Takeoff speed report, reviewer questions

Questions

Which sections of the report did you read?

Read in depth: Abstract, Short summary, Long summary, “Evidence about the size of the FLOP gap”, “Trading off runtime compute and training compute”, LessWrong post “A concise mathematical description of the FTM”, and the playground.

Skimmed: the entirety of the rest of the report.

Can you please briefly summarise the report’s central approach to estimating takeoff speed

Design and implement a macroeconomics-style computational model of GWP, ML software & hardware R&D, and ML training, taking into account positive feedback loops from AI automating human cognitive labor. Estimate parameters of the model, most importantly “FLOP gap” (ratio of ML training FLOP spanned during “takeoff”: going from being able to automate a small fraction of economically valuable cognitive labor to 100% of it) and “AGI training requirements” (mostly from Cotra’s bio-anchors), and report what the model predicts in terms of takeoff duration and AGI arrival date.

General impressions and comments.

- What are your main thoughts about the report?
- What are its core contributions? What are its big limitations?
- How convincing is it overall?

I found the report very thought-provoking and something close to the best that can be done with this mode of analysis, modulo some of the specific critiques I list below, especially under Missing considerations and critiques.

In general, I think these sorts of forecasts of unprecedented scientific, technological, and economic dynamics are inherently difficult and prone to dramatic misses: there are lots of places where specific assumptions are made about the model’s functional form, and those assumptions could just be completely off. The basic project feels like an effort in 1750 to forecast the trajectory and effects of the industrial revolution, without knowing anything about electricity, aircraft, computing, or nuclear weapons. To be clear, I think such a project has a lot of value, in the sense that it gives us a baseline to discuss and critique, and over time compare against reality. I’m especially interested in how the next 10 years of AI development compares to the

predictions of the FTM. I personally expect that reality will deviate substantially, although it's harder to say what specifically the reasons will be.

As someone who follows the ML literature very closely, my subjective sense is I wouldn't be shocked if some new "chain of thought" trick, prompting strategy, language model cascade, or architectural advance towards combinatorial generalization just pushes us rapidly over the finish line in the next few years. In your model, I guess this would be closest to "software R&D", but of a form which is very discontinuous compared to the FTM: e.g. chain of thought just seems like a qualitative advance over pretrained LMs, like going from system 1 to system 2 thinking, and that's not something you can capture with Hernandez and Brown (2020)-style "algorithmic efficiency" measures: it's actually just optimizing for a different distribution of data when prompted in this way.

I also don't really trust Cotra-style estimates of brain vs. ANN FLOP, and it honestly wouldn't surprise me if ANNs, maybe because they have less noise, faster serial processing speed (and hence greater ability to take advantage of the training v. runtime compute tradeoff), or the use of backpropagation, can achieve what the brain does with 1% the cortical synapse count — i.e. within 1 OOM of where ML systems are now.

Nevertheless, as I said I did find the report very provocative, in that reading through and carefully considering the analysis (which overall I found to be very thoughtful and comprehensive) has led me to deepen my understanding of the relevant considerations in predicting the trajectory of AI progress over the next years/decades.

Some further comments on clarity and organization:

- In general I would have preferred if the abstract and short summary contained more detail. I found myself fairly confused until I played with the playground and read the long summary. More upfront specificity on definitions and how exactly your model works, rather than just stating generally what your goals are, would be appreciated.
- Furthermore, as I've written in comments in the margins, *emphasize from the start that the main product of this analysis is the FTM itself*, include links to the playground prominently and early, and describe your "conclusions" as the conclusions of the model given certain parameter estimates, rather than the conclusions of an "analysis", which is vaguer.
- Be clear and consistent on definitions of startpoint and takeoff, and whenever using different definitions, explicitly call it out each time (for example, in "Takeoff speed can differ in different domains", you use a definition of takeoff speed in terms of ratios between successive doubling times, which is confusing given how most of the rest of the report defines this). Don't expect the reader to keep track of this stuff. (For startpoint, could possibly even set the default to the same as wake-up time, so the reader has one less concept to keep track of in the default analysis.)
- The counter-intuitiveness of shorter FLOP gap = longer AGI timelines rears its head repeatedly. I recommend explicitly noting this wherever it may come up (e.g. in the

playground's sensitivity analysis, or any comparisons between "more aggressive" and "less aggressive" model scenarios, etc.)

- It also seems like different sections of the report were written "organically" over time, so it sort of feels like each section is working from a slightly different conceptual framework.

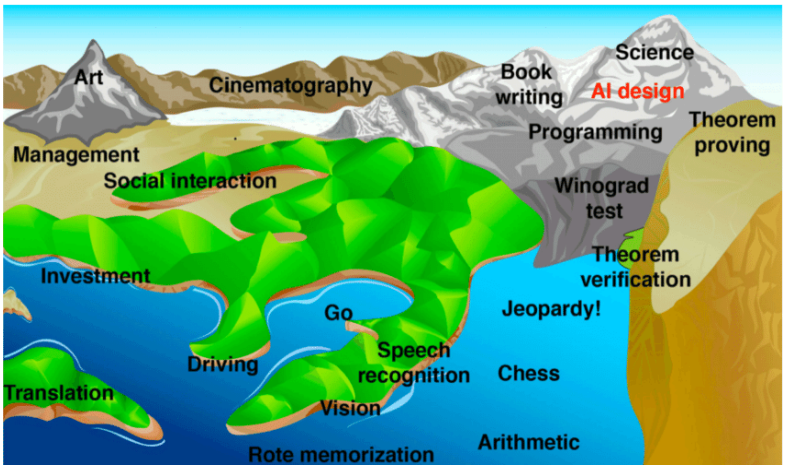
Missing considerations and critiques

FLOP Gap Evidence (severity: high)

The FLOP gap is probably the most important input to the FTM. I think there are conceptual problems/category errors with some of the offered lines of evidence for the magnitude of the FLOP gap, which should be examined carefully.

It's crucial to remember the question we're ultimately asking: what is the gap in training requirements between performing the *comparatively easiest* (for an AI) *economically useful* task and performing the *comparatively hardest* (for an AI) *economically useful* task.

In particular, it's far from obvious why it's relevant *at all* "**How AI capabilities vary with training FLOP *within one domain***". Indeed, the *very fact* that AI already have fully traversed the human range in many domains despite not even beginning to cross it in others is a perfect illustration of the inappropriateness of this sort of evidence for addressing the question at hand. On the one hand, the domain-specific range of human capabilities can *dramatically overestimate* the FLOP gap, by looking not at the range of *economically valuable* human skill (in Go, for example, only the very best Go players can make a living doing it), but inappropriately looking at the full range of human skill (which is affected by many factors not exclusively limited to innate ability, among them motivation and experience). On the other hand, looking within a single domain can dramatically *underestimate* the FLOP gap, by not accounting for the vast differences in relative AI v. human difficulty across domains. It just seems like fundamentally the wrong sort of thing to be pointing to. (Or if you're going to invoke this line of evidence, I think you need to explicitly connect the dots as to how it sheds light on the question you're asking, and personally I doubt that it does. In fact, I'd argue that the rationale for pointing to *any* evidence about the FLOP gap should be clearly laid out: don't just say "FLOP gap from 20 kyu to 9 dan in Go is N, therefore that's evidence for FLOP gap of N from X% automation to AGI", but explain what one has to do with the other.)



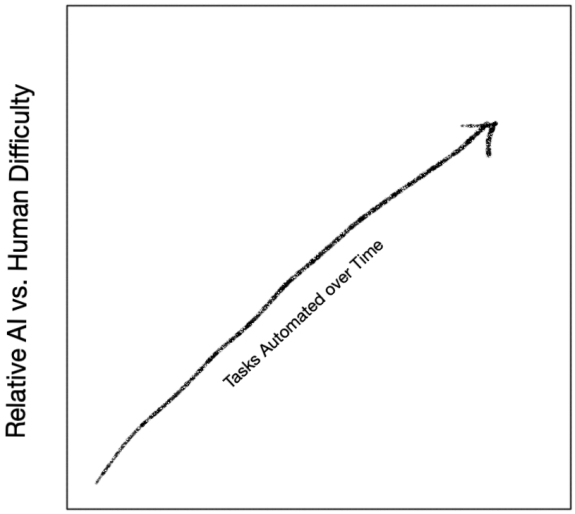
In principle, we need something like a topographical map of intellectual labor, where surface area is market size, and altitude is relative difficulty for AI compared to economically competitive humans, as measured in OOMs, and the FLOP gap is the maximum altitude minus the

minimum altitude on this map. This is essentially Hans Moravec's "**rising tide of AI capacity**" model (illustrated at left, taken from Life 3.0 p. 53), and I think an explicit reference to it, with illustration (perhaps on page 1 or 2 of your report) would be helpful for framing the question being asked.

While the report does flirt with acknowledging this issue in parts, I think taking it seriously involves removing a lot of the ostensible evidence (especially on within-domain variation), reorganizing/reweighting what remains, and clearly connecting the dots between any given offered evidence and the FLOP gap/"topographical altitude gap".

In general, I think *a lot more* weight should be put on the evidence in "**How AI capabilities vary with training FLOP *between different domains***", since it's actually the thing we care about. The lower bound from the "GPT-N task performance" subsection seems particularly compelling to me. Also, it **should be treated as a lower bound, not a point estimate**, but it seems you're currently treating it as a point estimate in your aggregation table ("effective FLOP gap ~4 OOMs"). Note also that the > 4 OOMs predicted in the "RL vs transformer training" subsection is potentially *additive* with other factors: that is, if there's the same > 4 OOM FLOP gap within deep RL tasks as within LM tasks, but deep RL tasks *also* come with an additional > 4 OOM penalty, then it seems like there could be > 8 OOMs between the easiest LM task and the hardest RL task (this matters, of course, only if all such tasks are economically relevant).

I accept that *some* weight should be put on "**How animal capabilities vary with brain size**" (although this is very ungrounded: we're basically just guessing whether chimp-like, mice-like etc. brains could perform economically useful cognitive tasks, given the right conditions) and "**Brain size - IQ correlations within humans**", because it does seem that looking at something very domain-general (and not substantially improvable with practice) like IQ is probably a reasonable way to put a (possibly very loose) *lower bound* on the FLOP gap: After all, even if ~85 IQ human beings probably can't make money playing Go, they *can* make money doing some things, and there is a continuum of increasingly more demanding intellectual labor with (loosely speaking) higher and higher minimum IQ requirements. We can thus imagine a 2D space, with the X axis representing something like the minimum viable IQ for humans to do the task, and the Y axis representing something like the relative difficulty of the task for machines compared to humans. In that case we can expect automation to proceed from the bottom-left corner to the top-right corner, which should take longer than automation just moving horizontally from left to right. This is why I say it's a (potentially extremely loose) lower bound, not a point estimate. Also, **we shouldn't take the average of multiple lower bounds as our "FLOP gap**



estimate". If the lower bounds we get from looking at animal capabilities or IQ relative to brain size are consistently lower than the lower bounds we get from looking at AI capability variations between domains, then we should roughly speaking take the larger of these lower bounds, rather than averaging between them, as you appear to be doing when you aggregate the different lines of evidence (even though in some places you acknowledge that these could be underestimates).

Three additional considerations related to the FLOP gap and the "rising tide" model:

1. Of the domains relevant to AI research itself, will those tend to be automated early or late in traversing the FLOP gap? If early, that points to the model overestimating takeoff duration; if late, the opposite. On the one hand, AI research seems like a "hard" job, requiring both deep understanding as well as creativity. But on the other hand:
    - It involves a lot of coding, a task for which ground truth training data can be generated without limit since code correctness can sometimes be verified automatically, and where we're beginning to see economic utility already with Codex, Copilot, and AlphaCode.
    - There are many ML benchmarks which make progress reasonably quantifiable, hence achievable by automated means.
    - It's very close to a "pure" cognitive labor task, where there may be fewer workflow rearrangement or robotics tech bottlenecks, and the industry is probably best-positioned and -incentivized out of just about any to tackle those bottlenecks which do emerge.
2. The "rising tide" model points to a possible rebuttal to the "chimp" argument: as we scale primate brains up, we expect them to surpass human beings roughly "all at once" (i.e. over a factor of 2 or so) since they all have a very similar cognitive architecture to our own; but AI has a different cognitive architecture and hence will likely become economically useful in numerous domains before surpassing human beings in all domains. (This is not a conclusive rebuttal: AI could theoretically *far* surpass human beings in enough domains that the reach of such intelligence makes up for the relative weaknesses in the other domains, the way human ingenuity allows us to overcome speed deficiencies relative to cheetahs, falcons, etc. But it's plausible.)
3. The "rising tide" model implies that, before all human cognitive labor is automated, there will be significant contributions to GWP (and maybe even R&D) from new types of cognitive labor that humans were unable to do at all. How hard would it be to include cognitive tasks of that nature in the CES production functions?

Training v. Runtime Compute Tradeoff (severity: high)

This is currently turned off in the model, but the model's outputs are so sensitive to this, and the case for it being an actual tradeoff that will actually be made in the real world is so strong, that I think it should be turned on with reasonable settings.

You write: “The empirical evidence typically suggests you should be able to do the tradeoff for at least a few OOMs, but no more without additional tricks. The theoretical argument seemingly supports being able to do the tradeoff ~indefinitely.”

I disagree that the two are in conflict. For reference, here’s the notebook I shared with you with a theoretical analysis of one model of this tradeoff which allows it to be done for no more than ~3 OOMs.

Multiplicative Interaction of Hardware & Software Progress (severity: high)

The FTM models effective compute as $HW \cdot SW$, but I think a more accurate model is HW^{SW}. Strongest evidence is from Droppo and Elibol (2021), where the empirical scaling laws for LSTMs and transformers have the same irreducible loss $L_\infty = 0.307$, but different exponents in the power law: -0.167 for LSTMs and -0.197 for transformers, with figure 5 showing intersection at training compute $C_0 = 0.04$ petaflop/s-days (3.5×10^{18} FLOP).

Rearranging, a transformer with training compute C_{Tran} has the same loss as an LSTM with training compute C_{LSTM} when $C_{LSTM}/C_0 = (C_{Tran}/C_0)^{1.180}$. Bottom line: the “effective FLOP multiplier” going from LSTMs to transformers (at least in this case) scales with the 1.180 power of the amount of compute being used, which means going from the current training regime of ~10^{24} FLOP to the AGI regime of ~10^{36} FLOP implies underestimating effective algorithmic efficiency by $(10^{36-24})^{0.180} \approx 150$.

That’s for *one* architectural transition: LSTMs → transformers. But it may be reasonable to expect algorithmic progress of that order every 5–10 years. If there are three such breakthroughs yet to come before AGI and we’re basing our estimates on today’s SOTA tasks, we could be underestimating the effective algorithmic improvements by a factor of several million, or much more so if we’re basing them off of 2012-era performance on ImageNet, as Hernandez and Brown (2020) does.

This HW^{SW} model predicts shorter algorithmic doubling times on more difficult tasks, since the HW base of the exponential is larger in that case. This is consistent with Hernandez and Brown (2020) table 2 and Dorner (2021) figure 3 and table 4 (although such a trend isn’t clear from table 5).

See also Tay et al (2022) figure 2 for further evidence that architectural differences affect the scaling *power* and don’t just give the equivalent of a constant compute multiplier.

Unmodeled Deployment Time Lags (severity: medium)

As I commented on "Practical difficulties with partially automating jobs: suggests a shorter FLOP gap”: How difficult would it be to model such delays explicitly in the FTM? After thinking through this, I'm concerned that this delay cannot, in some cases, be well-approximated by a shorter

FLOP gap. That is, this delay has the same *directional effect* on takeoff time as a shorter FLOP gap (i.e. reducing it), but the dynamics of takeoff may look completely different.

To take one example: consider scenarios exhibiting hyperbolic growth in software tech around the time we approach AGI (what you describe as a "software-only singularity"). Any singularity-like growth curve fundamentally relies on the delay between technological advancement and deployment approaching 0 (most basically, this is because the *speed* of technological growth has to approach ∞ at the same time the technology level itself does, but with a deployment delay bounded below at a level greater than 0, this can't happen). Do you expect deployment delays such as workflow rearrangement to continuously approach 0 as we approach AGI? This seems plausible to me, but if so, I would explicitly state and argue for that crucial assumption, and fill in the argumentative steps for why this then allows you to approximate/simulate automation delays with a smaller FLOP gap. If not, then I'd like to either see some other argument for the validity of this approximation, or that you just get rid of the approximation and model the deployment delays in the FTM.

Lack of Validation (severity: medium)

This is a complicated model with a lot of moving parts, and it would be good to try to validate it if and where possible. This is probably hard to do given the nature of the uncertainties here, and it doesn't make the model useless if it can't be done. But where possible, seems like it would be useful.

A few ideas:

- Do an alternate analysis in which you define % automation in terms of weighting by 2000 or 2010 economic value (taking into account that e.g. some tasks like machine translation might have artificially low values today due to how cheap and reliable they've become compared to then; as you put it: “when a task in the economy becomes more productive (relative to other tasks), its fraction of GDP declines”), and see if the model predicts similar dynamics going forward. This could reveal how sensitive it is to arbitrary choices of economic value weighting.
- If you run the model from 2012, how well does it predict the next 10 years of AI progress? (Admittedly, the parameters were probably fit to do so to some extent, but checking this is still better than nothing.)
- Identify the earliest (say ca. 2027) testable and surprising predictions the model makes, which could help us to have an early sense of whether we can trust it going forward.

I’m not in love with any of those ideas, but there may be other possibilities I'm not thinking of.

Application of Chinchilla-like Assumptions to Animal Brains (severity: low)

Discussing the FLOP gap in "**How animal capabilities vary with brain size**" and "**Brain size - IQ correlations within humans**", you apply Chinchilla-like assumptions that training data scales roughly linearly with brain size. There's some surprisingly good evidence that something

like this might be the case across species, but I don't think there's any evidence to apply this to within-species comparisons.

The strongest evidence I'm aware of in animals is that the duration of cognitive development seems to scale linearly with the number of cortical neurons, at least in Carnivora and primates. For example, in the timing of reaching stage 4 Piagetian object permanence, there's something close to a consistent 2 week : 1 billion cortical neuron ratio. This suggests to me something like a deliberately slowed-down plasticity/learning rate schedule, with the expectation of incorporating the lessons from more experiences over a longer period of time into the greater neuronal capacity available to some species.

Species	**Cortical Neurons**	**Stage 4 Object Permanence Time**[1]	**Weeks per billion cortical neurons**
Cat	250M	2 weeks	8
Dog	500M	2 weeks	4
Monkey	500M–3B	1–2 months	6
Non-Human Great Ape	7.5–9B	5 months	2.5
Human	16.3B	8 months	2

However, I'm not aware of any evidence like this *within* species, which should reduce the FLOP gap lower bound from your human v. human brain size comparisons by a factor of 2.

(In both the across-species and within-species cases, if you're going to make an argument like this, I think you should explain what the analog of "training data" is for biological brains in your view and why it would be larger for animals with bigger brains.)

Discuss How You Address Robotics Bottlenecks Earlier On (severity: low)

As discussed with you in the comments at the end of section 8, I think it's important to note up-front that you're effectively modeling potential robotics bottlenecks by increasing the capital share: at least as early as the "Physical capital bottlenecks" discussion in section 6, if not in the long summary.

Conclusions

- How does your view on takeoff speed differ from the conclusions of the report? What explains this difference?

[1] Gómez (2004). Apes, Monkeys, Children and the Growth of Mind pp. 67–8.

I mostly answer this in General impressions and comments. Basically, I think there's a good chance takeoff will be much faster and sooner than in this report, because the FTM can't really model the sorts of algorithmic tricks which I think might put us past the finish line this decade.

OTOH, if I'm wrong and an FTM-like model applies, then I think takeoff could be slower than the FTM predicts because I think it may be underestimating the FLOP gap, for reasons discussed here.

How should I change the sensitivity analysis?

I'd like to see one or more of the following:

- How much does takeoff speed / AGI arrival date depend on each parameter X, holding the rest of the parameters constant, in expectation over all those parameters' other values? (This is essentially asking about a causal relationship, and is different from what you have now, which is: how much does takeoff speed / AGI arrival date depend on parameter X, holding the rest *at their default values*? But if the sensitivity of takeoff speed or AGI arrival date to parameter X is not balanced with respect to the default values of other parameters, then this is misleading.)
- The same thing, but for the interaction term $X \cdot Y$ for each pair (X, Y).
- What's the correlation coefficient across all Monte Carlo simulations between takeoff speed / AGI arrival date and each parameter X? (This is not asking about a causal relationship, but something closer to mutual information. And this is different from asking what happens if you hold all other parameters at their median/default estimate and vary X, since we're allowing those parameters to covary with X.)
- The same thing, but for the interaction term $X \cdot Y$ for each pair (X, Y).

I also agree that these two could be important:

a. Model the importance of computational experiments to software R&D. Though as I wrote in a comment: "one possible counterpoint, at least for certain types of tests, is that scaling laws for architectural changes can be reasonably reliably estimated with multiple OOMs fewer FLOP than a SOTA training run would require. Indeed, it's plausible that the FLOP required to test an architecture's scaling properties might not actually increase over time, even as SOTA FLOP requirements do."

b. AI that can perform ~all software R&D tasks is much easier than AI that can perform ~all tasks

Infohazard advice

I lean towards "public as draft" or "publish" as I think this category of information is generally beneficial, but I don't feel confident about this.

Permissions

- Would you be willing for us to publish your answers to the above questions alongside the report, if we publish it?

Yes.

- Would it be OK for us to publish your name alongside your comments?

Yes.