predictions for a less/more ambitious goal (here)?
- How might other evidence make you update from your semi-informative prior (here)?

Appendix 12 is particularly important. It justifies the adequacy of the semi-informative priors framework, given this report's aims, in much greater depth. It argues that, although the framework models the AGI development process as a series of independent trials with an unknown probability success, the framework's legitimacy and usefulness does not depend upon this assumption being literally true. To reach this conclusion, I consider the *unconditional probability distributions over total inputs* (total time, total researcher-years, total compute) that the semi-informative priors framework gives rise to. This turns out to correspond to the family of beta-geometric distributions. Each semi-informative prior corresponds to one such beta-geometric distribution, and we can consider these distributions as fundamental (rather than

derivative on the assumption that AI R&D is a series of trials). I argue that this class of unconditional probability distributions is sufficiently expressive for the purposes of this report.

Three academics reviewed the report. I link to their reviews in appendix X.

Note: throughout the report, potential objections and technical subtleties are often discussed in footnotes to keep the main text more readable.

2 Uninformative priors over AGI timelines

2.1 The sunrise problem

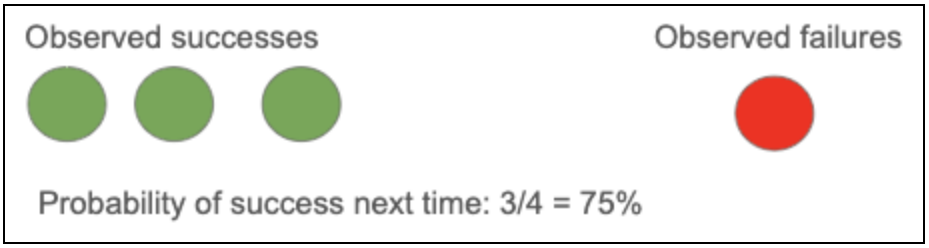
The polymath Laplace introduced the sunrise problem:

Suppose you knew nothing about the universe except whether, on each day, the sun has risen. Suppose there have been N days so far, and the sun has risen on all of them. What is the probability that the sun will rise tomorrow?

Just as we wish to bracket off information about precisely how AGI might be developed, the sunrise problem brackets off information about why the sun rises. And just as we wish to take into account the fact that AGI has not yet been developed as of the start of 2020, the sunrise problem takes into account the fact that the sun has risen on every day so far.

2.1.1 Naive solution to the sunrise problem

One might think the probability of an event is simply the fraction of observations you've made in which it occurs: (number of observed successes) / (number of observations).[19]



In the sunrise problem, we've observed *N* successes and no failures, so this naive approach would estimate the probability that the sun rises tomorrow as 100%. This answer is clearly unsatisfactory when *N* is small. For example, observing the sun rise just three times does not warrant certainty that it will rise the next day.

[19] This approach can be expressed in a Bayesian framework by placing a Haldane prior over the probability that each observation is a success. The frequentist Hans Reichenbach proposed this method.

2.1.2 Laplace's solution to the sunrise problem: the rule of succession

Laplace's proposed solution was his rule of succession. He assumes that each day there is a 'trial' with a constant but unknown probability *p* that the sun rises. To represent our ignorance about the universe, Laplace recommends that our initial belief about *p* is a uniform distribution in the range [0, 1]. According to this uninformative prior, *p* is equally likely to be between 0 and 0.01, 0.5 and 0.51, and 0.9 and 0.91; the expected value of *p* $E(p) = 0.5$.

When you update this prior on *N* trials where the sun rises and none where it does not,[20] your *posterior* expected value of *p* is:

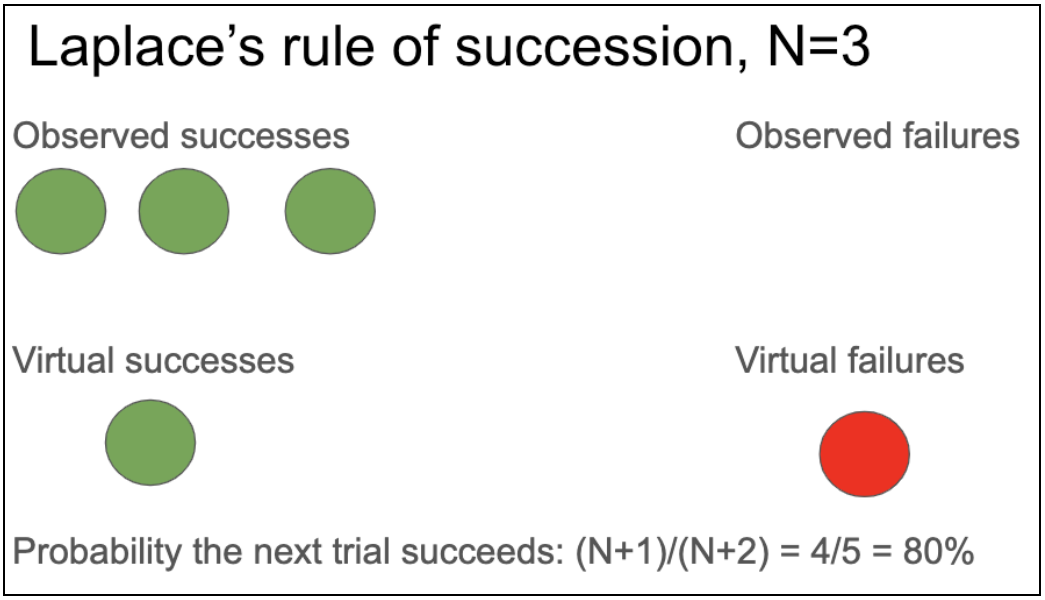
$$E(p) = (N + 1) / (N + 2)$$

In other words, after seeing the sun rise without fail *N* times in a row, our probability that it will rise on the next day is $(N + 1) / (N + 2)$.

One way to understand this formula is to suppose that, before we saw the sun rise on the first day, we made two additional *virtual observations*.[21] In one of these the sun rose, in another it didn't. Laplace's rule then says the probability the sun rises tomorrow is given by the *fraction of all past observations* (both virtual and actual) in which the sun rose.

[20] Mathematically, we are modelling each day as a discrete independent trial with probability *p* of success - a Bernoulli distribution. We then update our estimate of *p* according to Bayes' rule. See details here.

[21] I believe the idea of virtual observations was first introduced by Rudolf Carnap in *The Logical Foundations of Probability*. Though it is funny to imagine that we have already made *virtual observations* before we start making actual observations, I find it a useful way to think about Laplace's rule.



2.2 Applying Laplace's rule of succession to AGI timelines

We want to estimate pr(AGI by 2036).[22] Rather than observing that the sun has risen for *N* days, we have observed that AI researchers have not developed AGI with *N* years of effort. The field of AI research is widely held to have begun in Dartmouth in 1956, so it is natural to take *N* = 64. (The choice of a year - rather than e.g. 1 month - is arbitrary and made for expositional purposes. The results of this report don't depend on such arbitrary choices, as discussed in the next section.)

By analogy with the sunrise problem, we assume there's been some constant but unknown probability *p* of creating AGI each year. We place a uniform prior probability distribution over *p* to represent our uncertainty about its true value, and update this distribution for each year that AGI hasn't happened.[23]

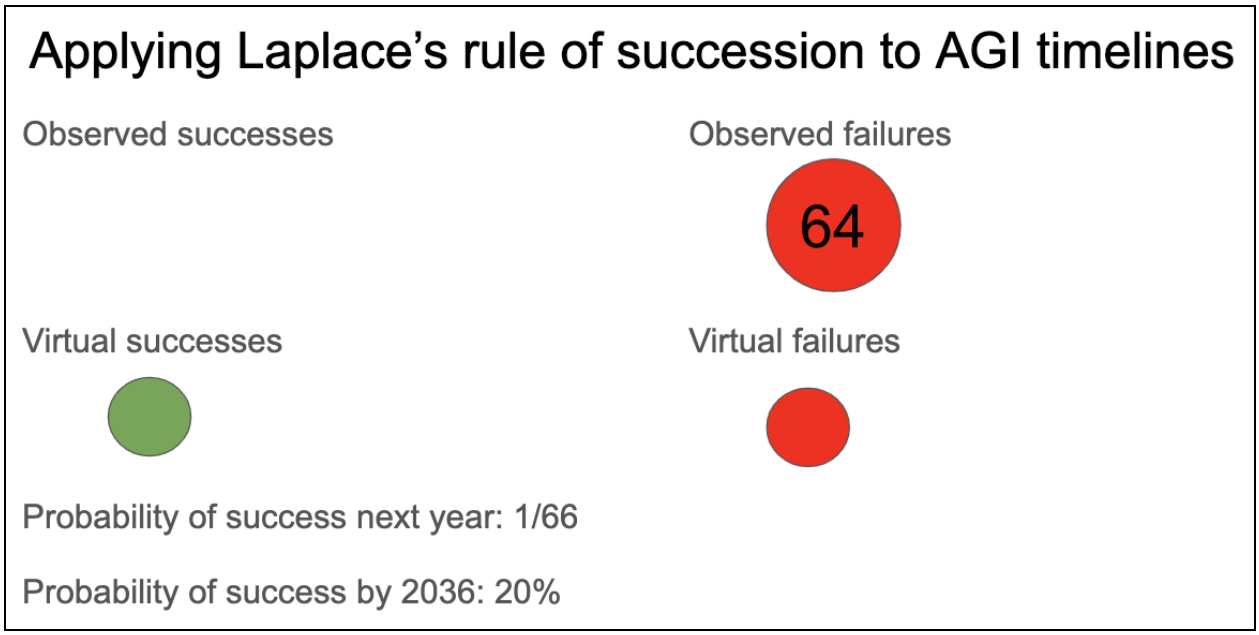
The rule of succession implies that the chance AGI will again not be developed on the next trial is *(N + 1) / (N + 2)* = 65/66. The chance it will not be developed in any of the next 16 trials is 65/66 * 66/67 * … * 81/82 = 0.8, and so **pr(AGI by 2036) = 0.2.**

---

[22] I am not the first person to apply an uninformative prior to AGI timelines. Firstly, Fallenstein and Mennen use Laplace's rule as one argument in favour of using a pareto probability distribution over pr(AGI by year X) - see section 3.3 of the linked paper. Their favoured distribution is the continuous limit of the one reached by applying Laplace's rule. Secondly, informal discussions about AGI timelines on the Effective Altruist Forum, and in the community more generally, have sometimes referenced the rule (e.g. here).

[23] This update is made according to Bayes' rule. Suppose that before observing a trial our initial credence in each value of *p* is given by $pr_i(p=x)$. After observing a failed trial X, our final credence in each possible value of *p* is $pr_f(p=x) = pr_i(p=x|X) = pr_i(p=x) * pr_i(X|p=x) / pr_i(X)$. The relative weight of our credence in *p*=x is boosted by the factor $pr_i(X|p=x) = (1 - p)$. We perform an update of this kind for every failed trial.

An equivalent way to think about this calculation is that, after observing 64 failed trials, our belief about chance of success in the next trial E(p) is 1/66. This is the fraction of our actual and virtual observations that are successes. So our probability of developing AGI next year is 1/66. We combine the probabilities for the next 16 years to get the total probability of success.



The next section discusses some significant problems with this application of Laplace's rule of succession. These problems will motivate a more general framework, in which this calculation is a special case.

3 Semi-informative priors over AGI timelines

This section motivates and explores the semi-informative priors framework in the context of AGI timelines. In particular:

- I introduce the framework by identifying various debatable inputs in our previous application of Laplace's rule (here).
- I explain how the semi-informative priors framework addresses problems with applying Laplace's rule to AGI timelines (here).
- I describe key properties of the framework (here).
- I perform a sensitivity analysis on how pr(AGI by 2036) depends on each input (here).

This lays the groundwork for sections 4-6 which apply the framework to AGI timelines.

3.1 Introducing the semi-informative priors framework

Our application of Laplace's rule of succession to calculate pr(AGI by 2036) had several inputs that we could reasonably change.

First, the calculation identified the start of a regime such that the failure to develop AGI before the regime tells us very little about the probability of success during the regime. This *regime start-time* was 1956. This is why we didn't update our belief about *p* based on AGI not being developed in the years prior to 1956. Though 1956 is a natural choice, there are other possible *regime start-times.*

Second, we assumed that each trial (with constant probability *p* of creating AGI) was a calendar year. But there are other possible *trial definitions*. Alternatives include 'a year of work done by one researcher', and 'a doubling of the compute used in AI R&D'. With this latter alternative, the model would assume that each doubling of compute costs was a discrete event with a constant but unknown probability *p* of producing AGI.[24]

Third, we assumed that an appropriate initial distribution over *p* was uniform over [0, 1]. But there are many other possible choices of distribution. The Jeffreys prior over *p*, another uninformative distribution, is more concentrated at values close to 0 and 1, reflecting the idea that many events are almost certain *to* happen or certain *not* to happen. It turns out that the difference between these two distributions corresponds to the *number of virtual successes* we

[24] We will see in the next subsection that it doesn't affect the results whether a trial is a 'doubling' of compute, a '10% increase', or a '0.1% increase'. As long as there have been many (>5) such trials since the regime start-time, and many trials (>5) between now and 2036, pr(AGI by 2036) will be very similar. For the same reason, it would make ~0 difference to change the trial definition from 'a calendar year' to "2 calendar years" or "10 minutes". Throughout I use fine-grained trial definitions (e.g. a 0.1% increase) to ensure none of the results are a product of an unrealistic discretisation of effort.

observed before the regime started. While Laplace has 1 virtual success (and 1 virtual failure), Jeffreys has just 0.5 virtual successes (and 0.5 failures) and so these virtual observations are more quickly overwhelmed by further evidence. The significance of this input is that the fewer *virtual successes*, the quicker you update E(*p*) towards 0 when you observe failed trials.

Lastly, and most importantly, both Laplace and Jeffreys initially have E(*p*) = 0.5, reflecting an initial belief that the first trial of the regime is 50% likely to create AGI. Call this initial value of E(*p*) the *first-trial probability*. The *first-trial probability* is the probability that the first trial succeeds. There are different initial distributions over *p* corresponding to different *first-trial probabilities*. Both Laplace's uniform distribution and the Jeffreys prior over *p* are specific examples of beta distributions,[25] which can in fact be parameterised by the *first-trial probability* and the *number of virtual successes*.[26] Roughly speaking, the *first-trial probability* represents how easy you expect developing AGI to be before you start trying; more precisely, it gives the probability that AGI is developed on the first trial.

If you find thinking about virtual observations helpful, the *first-trial probability* gives the fraction of virtual observations that are successes:

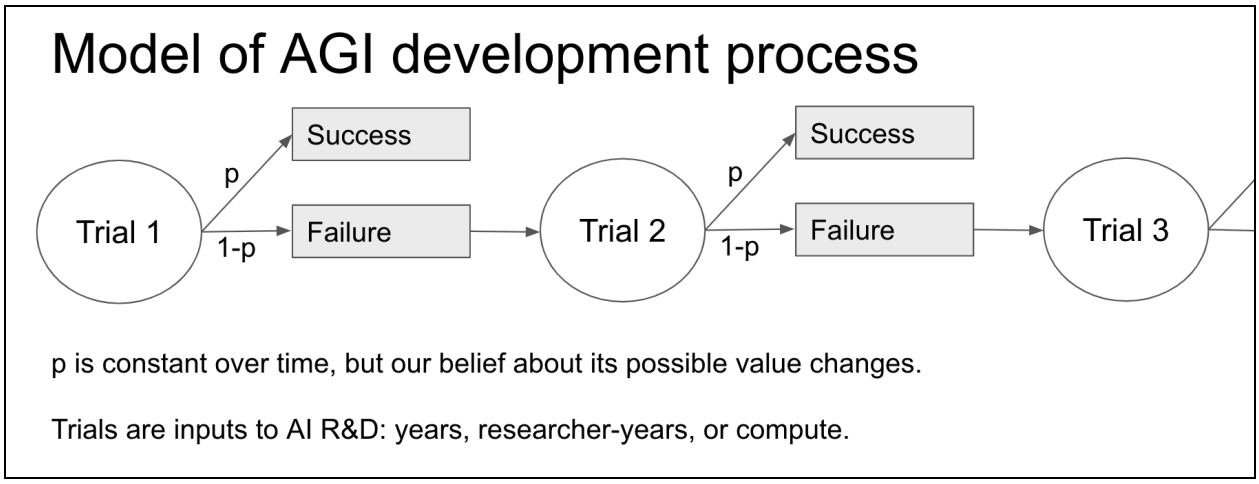
$$first\ trial\ probability\ =\ (\#\ virtual\ successes)\ /\ (\#\ virtual\ successes\ +\ \#\ virtual\ failures).$$

So we have 4 inputs to our generalised update rule (Laplace's values in brackets):

[25] Beta distributions are useful in this context because they're easy to update in response to evidence from Bernoulli trials. Formally, this is because beta distributions are conjugate priors to Bernoulli distributions. I am not the first to suggest replacing Laplace's uniform distribution with the more general, yet analytically tractable, beta distribution. For example see Huttegger (2017); Raman (2000); Bernardo and Smith (1994), p271-272, example 5.4, 2nd edition.

[26] The beta distribution is only well-defined if *virtual successes* > 0 and 0 < *first-trial probability* < 1.

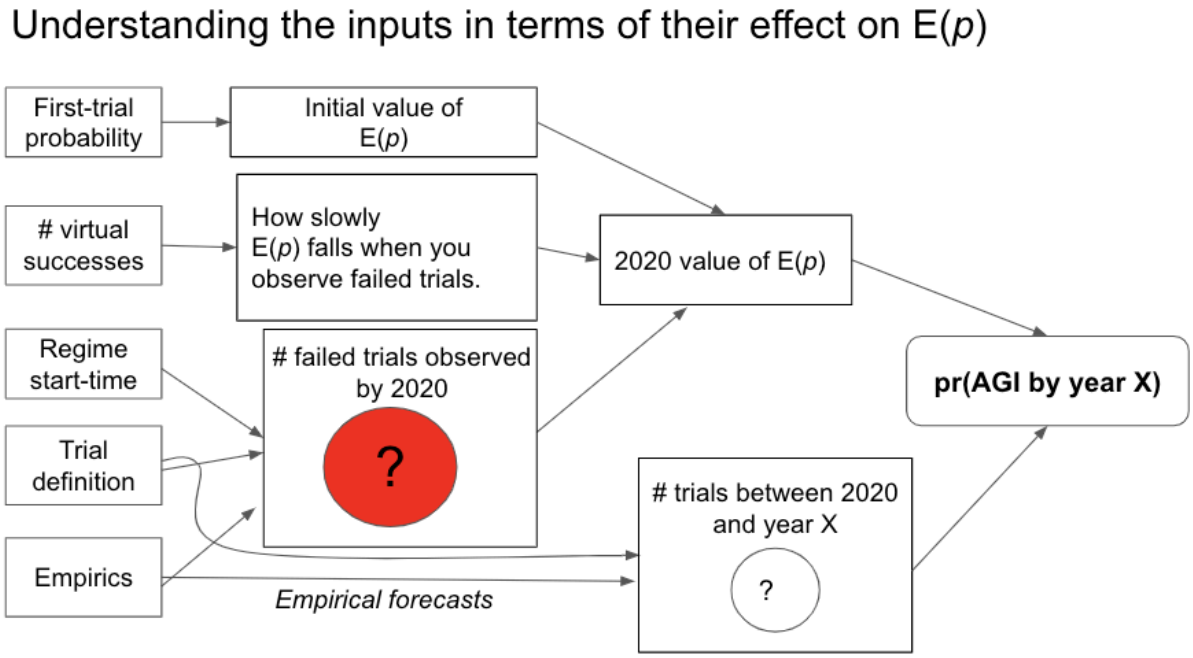
- *Regime start-time* (1956)
- *Trial definition* (calendar year)
- *Number of virtual successes* (1)
- *first-trial probability* (0.5)



I find it useful to think about these inputs in terms of how E(*p*), our belief about the probability of success in the next trial, changes over time.[27] The *first-trial probability* specifies the initial value of E(*p*) and the *number of virtual successes* describes how quickly E(*p*) falls when we observe failed trials.[28] The *regime start-time* and *trial definition* determine how many failed trials we've observed to date; for some trial definitions (e.g. 'one researcher-year') we also need empirical data. The *trial definition*, perhaps in conjunction with empirical forecasts, also determines the number of trials that will occur in each future year. Together the four inputs determine a probability distribution over the year in which AGI will be developed. When the choice of inputs are informed by commonsense and relevant reference classes for AGI, I call such a distribution a *semi-informative prior* over AGI timelines. We will see that some highly subjective judgements seem to be needed to choose precise values for the inputs.

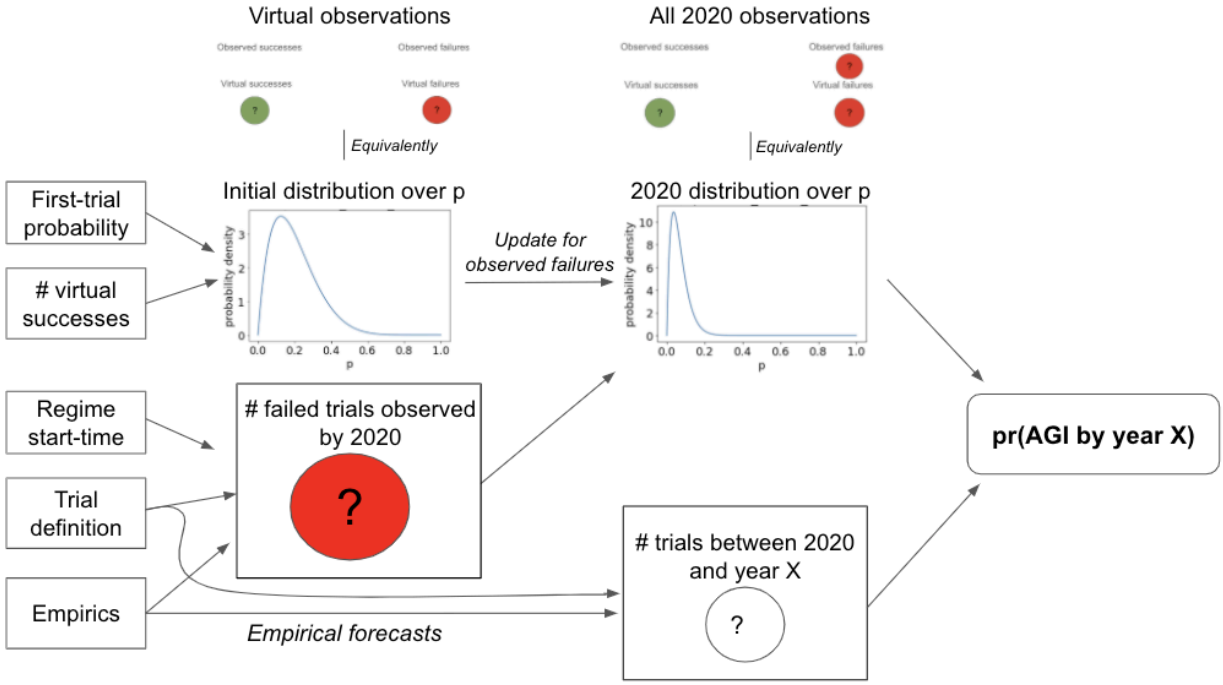
[27] Another equally valid perspective is that the inputs determine an unconditional probability distribution over the total amount of time needed to develop AGI. I take this alternative perspective in section 9, where I argue the family of unconditional probability distributions corresponding to the framework are sufficiently general for our purposes.

[28] Why not use the total number of *virtual observations* (both successes and failures) as an input, rather than just the number of *virtual successes*? Both influence the size of the update from failed trials. Essentially, because using *virtual observations* has the consequence that the *first-trial probability* also affects the size of the update from failed trials. Using *virtual successes*, we avoid this consequence and only the number of *virtual successes* affects the size of the update from failed trials. I explain this in greater detail here.



To use this framework to calculate pr(AGI by 2036) you need to choose values for each of the four inputs, estimate the number of trials that have occurred so far and estimate the number that will occur by 2036. I do this, and conduct various sensitivity analyses in sections 4, 5 and 6. The rest of section 3 explores the behaviour of the semi-informative framework in more detail.

The following diagram gives a more detailed mathematical view of the framework:



The **first-trial probability** and **# virtual successes** determine your initial probability distribution over *p*. This initial distribution corresponds to the number of virtual successes and virtual failures. The start-time and trial definition determine the number of observed failures by 2020. Updating on these failures creates you 2020 probability distribution over *p*. The 2020 distribution, together with the number of trials between 2020 and year X, determines pr(AGI by year X).]

3.2 The semi-informative priors framework can solve problems with using uninformative priors

This section identifies two problems with the application of Laplace's rule of succession to AGI timelines, and argues that both can be addressed by the semi-informative priors framework.

3.2.1 Uninformative priors are aggressive about AGI timelines

Before the first trial, an uninformative prior implies that E(*p*) is 0.5.[29] So our application of uninformative priors to AGI timelines implies that there was a 50% probability of developing in AGI in the first year of effort. Worse, it implies that there was a 91%[30] probability of developing AGI in the first ten years of effort.[31] The prior is so uninformative that it precludes the commonsense knowledge that highly ambitious R&D projects rarely succeed in the first year of effort![32]

The fact that these priors are *initially* overly optimistic about the prospects of developing AGI means that, after updating on the failure to develop it so far, they will *still* be overly optimistic. For if we corrected their initial optimism by reducing the *first-trial probability*, the derived pr(AGI by 2036) will also decrease as a result. Their unreasonable initial optimism translates into unreasonable optimism about pr(AGI by 2036).

To look at this from another angle, when you use an uninformative prior the *only* source of scepticism that we'll build AGI next year is the observed failures to date. But in reality, there are other reasons for scepticism: the bare fact that ambitious R&D projects typically take a long time means that the prior probability of success in any given year should be fairly low.

---

[29] This is not just true of Laplace's uniform prior. The same is true of the two other common choices of uninformative prior over *p* for a Binomial distribution: the Jeffreys prior and the Haldane prior.

[30] 91% = 1 - (1/2 * 2/3 * … * 10/11)

[31] Again this problem is not particular to Laplace's uninformative prior. The Jeffreys prior similarly implies 82% likely to develop AGI in the first 10 years. (1 - ½ * ¾ * ⅝ *... * 19/20). This is lower, but still unreasonably high. The uninformative Haldane prior suffers from a different problem. Like the naive frequentist approach, it is certain that we won't ever develop AGI simply because we do not do so in the first year of effort.

[32] One way to think about this is that the uninformative prior abstracts away so much knowledge that it treats the proposition "we *do* develop AGI" in exactly the same way as it treats "we *don't* develop AGI", initially estimating E(*p*) = 0.5 for both.

In the semi-informative priors framework, we can address this problem by choosing a lower value for the *first-trial probability*. In this framework there are two sources of scepticism that we'll build AGI in the next trial: the failure to develop AGI to date *and* our initial belief that a given year of effort is unlikely to succeed.

3.2.2 The predictions of uninformative priors are sensitive to trivial changes in the trial definition

A further problem is that certain predictions about AGI timelines are overly sensitive to the *trial definition*. For example, if I had defined a trial as two years, rather than one, Laplace's rule would have predicted a 83%[33] probability of AGI in the first 10 years rather than 91%. If I had used one month, the probability would have been 99%.[34] But predictions like these should not be so sensitive to trivial changes in the trial definition.[35] Further, there doesn't seem to be any privileged choice of trial definition in this setting.

This problem can be addressed by the semi-informative priors framework. We can use a procedure for choosing the *first-trial probability* that makes the framework's predictions invariant under trivial changes in the trial definition. For example, we might choose the *first-trial probability* so that the probability of AGI in the first 20 years of effort is 10%. In this case, the model's predictions will not materially change if we shift our trial definition from 1 year to (e.g.) 1 month: although there will be more trials in each period of time, the *first-trial probability* will be lower and these effects cancel.[36]

In fact, using common sense and analogous reference classes to select the *first-trial probability* naturally has this consequence. Indeed, all the methods of constraining the *first-trial probability* that I use in this report are robust to trivial changes in the trial definition.

---

[33] = 1 - (1/2 * 2/3 * … * 5/6) = 99.2%

[34] = 1 - (1/2 * 2/3 * … * 120/121)

[35] By a 'trivial change' I mean a change where we keep the quantity the same (in this case *time*) but divide this quantity into slightly more (or less) fine-grained chunks. Such a change does not change the ratio (# trials in period 1) / (# trials in period 2) for any two periods of time. An example of *non*-trivial change is from 'trial = a calendar year' to 'trial = one researcher-year'. This isn't trivial because the underlying quantity has changed.

[36] Trivial changes in trial definition can still have small effects on the model's predictions by discretizing a continuous quantity in different-sized chunks. Throughout the report I use small enough chunks that this effect disappears.
In more detail, the effect of any trivial change is negligible as long as both trial definitions are sufficiently fine-grained. In particular, if for both trial definitions there have been >5 trials since the regime start-time and >5 trials between now and 2036, then changing the trial definition will have a negligible effect on pr(AGI by 2036). By "negligible" I mean that pr(AGI by 2036) will be the same to 2 significant figures. Throughout this report I subdivide trials into small parts to ensure none of the results are a product of an unrealistic discretisation of effort. The subdivisions are small enough that further subdivisions would make no difference to any of the numbers in the report. Similarly, taking the framework to the continuous limit would make no difference to the results.

3.3 How does the semi-informative priors framework behave?

There are a few features of this framework that it will be useful to keep in mind going forward.

- **If your first-trial probability is smaller, your update from failure so far is smaller.** If it takes 100 failures to reduce E(*p*) from 1/100 to 1/200, then it takes 200 failures to reduce E(*p*) from 1/200 to 1/400, holding the number of *virtual successes* fixed.[37]
- **The *first-trial probability* is related to the median number of trials until success.** Suppose your *first-trial probability* is 1/N and there's 1 *virtual success*. Then, it turns out, the probability of success within the first (N - 1) trials is 50%.[38]
- **E(*p*) is initially dominated by the *first-trial probability*; after observing many failures it's dominated by your observed failures.** Suppose your *first-trial probability* is 1/N and you have *v virtual successes*. After observing *n* failures, it turns out that $E(p) = 1/(N + n/v)$. For small values of n, E(*p*) is approximately equal to the *first-trial probability*. For large values of n, $n/v >> N$, E(*p*) is dominated by the update from observed failures.

3.4 Strengths and weaknesses

Here are some of the framework's strengths:

- **Quantifies the size of the negative update from failure so far.** We can compare the initial value of E(p) with its value after updating on the failed trials observed so far. The ratio between these values quantifies the size of the negative update from failure so far.
- **Highlights the role of intuitive parameters.** The report's analysis reveals the significance of the *first-trial probability*, *regime start-time*, the *trial definition*, and empirical assumptions for the bottom line. These are summarised in the conclusion.
- **Arguably appropriate for expressing deep uncertainty about AGI timelines.**
    - The framework produces a long-tailed distribution over the total time for AGI, reflecting the possibility that AGI will not be developed for a very long time. More.
    - The framework can express Pareto distributions (more), exponential distributions (more), and uninformative priors as special cases.
    - The framework spreads probability mass fairly evenly over trials.[39] For example, it couldn't express the belief that AGI will probably be developed between 2050 and 2070, but not in the periods before or after this.

---

[37] This example assumes 1 *virtual success*. If there are 0.5 *virtual successes*, it takes 50 failures to reduce E(*p*) from 1/100 to 1/200, and 100 failures to reduce E(*p*) from 1/200 to 1/400.

[38] There are similar results if we vary the *number of virtual successes*. For example, if there's 0.5 *virtual successes*, then 1 / *first-trial probability* is roughly the time until there's a 42% chance of success, rather than 50%. More generally, the fewer *virtual successes* we use, the smaller *x* is in the following: 1 / *first-trial probability* is the time until x% chance of success.

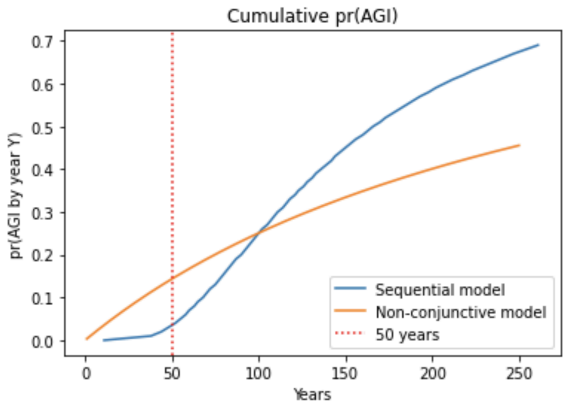
[39] The one exception to this if you use a large *first-trial probability*, >1/10. In this case, probability is concentrated in the first few trials. However, I recommend using a *first-trial probability* below 1/100.

    - The framework avoids using anything like “we’re x% of the way to completing AGI” or “X of Y key steps on the path to AGI have been completed.” This is attractive if you believe we are not in a position to make more direct judgments about these things.

Here are some of the framework’s weaknesses:

- **Incorporate limited kinds of evidence.**
    - The framework excludes evidence relating to how close we are to AGI and how quickly we are getting there. For some, this is the most important evidence we have.
    - It excludes knowledge of an *end-point*, a time by which we will have probably developed AGI. So it cannot express (log-)uniform distributions. More.
    - Evidence only includes the binary fact we haven’t developed AGI so far, and information from relevant reference classes about how hard AGI might be to develop.
- **Near term predictions are too high**. Today’s best AI systems are not nearly as capable as AGI, which should decrease our probability that AGI is developed in the next few years. But the framework doesn’t take this evidence into account.
- **Insensitive to small changes in the definition of AGI.** The methods I use to constrain the inputs to the framework involve subjective judgements about vague concepts. If we changed the definition of AGI to make it slightly easier/harder to achieve, the judgements might not be sensitive to these changes.
- **Assumes a constant chance of success each trial**. This is of course unrealistic; various factors could lead the chance of success to vary from trial to trial.
    - The assumption is more understandable given that the framework purposely excludes evidence relating to the details of the AI R&D process.
    - Appendix 12 argues that our results are driven by our choice of inputs to the framework, not by the framework itself. If this is right, then relaxing the problematic assumption would not significantly change our results.
    - Indeed, I analysed sequential models in which multiple steps must be completed to develop AGI. pr(next trial succeeds) is very low in early years, rises to a peak, and then slowly declines. I compared my framework to a sequential model, with the inputs to both chosen in a similar way. Although pr(next trial succeeds) is initially much lower for the sequential model, after a few decades the models agree within a factor of 2. This is shown by the similar steepness of the lines.[40]

[40] Inputs for both models are chosen to give the same value for pr(AGI within the 100 years). After about 50 years, the sequential model has a higher value for pr(next trial succeeds), but still within a factor of 2.



The reason is that the sequential models are agnostic about how many steps still remain in 2020; for all they know just 1 step remains! Such agnostic sequential models have similar pr(AGI by year X) to my framework once enough time has passed that all the steps might have been completed.

- That said, the argument in appendix 12 is not conclusive and I only analysed a few possible types of sequential model. It is possible that other ways of constructing sequential models, and other approaches to outside view forecasting more generally, may give results that differ more significantly from our framework.

3.5 How do the inputs to the framework affect pr(AGI by 2036)?

How does pr(AGI by year X) depend on the inputs to the semi-informative priors framework? I did a sensitivity analysis around how varying each input within a reasonable range alters pr(AGI by 2036); the other inputs were left as in the initial Laplacean calculation.

The values in this table are not trustworthy because they use a *first-trial probability* of 0.5, which is much too high. I circle back and discuss each input's effect on the bottom line in section 8. Nonetheless, the table illustrates that the *first-trial probability* has the greatest potential to make the bottom line very low, and its uncertainty spans multiple orders of magnitude. This motivates an in-depth analysis of the *first-trial probability* in the next section.

Input	**Values tested**	**Range for pr(AGI by 2036)**	**Comments**
Regime start-time	1800 - industrial revolution 1954 - Dartmouth conference 2000 - brain-compute affordable (explained in section 5)	[0.07, 0.43]	I discuss that even earlier regime start-times in section 5. 0.43 corresponds to '2000'. When the *first-trial probability* is constrained within reasonable bounds, this range is much smaller.
Trial definition	- A calendar year - A researcher-year - 1% increase in total researcher-years so far (See explanations of these definitions here)	[0.14, 0.71]	0.71 corresponds to 'a researcher-year' When the *first-trial probability* is constrained within reasonable bounds, this range is much smaller.
Number of virtual successes	0.5, 1	[0.11, 0.2]	I explain why I prefer the range [0.5, 1] for the case of AGI in section 5.
first-trial probability	0.5, 0.1, 1e-2, 1e-3, 1e-4	[1/1000, 0.2]	

The next section, section 4, discusses how we might constrain the *first-trial probability* for AGI; it also implicitly argues that it was reasonable for me to countenance such small values for *first-trial probability* in this sensitivity analysis. After this, section 5 revisits the importance of the other inputs. Both sections 4 and 5 assume that a trial is a calendar year; in section 6 we consider other trial definitions.

4 Constraining the *first-trial probability*

The sensitivity analysis in the previous section suggested that the *first-trial probability* was the most important input for determining pr(AGI by 2036). This section explains my preferred methodology for choosing the *first-trial probability* (here) and then makes an initial attempt to put this methodology into practice in the case of AGI (here).

4.1 How to constrain the *first-trial probability* in principle

One compelling way to constrain the *first-trial probability* for a project's duration would be as follows:

1. List different reference classes that seem potentially relevant to the project's likely difficulty and duration. Each reference class will highlight different features of the project that might be relevant.
2. For each of these reference classes, try to constrain or estimate the *first-trial probability* using a mixture of data and intuitions. This leaves you with one constraint for each reference class. These constraints should be interpreted *flexibly*; they are merely suggestive and can be overridden by other considerations.
3. Weight each constraint by how relevant you think its reference class is to the project. Then, either by taking a formal weighted sum or by combining the individual constraints in an informal way, arrive at an all-things-considered constraint of the *first-trial probability*.

To illustrate this process, I'll give a ***brief toy example*** with ***made-up numbers*** to show what these steps might look like when the project is developing AGI. To make the example short, I've removed most of the reasoning that would go into a comprehensive analysis, leaving only the bare bones.

1. List multiple different reference classes for the development of AGI:
    a. '*Hard computer science problem*' - the frequency with which such problems are solved is potentially relevant to the probability that developing AGI, an example of such a problem, is completed.
    b. '*Development of a new technology that leads to the automation of a wide range of tasks*' - the frequency at which such technologies are developed is potentially relevant to the probability that AGI, an example of such a technology, is developed.
    c. '*Ambitious milestone that an academic STEM field is trying to achieve*' - the time it typically takes for such fields to succeed is potentially relevant to the probability that the field of AI R&D will succeed.
2. Constrain the *first-trial probability* for each reference classes:

   a. Data about hard computer science problems suggests about 25% of such problems are solved after 20 years of effort. (These numbers are made up.) On the basis of this reference class, we should choose AGI's *first-trial probability* so that the chance of success in the first 20 years is close to 25%. This corresponds to a *first-trial probability* of 1/61. **So this reference class suggests that the *first-trial probability* be close to 1/60.**
   b. Data about historical technological developments suggest that developments with an impact on automation comparable to AGI occur on average less often than once each century.[41] So our probability that such a development occurs in a given year should be less than 1%. On the basis of this reference class, we should choose AGI's *first-trial probability* so that the chance of success each year is <1%. **So this reference class suggests that the *first-trial probability* be <1/100**.
   c. Data about whether STEM fields achieve ambitious milestones they're trying to achieve seems to suggest it is not that rare for fields to succeed after only a few decades of sustained effort. On the basis of this reference class, we should choose AGI's *first-trial probability* so that the chance of success in the first 50 years is >5%. This implies *first-trial probability* >1/950. **So consideration of this reference class suggests that the *first-trial probability* should be >1/1000**.
3. To reach an all-things-considered view on AGI's *first-trial probability*, weigh each constraint by how relevant you think the associated reference class is to the likely difficulty and duration of developing AGI. For example, someone might think the latter two classes are both somewhat relevant but put less weight on "*hard computer science problem*" because they think AGI is more like a large collection of such problems than any one such problem. As a consequence, their all things-considered view might be that AGI's *first-trial probability* should be >1/1000 and <1/100.

This is just a brief ***toy example*** (again, with ***made-up numbers***) to illustrate what my preferred process for constraining the *first-trial probability* might look like. Clearly, difficult and debatable judgement calls must be made in all three steps. In the first step, a short list of relevant reference classes must be identified. In the second step, data about the reference class must be interpreted to derive a constraint for the *first-trial probability*. In the third step, judgement calls must be made about the relevance of each reference class and the individual constraints must be combined together.

It may be that no reference class both has high quality data and is highly relevant to the likely duration of developing AGI. In this case, my preference is to make the most of the reference classes and data that is available, interpreting the derived constraints as no more than suggestive. It may be that by making many weak arguments, each with a different reference class, we can still obtain a meaningful constraint on our all-things-considered *first-trial*

[41] Of course, if you think that AGI automation would be more transformative than any previous technological development then this would complicate this argument. I discuss this consideration in this supplement, which discusses a methodology for this reference class in more detail.

probability. Even if we do not put much weight in any particular argument, multiple arguments collectively may help us triangulate what values for the *first-trial probability* are reasonable.

4.2 Constraining AGI's *first-trial probability* in practice

The *first-trial probability* should of course depend on the *trial definition*. For example, the *first-trial probability* should be higher if a trial is '5 calendar years' than if it's '1 calendar year'; it should be different again if a trial is 'a researcher-year'. In this section I assume that a trial is 'one calendar year of sustained AI R&D effort',[42] which I abbreviate to '1 calendar year'. I also assume that the *regime start-time* is 1956 and the *number of virtual successes* is 1; I consider the effects of varying these inputs in the next section.

The focus of this project has been in the articulation of the semi-informative priors framework, rather than in finding data relevant for constraining the *first-trial probability*. As such, I think all of the arguments I use to constrain the *first-trial probability* are fairly weak. In each case, either the relevance of the reference class is unclear, I have not found high quality data for the reference class, or both. Nonetheless, I have done my best to use readily available evidence to constrain my *first-trial probability* for AGI, and believe doing this has made my preferred range more reasonable.

I currently favour values for AGI's *first-trial probability* in the range [1/1000, 1/100], and my central estimate is 1/300.

This preferred range is informed by 4 reference classes. In each case, I use the reference class to argue for some constraint on, or estimate of, the *first-trial probability*. The 4 reference classes were not chosen because they are the most relevant reference classes to AGI, but because I was able to use them to construct constraints for AGI's *first-trial probability* that I find somewhat meaningful. While I extract inequalities or point estimates of the *first-trial probability* from each reference class, my exact numbers shouldn't be taken seriously and I think one could reasonably differ by at least a factor of 3 in either direction, perhaps more. Further, people might reasonably disagree with my views on the relevance of each reference class.

I explain my thinking about each reference class in detail in supplementary documents that are linked individually in the table below. These supplementary documents are designed to *help the reader use their own beliefs and intuitions to derive a constraint* from each reference class. I encourage readers use these to construct their own constraints for AGI's *first-trial probability*. **Much more work could be done finding and analysing data to better triangulate the *first-trial probability*, and I'd be excited about such work being done.**

[42] This means that if no research happens in some year, no trial occurs in that year.

The following table summarizes how the 4 reference classes inform my preferred range for the *first-trial probability*. Please keep in mind that I think all of these arguments are fairly weak and see all the constraints and point estimates as merely suggestive.

Reference class	**Argument deriving a constraint on the *first-trial probability (ftp)***	**Constraints and estimates of ftp**	**My view on the informativeness of this reference class**
Ambitious but feasible technology that a serious STEM field is explicitly trying to develop. (more)	Scientific and technological R&D efforts have an excellent track record of success. Very significant advances have been made in central and diverse areas of human understanding and technology: physics, chemistry, biology, medicine, transportation, communication, information, and energy. I list 11 examples, with a median completion time of 75 years. Experts regard AGI as feasible in principle. Multiple well-funded and prestigious organisations are explicitly trying to develop AGI. Given the above, we shouldn't assign a very low probability to the serious STEM field of AI R&D achieving one of its central aims after 100 years of sustained effort.	Lower bound: ***ftp* > 1/3000** -- pr(AGI within 100 years of effort) >3%, or pr(AGI within 30 years of effort) >1%. Conservative estimate: ***ftp* = 1/300** -- pr(AGI within 100 years of effort) = 25%. Optimistic estimate: ***ftp* = 1/50** -- pr(AGI within 50 years of effort) = 50%.	In my view, this is the most relevant reference class of the 4 that I consider. The fact that a serious STEM field is trying to build AGI is clearly relevant to AGI's probability of being developed. That said, STEM fields vary in their degree of success and AGI may be an especially ambitious technology, reducing the relevance of this reference class. There is also a selection bias in the list of successful STEM fields (that I try to adjust for in the conservative estimate).
Possible future technology that a STEM field is trying to build in 2020. (more)	This report focuses on AGI and its core reason for having a non-tiny *first-trial probability* is that a STEM field is trying to build AGI. But we could apply the same framework to multiple different technologies that STEM fields are trying to build in 2020. It would be worrying if, by doing this many times, we could deduce that the expected	Conservative upper bound: ***ftp* < 1/100** -- assumes that STEM fields are trying to build 10 highly impactful technologies in 2020, but we expect <0.5 technologies with this level of	In principle, I think this reference class is highly relevant. We shouldn't trust this methodology if applying it elsewhere leads to unrealistic predictions. In practice, however, it's hard to make this objection cleanly for various reasons. As

	number of transformative technologies that will be developed in a 10 year period is very large. We can avoid this problem by placing an upper bound on the *first-trial probability*.	impact to be developed in a ten year period on average) Aggressive upper bound: ***fpt < 1/300*** -- As above but expect <0.25 to be developed	such, I put very little stock in the precise numbers derived. I'm unsure what constraint a more comprehensive analysis would suggest.
Technological development that has a transformative effect on the nature of work and society. (*more*)	Some people believe that AGI would have a transformative effect on the nature of work and society. We can use the history of technological developments to estimate the frequency with which *transformative developments* like AGI occur. This frequency should guide the probability p_{transf} we assign to a transformative development occurring in a given year. Our annual probability that AGI is developed should be lower than p_{transf}, as it's less likely that AGI in particular is developed than that *any* transformative development occurs.	Upper bound: *ftp < 1/130* -- Assume two transformative events have occurred. Assume the probability of a transformative development occurring in a year is proportional to the amount of technological progress in that year.	I believe that a technology's impact is relevant to the likely difficulty of developing it (more). So I find this reference class somewhat informative. Further, a common objection to AGI is that it would have such a large impact so is unrealistic. This reference class translates this objection into a constraint on the *ftp*. However, there are very few (possibly zero) examples of developments with impact comparable to AGI; this makes this reference class less informative.
Notable mathematical conjectures. (*more*)	AI Impacts investigated how long notable mathematical conjectures, not explicitly selected for difficulty, take to be resolved. They found that the probability that an unsolved	*ftp ~ 1/170*	The data for this reference class is better than for any other. However, I doubt that resolving a mathematical conjecture is similar to developing AGI. So I

	conjecture is solved in the next year of research is ~1/170.		view this as the least informative reference class.

The following table succinctly summarizes the most relevant inputs for forming an all-things-considered view.

Reference class	**Constraints and point estimates of the *first-trial probability (ftp)***	**Informativeness**
Ambitious but feasible technology that a serious STEM field is explicitly trying to build. (link)	Lower bound: *ftp* > 1/3000 Conservative estimate: *ftp ~ 1/300* Optimistic estimate: *ftp ~ 1/50*	Most informative.
High impact technology that a serious STEM field is trying to build in 2020. (link)	Conservative upper bound: *ftp < 1/100* Aggressive upper bound: *fpt < 1/300*	Weakly informative.
Technological development that has a transformative effect on the nature of work and society. (link)	Upper bound: *ftp < 1/130*	Somewhat informative.
Notable mathematical conjectures. (link)	*ftp ~ 1/170*	Least informative.

I did not find it useful to use a precise formula to combine the constraints and point estimates from these four reference classes. Overall, I favour a *first-trial probability* in the range **[1/1000, 1/100]**, with a preference for the higher end of that range.[43] If I had to pick a number I'd go with **~1/300**, perhaps higher.

The numbers I've derived depend on *subjective* choices about which references classes to use (reviewers suggested alternatives[44]), how to interpret them (the reference classes

[43] One other reason to prefer this range is that *first-trial probabilities* much higher than 1/100 lead to very large updates towards thinking AGI is impossible, based on the failure to develop it by 2020. See this table.

[44] Here are some alternative reference classes that reviewers suggested: crazy-hard ancient ambitions (e.g. immortality, teleportation, mind control, soothsaying, turning lead in gold cheaply, flying to the moon); hard business problems; years between inception of a technology and realisation (e.g. computers – Babbage to Turing, flight – ancient Greeks to Wright Brothers); process of automation (e.g. identify AGI with the point at which 90% of current-day-jobs are automated); growth of the world economy (e.g. identify AGI with passing some milestone in Gross World Product). I'm not sure exactly what first-trial probabilities these reference classes would suggest if someone worked them through.

are somewhat vague[45]), and how relevant they are to AGI. I did my best to use a balanced range of reference classes that could drive high and low values. These subjective judgements would probably not be sensitive to small changes in the definition of AGI (see more).

The following table shows how different *first-trial probabilities* affect the bottom line, assuming that 1 *virtual success* and a *regime start-time* of 1956.[46]

first-trial probability	**pr(AGI by 2036)**
1/50	12%
1/100	8.9%
1/200	5.7%
1/300	4.2%
1/500	2.8%
1/1000	1.5%
1/2000	0.77%
1/3000	0.52%

(Throughout this report, I typically give results to 2 significant figures as it is sometimes useful for understanding a table. However, I don't think precision beyond 1 significant figure is meaningful.)

Based on the table and my preferred range for the *first-trial probability*, my preferred range for pr(AGI by 2036) is **1.5 - 9%**, with my best guess around 4%. I will be refining this preferred range over the course of the report. (At each time, I'll refer to the currently most refined estimate as my "preferred range," though it may continue to change throughout the report.)

[45] They're defined using vague words like "ambitious", "high impact", "serious".

[46] Notice that when the first-trial probability is very small, doubling it roughly doubles pr(AGI by 2036) - this is because the update from failure so far makes very little difference. But after the first-trial probability exceeds 1/300, doubling it less than doubles pr(AGI by 2036) - this is because the update from failure is more significant when the first-trial probability is bigger. Another way to think about this is that the update from failure puts a cap at how large pr(AGI by 2036) can be. That cap is 20%, which is reached for very high *first-trial probabilities* like that of Laplace's rule of succession. When the cap is near to being approached, doubling the *first-trial probability* less than doubles pr(AGI by 2036). For very high *first-trial probabilities* above 1/20, doubling them makes very little difference to pr(AGI by 2036).

5 Importance of other inputs

The semi-informative priors framework has four inputs:

- *Regime start-time*
- *Trial definition*
- *Number of virtual successes*
- *first-trial probability*

In the previous section we assumed that the *regime start-time* was 1956, the *number of virtual successes* was 1, and the *trial definition* was a 'calendar year'. I then suggested that a reasonable *first-trial probability* for AGI should probably be in the range [1/1000, 1/100]. This corresponded to a bottom line pr(AGI by 2020) in the range [1.5%, 9%].

In this section, I investigate how this bottom line changes if we allow the *regime start-time* and the *number of virtual successes* to vary within reasonable bounds, still using the trial definition 'calendar year'. My conclusion is that these two inputs don't affect the bottom line much if your *first-trial probability* is below 1/100. They matter even less if your *first-trial probability* is below 1/300. The core reason for this is that *if your first-trial probability is lower, you update less from observed failures*. Both the *regime start-time* and the *number of virtual successes* affect the size of the update from observed failures; if this update is very small to begin with (due to a low *first-trial probability*), then these inputs make little difference.

Overall, this section slightly widens my preferred range to [1%, 10%]. If this seems reasonable, I suggest skipping to section 6.

The section has three parts:

- I briefly explain with an example why having a lower *first-trial probability* means that you update less from observed failures (here).
- I investigate how the *number of virtual successes* affects the bottom line (here).
- I investigate how the *regime start-time* affects the bottom line (here).

5.1 The lower the *first-trial probability*, the smaller the update from observing failure

To illustrate this core idea, let's consider a simple example:

You've just landed in foreign land that you know little about and are wondering about the probability p that it rains each day in your new location. You've been there 10 days and it hasn't rained yet.

Let's assume each day is a trial and use 1 *virtual success*. Ten failed trials have happened. We'll compare the size of the update from these failures for different possible *first-trial probabilities*.

If your *first-trial probability* was ½, then your posterior probability that it rains each day is $E(p) = 1/(2 + 10) = 1/12$ (see formula). You update $E(p)$ from $1/2$ to $1/12$.

But if your *first-trial* is 1/50 -- you initially believed it was very unlikely to rain on a given day -- then your posterior is $E(p) = 1/(50 + 10) = 1/60$. You update $E(p)$ from 1/50 to 1/60. This is a smaller change in your belief about the probability that it rains $E(p)$, both in absolute and percentage terms.[47]

A similar principle is important for this section. If you have a sufficiently low *first-trial probability* that AGI will be developed, then the update from failure to develop it so far will make only a small difference to your probability that AGI is developed in future years. Changing the *number of virtual successes* and the *regime start-time* changes the exact size of this update; but if the update is small then this makes little difference to the bottom line.

5.2 Number of virtual successes

In this section I:

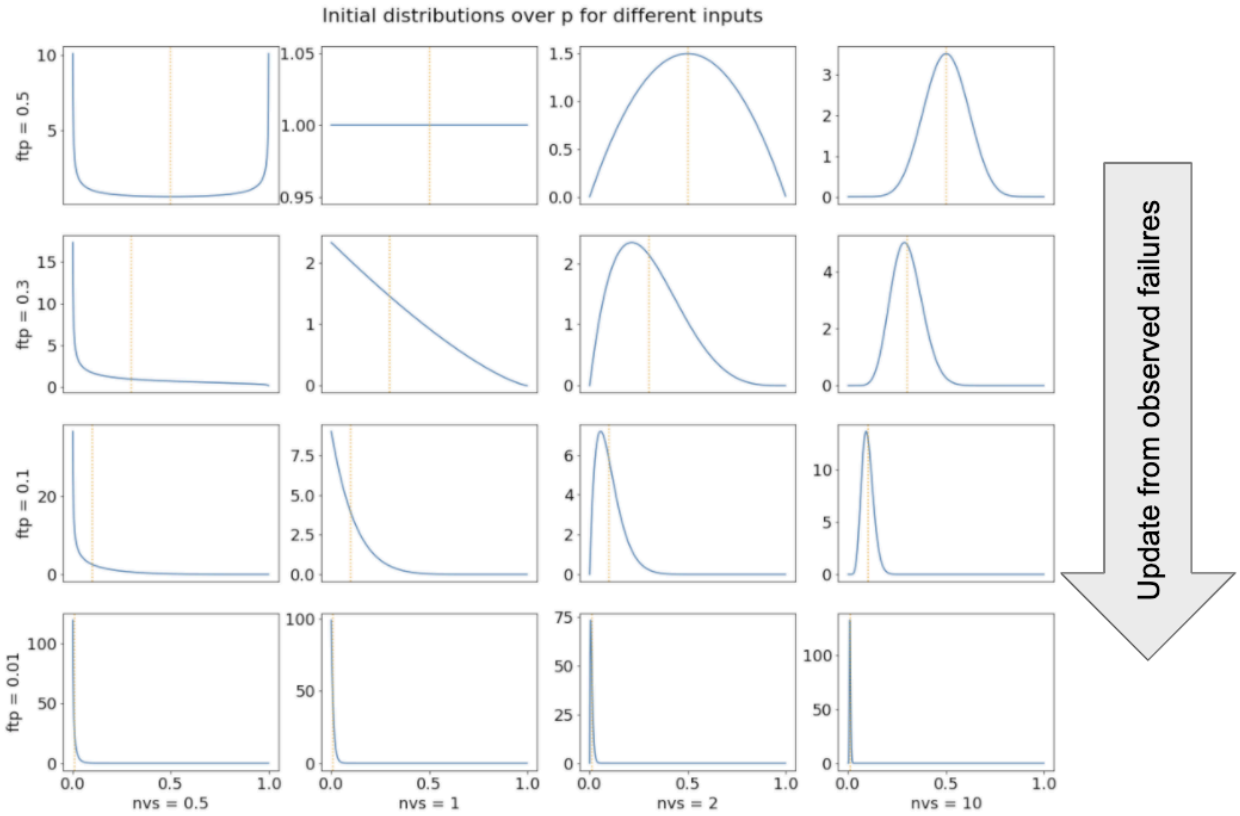
- Discuss the meaning of the *number of virtual successes* (here).
- Explain what I range I prefer for this parameter (here).
- Analyse the effect of varying this parameter on the bottom line (here).

5.2.1 What is the significance of the number of virtual successes?

Recall that, in this model, *p* is the constant probability of developing AGI in each trial. Intuitively, *p* represents the difficulty of developing AGI. We are unsure about the true value of *p* so place a probability distribution -- a beta distribution, in fact -- over its value. E(*p*) is our expected value of *p*, our overall belief about how likely AGI is to be developed in one trial, given the outcomes (failures) in any previous trials.

The *number of virtual successes*, together with the *first-trial probability*, determines your initial probability distribution over *p*. The following graphs show this initial distribution for different values of these two inputs, which I shorten to *nvs* and *ftp* on the graph labels.

[47] We can also consider the change in your belief about the median number of trials until it first rains. When the *first-trial probability* is 1/2, this shifts from 1 day to 11 days; when the *first-trial probability* is 1/50, this shifts from 49 days to 59 days. So the absolute change in the median is constant between the two examples, but the percentage change is smaller when your *first-trial probability* is smaller.



The vertical orange dotted lines shows the value of the *first-trial probability*. More *virtual successes* makes the distribution spike more sharply near the *first-trial probability;* this represents increased confidence about how difficult AGI is to develop. Conversely, fewer *virtual successes* spreads out probability mass towards extremal values of *p*; this represents more uncertainty about the difficulty of developing AGI. In other words, the number of *virtual successes* relates to the *variance* of our initial estimate of *p*. More *virtual successes* -> less variance.

We can relate this to the reference classes discussed in section 4. If there is a strong link between AGI and one particular reference class, and items in that reference class are similarly difficult to one other, this suggests we can be confident about how difficult AGI will be. This would point towards using more *virtual successes*. Conversely, if there are possible links to multiple reference classes, these reference classes differ from each other in their average difficulty, and the items within each reference class vary in their difficulty, this suggests we should be uncertain about how difficult AGI will be. This would point towards using fewer *virtual successes*.[48]

As discussed in section 3, fewer *virtual successes* means that E(*p*) changes more when you observe failed trials (holding *ftp* fixed). So we can think of virtual successes as representing the degree of *resiliency* of our belief about *p*. An alternative measure of resiliency would be the total

[48] Thanks to Alan Hajek and Jeremy Strasser for making the link with reference classes.

number of virtual observations: virtual successes + virtual failures. I explain why I don't use this measure in an appendix.

We can also use the above graphs to visualize what happens to our distribution over *p* when we observe failed trials. The distribution changes just as if we had decreased the *first-trial probability*.[49] If our initial distribution is one of the top graphs then as we observe failures it will morph into the distributions shown directly below it.[50]

5.2.2 What is a reasonable range for the *number of virtual successes*?

This section briefly discusses a few ways to inform your choice of this parameter.

I favour values for this parameter in the range [0.5, 1], and think there are good reasons to avoid values as high as 10 or as low as 0.1.

Eyeballing the graphs

One way to inform your choice of *number of virtual successes* is to eyeball the above collection of graphs, and favour the distributions that look more reasonable to you. For example, I prefer the probability density to increase as *p* approaches 0 -- e.g. I think *p* is more likely to be between 0 and 1 / 10,000 than between 1 / 10,000 and 2 / 10,000. This implies that the *number of virtual successes* ≤ 1.[51]

Such considerations aren't very persuasive to me, but I give them some weight.

Consider what a reasonable update would be

Suppose your *first-trial probability* for AGI is 1/100. That means that initially you think a year of research has a 1/100 chance of successfully developing AGI: E(*p*) = 1/100. Suppose you then learn that 50 years of research have failed to produce AGI. Later, you learn that a further 50 years have again failed. The following table shows your posterior value of E(*p*) after these updates.[52]

Number of virtual	0.1	0.5	1	2	10

[49] I prove this in this appendix.
[50] If you have fewer *virtual successes* then it will take fewer observed failures for your distribution to morph into the one directly below it. For example, the top left distribution (0.5 *virtual successes*) will morph into the bottom left after 50 observed failures, while the top right distribution (10 *virtual successes*) would take 1000 failures to morph into the bottom right.
[51] With 1 *virtual success*, the probability density tends upwards towards a constant as *p* tends 0; with <1 *virtual successes*, the probability density tends to infinity as *p* tends to 0.
[52] The table uses the equation E(p) = 1/(N + n/v), where *first-trial probability* = 1/N, n is the number of observed failures, *v* is the number of virtual successes.

successes					
Initial E(*p*)	1/100	1/100	1/100	1/100	1/100
E(*p*) after 50 failures	1/600	1/200	1/150	1/125	1/105
E(*p*) after 100 failures	1/1100	1/300	1/200	1/150	1/110

I recommend choosing your preferred *number of virtual successes* by considering which update you find the most reasonable. I explain my thinking about this below.

Intuitively, I find the update much too large with 0.1 *virtual successes*. If you initially thought the annual chance of developing AGI was 1/100, 50 years of failure is not *that* surprising and it should not reduce your estimate down as low as 1/600.[53] Such a large update might be reasonable if we initially knew that AGI would either be very easy to develop, or it would be very hard. But, at least given the evidence this project is taking into account, we don't know this.

Similarly, I intuitively find the update with 10 *virtual successes* much too small. If you initially thought the annual chance of developing AGI was 1/100, then 100 years of failure is somewhat surprising (~37%) and should reduce your estimate down further than just to 1/110.[54] Such a small update might be reasonable if we initially had reason to be very confident about exactly how hard AGI would be to develop (e.g. because we had lots of very similar examples to inform our view). But this doesn't seem to be the case.

I personally find the updates most reasonable when the *number of successes* is 1, followed by those for 0.5. This and the previous section explains my preference for the range [0.5, 1]. I expect readers to differ somewhat, but would be surprised if people preferred values far outside the range [0.5, 2].

[53] The reason why this happens is because, when the *number of virtual successes* is that lower, your initial distribution over *p* is more concentrated at values of *p* close to 0 and 1. The observed failures then very significantly reduce your probability that *p* is indeed close to 1, only leaving the possibility that *p* is very close to 0. Intuitively, this corresponds to you initially thinking that AGI is either easy or very hard and then updating to thinking it must be very hard.

[54] The reason why this happens is because, when the *number of virtual successes* is that lower, your initial distribution over *p* is more concentrated at values of *p* close to 0 and 1. The observed failures then very significantly reduce your probability that *p* is indeed close to 1, only leaving the possibility that *p* is very close to 0. Intuitively, this corresponds to you initially thinking that AGI is either easy or very hard and then updating to thinking it must be very hard.

A pragmatic reason to prefer *number of virtual successes* = 1

The mathematical interpretation of the *first-trial probability* is easier to think about if there is 1 *virtual success*.

In this case, if the *first-trial probability* = 1 / N then it turns out that there's a 50% chance of success within the first N - 1 trials. This makes it easy to translate claims about the *first-trial probability* into claims about the median expected time until success. This isn't true for other numbers of *virtual successes*.

This consideration could potentially be a tiebreaker.

5.2.3 How does varying the *number of virtual successes* affect the bottom line?

The following table shows pr(AGI by 2036) for different *numbers of virtual successes* and *first-trial probabilities*. I use a *regime start-time* of 1956.

Number of virtual successes	*first-trial probability*		
	1/100	*1/300*	*1/1000*
0.1	2.0%	1.6%	0.93%
0.25	4.1%	2.7%	1.2%
0.5	**6.4%**	**3.6%**	**1.4%**
1	**8.9%**	**4.2%**	**1.5%**
2	11%	4.7%	1.5%
4	13%	4.9%	1.6%
10	14%	5.1%	1.6%

There are a few things worth noting:

- Fewer *virtual successes* means a lower pr(AGI by 2036) as you update more from failures to date.
- Varying the *number of virtual successes* within my preferred range [0.5, 1] makes little difference to the bottom line.
- Varying the *number of virtual successes* makes less difference when the *first-trial probability* is lower.[55]
- Using very large values for the *number of virtual successes* won't affect your bottom line much, but using very small values will.[56] For example, the increase from 4 to 10 has very little effect, while the decrease from 0.25 to 0.1 has a moderate effect.

[55] The reason for this stems from the fact that when your *first-trial probability* is low, your update from observed failures is low. Halving / doubling the *number of virtual successes* can double / halve the size of these updates. But if these updates are sufficiently small to begin with, then doubling or halving their size makes little difference to the bottom line.

[56] The reason for this is that as the *number of virtual successes* tends to infinity, the update from the observed failures so far tends to 0. Once the update is already small, further increases in the *number of virtual successes* can have very little effect. By contrast, as the *number of virtual successes* tends to 0, the update from the observed failures grows without limit and the posterior E(*p*) tends to 0. So decreases in the *number of virtual successes* can continue to halve your bottom line without limit.

In fact, the above table may *overestimate* the importance of the *number of virtual successes*. This is because using fewer *virtual successes* may lead you to favour a larger *first-trial probability*, and these effects partially cancel out.

In particular, when choosing the *first-trial probability* one useful tool is to constrain or estimate the cumulative probability of success within some period. A smaller *number of virtual successes* will lower this cumulative probability, so you will need a larger *first-trial probability* in order to satisfy any given constraint.

	0.5 virtual successes				*1 virtual success*		
first-trial probability	*1/50*	*1/100*	*1/300*	*1/1000*	*1/100*	*1/300*	*1/1000*
pr(AGI in first 50 years)	43%	30%	13%	4.7%	34%	14%	4.8%
pr(AGI in first 100 years)	**56%**	**43%**	23%	8.7%	**50%**	25%	9.1%
pr(AGI by 2036 | no AGI by 2020)	**8.0%**	**6.4%**	3.6%	1.4%	**8.9%**	4.2%	1.5%

For example, suppose you constrain the probability of success in the first 100 years of research to be roughly 50%. If you use 1 *virtual success*, your *first-trial probability* will be close to 1/100; but if you use 0.5 *virtual successes*, your *first-trial probability* will be closer to 1/50. As a consequence, using 0.5 *virtual successes* rather than 1 only decreases pr(AGI by 2036) by about 8.9% - 8.0% = 0.9%, rather than the 8.9% - 6.4% = 2.5% that it would be if you kept the *first-trial probability* constant.

(Using a table like this is in fact another way to inform your preferred *number of virtual successes.* Keeping the reference classes discussed in section 4 in mind, you can decide which combination of inputs give the most plausible values for pr(AGI in first 50 years) and pr(AGI in first 100 years).)

Summary - How does varying the number of virtual successes affect the bottom line?
I prefer a range for the *number of virtual successes* of [0.5, 1]. If the *first-trial probability* ≤ 1/300, changes with this range make <1% difference to the bottom line; if the *first-trial probability* is as high as *1/100*, changes in this range make <2% difference to the bottom line.[57] Throughout the rest of the document, I use 1 *virtual success* unless I specify otherwise.

[57] If your *first-trial probability* is 1/50, changing the *number of virtual successes* from 1 to 0.5 can affect the bottom line by about 3%. For example, suppose that you choose the *first-trial probability* so that the probability of success in the first 50 years is 50%. Then with 1 *virtual success* you'll choose *first-trial probability* = 1/50 and have pr(AGI by 2036) = 12%. While with 0.5 *virtual successes* you'll choose *first-trial probability* = 1/30 and have pr(AGI by 2036) = 8.9%.

5.3 Regime start time

The *regime start time* is a time such that the failure to develop AGI before that time tells us very little about the probability of success after that time. Its significance in the semi-informative priors framework is that we update our belief about *p* -- the difficulty of developing AGI -- based on failed trials after the *regime start time* but not before it.

A natural choice of *regime start time* is 1956, the year when the field of AI R&D is commonly taken to have begun. However, there are other possible choices:

- 1945, the date of the first digital computer.
- 2000, roughly the time when an amount of computational power that's comparable with the brain first became affordable.[58]
- 1650, roughly the time when classical philosophers started trying to represent rational thought as a symbolic system.

What about even earlier *regime start times*? Someone could argue:

> '*Humans have been trying to automate parts of their work since society began. AGI would allow all human work to be automated. So people have always been trying to do the same thing AI R&D is trying to do. A better start-time would be 5000 BC.'*

The following table shows the bottom line for various values of the *first-trial probability* and the *regime start-time*.

Pr(AGI by 2036) for different inputs					
	Regime start-time				
first-trial probability	*2000*	*1956*	*1945*	*1650*	*5000 BC*
1/50	19%	12%	11%	3.7%	0.23%
1/100	12%	8.9%	8.4%	3.3%	0.22%
1/300	4.8%	4.2%	4.1%	2.3%	0.22%
1/1000	1.5%	1.5%	1.5%	1.2%	0.20%

[58] I operationalise this as the first year when you could first buy machines that can do 1e15 FLOP/s for $1 billion. My colleague Joe Carlsmith has written a report on how much computational power might be needed to match the human brain (see blog); his median estimate is that it would take 1e15 FLOP/s to match the brain's cognitive task performance. This table from AI impacts suggests that the first time 1e15 FLOP/s cost $1 billion was around 2000.

A few things are worth noting:

- If your *first-trial probability* is lower, changes in the *regime start time* make less difference to the bottom line.
- The highest values of pr(AGI by 2036) correspond to large *first-trial probabilities* and late *regime start-times.*
- Very early *regime start-times* drive very low pr(AGI by 2036) no matter what your *first-trial probability*.

However this last conclusion is misleading. The above analysis ignores the fact that the world is changing much more quickly now than in ancient times. In particular, technological progress is much faster.[59] As a result, even if we take very early *regime start-times* seriously, we should judge that the annual probability of creating AGI is higher now than in ancient times. But our above analysis implicitly assumes that the annual probability *p* of success was the same in modern times as in ancient times. As a consequence, its update from the failure to build AGI in ancient times was too strong.

In response to this problem we should down-weight the number of trials occurring each year in ancient times relative to modern times. There are a few ways to do this:

- Weight each year by the **global population** in that year. The idea here is that twice as many people should make twice as much technological progress.
- Weight each year by the amount of **economic growth** that occurs in each year, measured as the percentage increase in Gross World Product (GWP). Though GWP is hard to measure in ancient times, economic growth is a better indicator of technological progress than the population.
- Weight each year by the amount of **technological progress** in frontier countries, operationalised as the percentage increase in French GDP per capita.[60]

As we go down this list, the quantity used to weight each year becomes more relevant to our analysis but our measurement of the quantity becomes more uncertain. I will present results for all three, and encourage readers to use whichever they think is most reasonable.

Each of these approaches assigns a weight to each year. I normalise the weights for each approach by setting the average weight of 1956-2020 to 1 - this matches our previous assumption of one trial per calendar year since 1956. Then I use the weights to calculate the

[59] The rate of technological progress is faster today than in ancient times by all measures of technological progress that I'm aware of: percentage increase in TFP, percentage increase in MFP, GDP/capita, and land productivity. This is true *despite* the fact that we may already have discovered most of the low-hanging fruit technologies.

[60] I chose France on the basis of data availability and its proximity to the frontier in both modern and Roman times. In many standard economic growth models the long-run growth rate in the level of technology is the same as the long-run growth rate of GDP per capita. However, I expect this measure to underestimate the amount of technological progress in ancient times, as such progress often contributed to a larger population rather than a higher quality of life.

number of trials before 1956. The following table shows the results when the regime start-time is 5,000 BC.

	Approach to weighting each year			
	Population	***Economic growth (%)***	***Technological progress (%)***	***Zero weight before 1956***[61]
Trials between 5000 BC and 1956	168	220	139	0
first-trial probability	*Pr(AGI by 2036)*			
1/2	6.4%	5.3%	7.3%	20%
1/100	4.6%	4.0%	5.0%	8.9%
1/300	2.9%	2.7%	3.1%	4.2%
1/1000	1.3%	1.2%	1.3%	1.5%

All three approaches to weighting each year give broadly similar results. They imply that a few hundred trials occurred before 1956, rather than thousands, and so pr(AGI by 2036) is only moderately downweighted. The effect, compared with a regime start-time of 1956, is to push the bottom line down into the 1 - 7% range regardless of your *first-trial probability*.[62] So if you regard very early regime start-times as plausible, this gives you a reason to avoid the upper-end of my preferred range of 1 - 9%.[63]

[61] This is what we did in section 4: trials as calendar years and a regime start-time of 1956.

[62] This analysis used 1 *virtual success.* I did the analysis with 0.5 *virtual successes*, and the equivalent table suggested a range of 1 - 5%; this range is lower because we update more from the 'failed trials' between 5000BC and 1956.

[63] We could also keep the *regime start-time* at 1956 and then use the 2nd row of the table 'Trials between 5000 BC and 1956' to inform our choice of the *first-trial probability*. In the case of technological progress, rather than *regime start-time* = 5,000 BC and *first-trial probability* = 1/100, we could use *regime start-time* = 1956 and *first-trial* = 1/(100 + 139) = 1/239. These two approaches give identical results: our distribution over *p* in 1956 is the same on either approach (see more on why here).

The above example assumes 1 *virtual success*, but the result applies more generally: we can use a *regime start-time* of 1956 and use the above analysis to adjust our *first-trial probability*. There's a formula relating (*first-trial probability* for 1956), (*first-trial probability* for 5,000 BC), (failed trials between 5,000 BC and 1956), and *v* the *number of virtual successes*. The formula is 1/(*first-trial probability* for 1956) = *v* * (failed trials before 1956) + 1/(*first-trial probability* for 5,000 BC). The exact value of (failed trials between 5,000 BC and 1956) depends on the approach to weighting each year, as in the main text.

Summary - How does varying the *regime start-time* affect the bottom line?
Overall, the effect of very early *regime start-times* is to bring down the bottom line into the range 1 - 7% even if you have a very large *first-trial probability*. Late *regime start-times* would somewhat increase the higher end of my preferred range, potentially from 9 to 12%.

5.4 Summary - importance of other inputs

In section 4 we assumed that there was 1 *virtual success* and that the *regime start-time* was 1956. On this basis my preferred range for pr(AGI by 2036) was **1.5 - 9%**.

This basic picture changes surprisingly little when we consider different values for the *number of virtual successes* and the *regime start-time*.

- If your bottom line was towards the top of that range, then fewer *virtual successes* or an earlier *regime-start time* can push you slightly towards the bottom of that range. Conversely, a late *regime start-time* could raise your bottom line slightly.
- But if you were already near the bottom of that range, then varying these two inputs has very little effect. This is because when your *first-trial probability* is lower, you update less from the failure to develop AGI to date.

On this basis, my preferred range for pr(AGI by 2036) is now **1 - 10%**,[64] and my central estimate is still around 4%.[65]

All the analysis so far assumes that a trial is a calendar year. The next section considers other trial definitions.

[64] The increase in the upper bound is due to having some weight on *first-trial probability* = 1/100 and a late *regime start-time*.
[65] To reiterate from earlier, I'm refining my preferred range and estimate over the course of the report. At each time, I'll refer to the currently most refined estimate as my "preferred range," though it will continue to change throughout the report.