General questions

1. Please briefly summarize the key claims and arguments of the report.

The focus of the report is in trying to estimate whether Artificial General Intelligence will arise by 2036. The approach is Bayesian, so the key question is what the prior should be. The author considers a family of priors that he calls beta-geometric distributions (I've also seen them called alpha-beta priors). That is because he is viewing whether AGI arises in any particular year as a Bernoulli trial. An alpha-beta prior puts a probability on the success probability of the Bernoulli trial. The key question of the paper is estimating the appropriate value for the parameters of the alpha-beta prior; this, in turn, will determine the prior that should be used, which will then allow us to estimate the probability that AGI will arise by 2036.

The initial settings of the alpha-beta prior (what the author calls the first-trial probability) is the prior probability of success, before seeing evidence. Other parameters that need to be determined are the regime start time (since we haven't seen AGI so far, this amounts to asking how many failures there have been up to now); the author takes that to be 1956. We also need to determine how many trials there have been and what counts as a trial. The author discusses these issues and makes estimates.

2. Is the central question of the report framed in a way that makes sense to you? If not, what is unclear or confusing?

The central question of the report is completely clear.

3. What do you see as the most interesting/important contributions of the report?

This seems to be the most serious attempt to estimate when AGI will be developed that I've seen. That makes it interesting, although I wouldn't bet money on the conclusions (see below).

4. The report concludes that, given the framing of the report, the probability of AGI by 2036 is 1 - 17%, with a central estimate of around 7%. Is there some other range/central estimate that would have seemed more appropriate to you? If so, why?

If our goal were a single estimate, then this is probably as reasonable as any other. I have problems with the goal (see below).

5. Were the main considerations bearing on this that you're aware of covered in the report? If not, what were the most important considerations that were missing?

While I can't think of any obvious thing that should have been covered, I'm willing to bet that there are some things that should have been covered that I can't think of (what Donald Rumsfeld called the "unknown unknowns". I strongly suspect that the more the author thought about it, the more possible factors he generated (e.g., the time period being a calendar year, the time period depending on the number of research-years, taking computing power into account, and so on). My guess is that if we put together 10 smart people all trying to think of relevant factors, they would come up with more. Things get much more complicated when we time into account. The author does consider a more dynamic model in Section 7.1.5, but this makes the approach far more complicated. For example, if AGI will require a sequence of 3 breakthroughs, then after each breakthrough, the parameters will have to be recalculated. Once we take time into account, there are other things that need to be considered. For example, suppose that you might get a government that's anti-AI research. This will clearly affect the success probability of a Bernoulli trial. How will this be taken into account. Or what about if one country seems to have made a major breakthrough so country pre-emptively attacks it, to prevent it from getting AGI first. All this will lead the success probability to change. Even if the model makes it possible to take such factors into account, it is easy to generate scenarios and less easy to be sure that one has covered all the possibilities. This makes it important to model "unknown unknowns" appropriately.

6. Do you find the report's way of addressing these considerations clear and convincing? If not, what is unclear/unconvincing?

As I said above, I have serious concerns about the way that dynamic issues are being handled. Although the author is certainly aware of the broad issue and discusses it, I don't think it's getting the attention that it deserves.

7. Are there any major issues that you think should be addressed before we publish this report?

I am not comfortable with modeling uncertainty in this case using a single probability measure. We want a way of measuring confidence in uncertainty so that we can distinguish, for example, a coin that we know to be fair from one about whose bias we're completely uncertainty. If we had to represent the latter situation by a single probability measure, the obvious one to choose is the one that makes the success probability (i.e., the probability of getting heads) ½, by symmetry. The author deals with this issue to some extent by considering a set of semi-informative priors in Section 7 and putting a hyperprior on them. But this leads me to my major concern: I don't think that the question of the probability of AGI by 2036 is all that interesting in and of itself. What we want to do is to use the estimate of likelihood (not necessary a probability) to make decisions. To me the big gap in the paper is not considering the question of likelihood in the context of decision theory. If you represent your likelihood in terms of a single probability measure, then the obvious thing to do is to maximize expected utility. But if you have a set of probability measures as in

Section 7, what the "right" decision rule should be becomes much less clear. The situation gets even worse with a hyperprior on these measures. The obvious thing to do (which is essentially what the author does) is to use the hyperprior to convert the set of probability measures into a single probability measure, but this means that we've lost the information about (lack of) confidence. I think that there are better decision rules out there that take the lack of confidence into account. Perhaps not surprisingly, my preferred rule is minimizing weighted regret (J. Y. Halpern and S. Leung, Weighted sets of probabilities and minimax weighted expected regret: a new approach for representing uncertainty and making decisions, *Theory and Decision* **79**:3, 2015, pp. 415-450). We have a set of probability distributions, each with a weight (which can be thought of as a hyperprior, although that's not how we think of it). The weights are updated with evidence, just as the hyperprior is. The difference is how the weights are used in making decisions; the decision rule involves minimizing weighted regret, not maximizing expected utility. As the decision maker becomes more and more certain of the true probability (if the evidence in facts leads the decision maker to become more certain, which may not be the case, so the hyperprior converges to putting probability 1 on some hypothesis), this rule converges to maximizing expected utility, But while there is uncertainty about what the true probability is (which is certainly the case when we are trying to determine when AGI will occur), the decision rule takes that into account (while maximizing expected utility does not).

Do you have any suggestions for improving the presentation of the report, especially the executive summary and the introduction?

As I said, I think that the biggest improvement would be to take the role of decision making (and the question of the right approach to decision making) much more seriously. I'll use the rest of my answer to this question to make a few other random comments on the report:

- It seems to me that there's been quite a bit of work on determining the likelihood of disasters occurring which seems to me similar in spirit to the problem of determining whether AGI will occur by 2036. One obvious example is Doomsday clock? This is surely analogous. Philosophers have also considered questions like "How likely is the universe (or my part of it) to die out in the next N years?" I believe that economists have also considered the likelihood of breakthroughs. See, for example, https://arxiv.org/pdf/1807.07285.pdf, which uses other families of priors. There's certainly a lot of work on identifying breakthrough; see, e.g., https://conference.druid.dk/acc_papers/0ig10k89vf41dhpfsiabslf3a11m.pdf**.** While certainly not identical, this work seems similar in spirit, and should be discussed and contrasted with what is being down here.

- In the discussion of notable mathematical conjectures, the author observes that resolving a mathematical conjecture is unlikely to be either necessary or sufficient for AGI. While this is true, the probability of a breakthrough of some other form may well be highly relevant. I take the data for this reference class to be relevant to the question of the likelihood of a breakthrough. Of course, that probability depends very much on what you consider to be a breakthrough. (These concerns are part of why I have very little confidence in any of these

numbers, and think that it's important to consider sets of probability measures, rather than a single one.)

- In Section 6.1 (bottom of 2 in Section II) it seems strange to talk about the number of AI researchers growing at a constant exponential rate. I'm not sure what "equilibrium" means in this context (in equilibrium, things tend to be constant, so at best this is some kind of dynamic equilibrium), but if this assumption were true, then in "equilibrium" we'd have more AI researchers than people. Edited for clarity.

- At the end of Section 10.1, it says "As long as this collection of distributions is adequate, it does not matter if they were derived from an unrealistic assumption." What does "adequate" mean here? How can we tell if a set of distributions is adequate? Edited for clarity. What I meant was that the results wouldn't change by much (more than a factor of 2) if I used a different collection of distributions, choosing its parameters in a similar way to how I chose the inputs to the report's framework.

Comments on particular claims from the report

Here are some of the key claims/decisions of the report presented in a way that might highlight any important disagreements. Please let us know if you have any major disagreements with the below, or find any of the claims unclear.

1. The report's conclusion identifies the most important inputs as the *first-trial prior*, the *trial definition*, *late regime start-times* and *empirical forecasts*. The following table summarizes the effect of these inputs on pr(AGI by 2036):

Trial definition	*Low ftp, conservative forecasts*	*Central ftp, central forecasts*	*High ftp, aggressive forecasts*	*High ftp, aggressive forecasts Late start-time: 2000*
Calendar year	1.5%	4%	9%	12%
Researcher-year	2%	8%	15%	25%
Compute[1]	2%	15%	22%	28%

As I indicated above, I have a problem with using a single probability to represent uncertainty. I would prefer a set of priors, with a confidence associated with each one, and a way of making clear how you're dealing with unknown unknowns.

2. The report suggests that the *number of virtual successes*, *very early regime start-times*, and *modelling AGI as conjunctive* are less important to the bottom line.

[1] For this trial definition the report modelled late start-times by using the AGI-hard and log-uniform models, rather than by holding the other inputs fixed and changing the *regime start-time*.

While the conjunctive part is perhaps not so important, dealing with the sequential aspect of breakthroughs and how the parameters change over time is (in my opinion).

3. The report uses a beta-binomial probability distribution (and variations on this distribution for the conjunctive model and the log-uniform model). Do you think we should have considered other distributions? (Appendix 12 discusses this question.)

Again, to me, this is the wrong question. I think that sets of distributions should have been considered (which might involve distributions other than the beta-binomial), with a weight on each one. Particular attention needs to be paid to the decision rule used and to modeling unknown unknowns.

4. The report suggests that an all-things-considered judgement should be a weighted sum of update rules, where the weights are updated in response to the failure to develop AGI to date.

See above. While I'm all in favor of using weighted sets of probabilities to represent uncertainty, I would use a different approach to making decisions than that implicitly suggested here. I note that although the author talks about a central tendency and talks about ranges, he does not say anything about his degree of confidence that these ranges are correct. Would he really be willing to bet large sums of money on AGI occurring within the range of years suggested by his approach?

Response from the author of the report

Thanks for your interesting feedback. I appreciated you pointing out that the report makes a non-trivial decision to combine together various probability measures into an all-things-considered probability measure. While doing this is fine if we intend to maximise expected utility, there are other decision rules for which it is not appropriate. I also like the point that the report likely leaves out 'unknown unknown' considerations; its results should be interpreted as just one perspective on when AGI will be developed rather than a comprehensive analysis.

Expected utility vs. weighted expected regret minimization

Your main suggestion concerns the fact we combine together the different probability measures. This is only appropriate if we plan to maximise expected utility (EU); but you have argued in favor of minimizing weighted expected regret (MWER). Why do we not not consider alternative decision rules like MWER in the report?

The first answer is that I wanted to start with the most default, simple version of the analysis. I often find it's easier and more persuasive to make the simplest points first.

In order to compare the effects of different decision rules, the report would have to be much more ambitious. It would have to consider the different possible actions we could take relating to AGI, and what the utility of these actions would be conditional upon AGI arriving at certain times. Only if we extended the model in this way, could we compare the results of different decision rules like MWER and EU.

However, I think it would be exciting for further work to address alternative approaches and action guidance more thoroughly.

What might MWER imply here?

For now, it may be interesting to think through a toy example of how these decision rules may differ. Suppose we're deciding whether or not to invest in AI safety research. If AGI is far away, it is better not to invest. But if AGI is near, it is much better to invest. We can represent this situation with a pay-off matrix:

	AGI is near	AGI is far away
Invest in AI safety	3	0
Don't invest	1	1

Suppose that we have equal weight on five different probability measures. Four of these measures think AGI is far away; they judge the regret from investing to AI safety to be 1. One of the measures thinks AGI is near; it judges the regret from *not* investing to be 3-1 = 2.

In this situation, MWER and EU give different results. Intuitively, EU says "Don't invest! Even though you might really regret it, you'll probably be fine. It's worth taking the risk."

But MWER says "Invest! If you don't and it turns out that that fifth probability measure is correct, you'll really regret it. You should minimise your regret in this worst-case scenario." In essence, MWER is cautious and wants to avoid a worst-case scenario.

In some more technical detail:

EU calculates that the expected utility of "don't invest" is 1, whereas the expected utility from investing is only ⅗. (We divide by 5 because there are 5 measures of equal weight, only one of which assigns a pay-off of 3 to investing.)

MWER is more cautious. It wants to minimize the regret of the worst-case probability measure.

- If we **invest**, the worst-case probability measure says that the regret is **1**

    - In fact, four of the five measures say this. The fifth measure says the regret is 0.
- If we **don't invest**, the worst-case probability measure says the regret is **2**
    - Four of the five measures say the regret is 0, the fifth measure says the regret is 2.
- So investing minimizes the regret of the worst-case probability measure.

This toy example suggests a general lesson: Suppose there's a reasonable perspective (i.e. a probability measure with similar weight to our other probability measures) according to which preparing to AGI is very important, and other reasonable perspectives (i.e. other probability measures with similar weight) think the stakes are less extreme. In cases like this, MWER may be more inclined than EU to recommend preparing for AGI because it cautiously guards against the worst-case scenario in which we don't prepare but should have.

In any case, extending this project to study the effect of different decision rules like MWER would be an interesting and instructive exercise.