Takeoff models often assume that some speed-ups before automated R&D aren't priced in

Here's a pattern that shows up a lot in AI timelines and takeoff models:

- The model describes some proxy metric for AI progress
- The model makes an inside-view argument for why this metric will increase much faster than it has previously, even before we get to Top-human-Expert-Dominating AI (TEDAI)
- This argument is used to argue for shorter timelines than the first-order predictions from historical data of that metric

Two examples of this pattern:

- AI 2027 uses time horizons as a metric, and uses a mix of empirical[1] and conceptual[2] arguments to argue for a superexponential increase in time horizon (instead of an exponential).
- What a Compute-Centric Framework Says About Takeoff Speeds uses multiple factors, including growth of annual hardware R&D spending, and argues that relevant actors "waking up" to the potential of AGI will increase this from 5% to 17%.

I think these models don't really justify the assumption that speed-ups like this, made from inside-view arguments, aren't already priced into the historical trend of growth. I think there are good historical reasons to expect reasonably-long-running trends to have involved several such big-if-true advances which were balanced out by other factors to result in the trend we see.

Related posts: Trust me bro, just one more RL scale up, this one will be the real scale up with the good environments, the actually legit one, trust me bro, Does Reality Drive Straight Lines On Graphs, Or Do Straight Lines On Graphs Drive Reality?

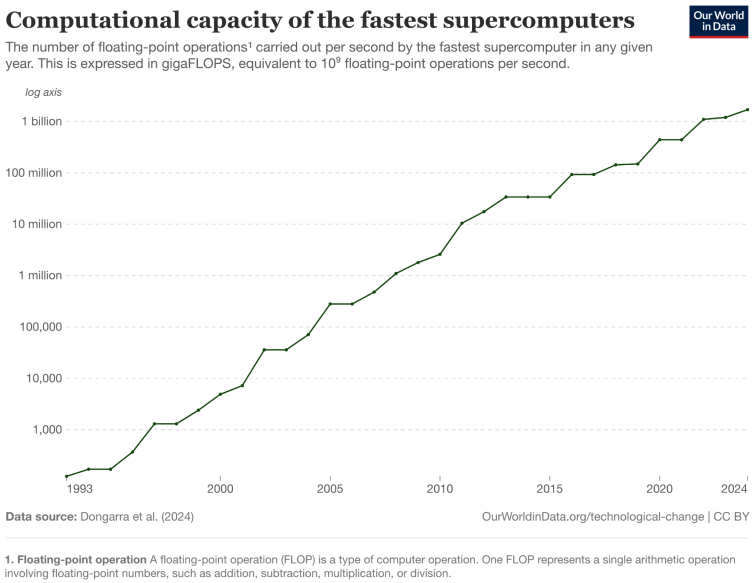
Three models in other domains

Here are three models of growth in other domains, in decreasing order of "the law of straight lines" holding:

[1] Time horizons seemingly increasing faster from 2024.
[2] The relative difficulty gaps for humans of tasks between time horizon doublings seemingly going down over time (i.e. that the gap in difficulty between 1 month and 2 month tasks is lower than between 1 day and 2 day tasks).

1. **Moore's law** has been driven by several pretty distinct inventions such as DRAM and deep UV excimer laser photolithography. In each case, there was probably a good argument that these inventions were a big deal and would cause things to grow faster than before. Conversely, there were also several points at which it seemed like the trend would break, only for some new invention to come along and continue the trend like before.

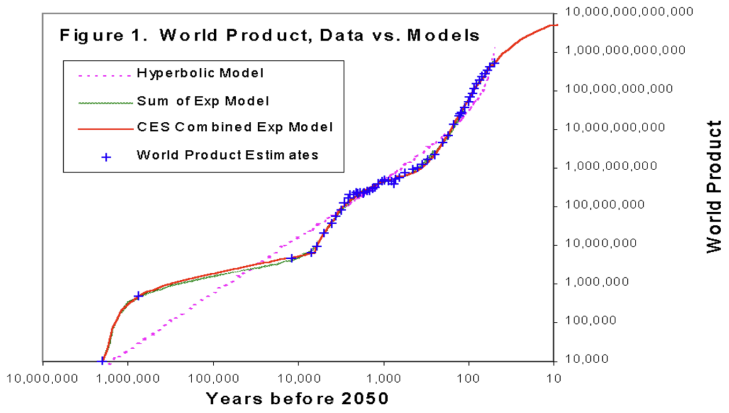
   Takeaway: Object-level arguments for speedup (and slowdown) kept happening throughout this period, and they were all priced in.



2. **Global GDP over time** is pretty interesting. Robin Hanson's model posits three separate periods of growth, each dramatically faster than the last, with a better fit for historical data. In this case specific object-level advances (hunting, farming, and industry) was in fact a step-change in progress relative to immediately before them. But the overall pattern is one of recurring discrete step-changes, so the next step-change wouldn't be discontinuous to the overall pattern.

   In a counterfactual world without AI (or strong human uplift), this model would've plausibly broken. But (a very coarse projection of) AI progress was already priced into the model. Object-level arguments about a particular advance would only have worked to distinguish "step-change growth vs normal exponential growth", and not "step-change growth vs greater-than-step-change growth".

   Takeaway: Even step-changes can be continuous with a pattern of step-changes. In such cases, object-level arguments would help you predict 'step-change vs normal exponential,' but not 'this step-change vs the trend of step-changes.'

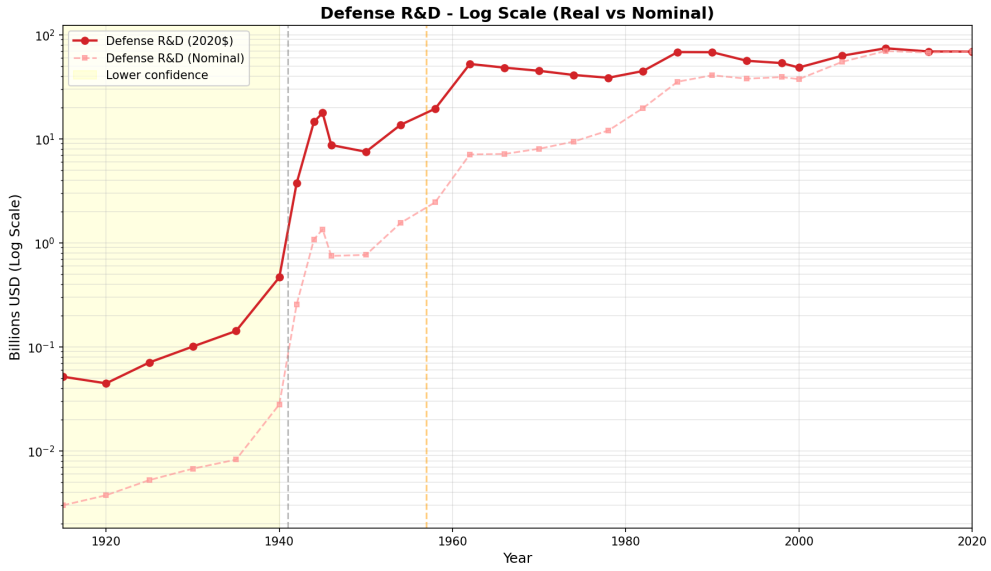


3. **Growth in US defense and space R&D**. Davidson's model lists this as a justification for why growth in annual hardware R&D spending might increase after relevant actors wake up, when there's suddenly large demand. The model argues that ~~similar~~ growth in US defense R&D during and after WW2, during a similar period of suddenly large demand, suggests that annual hardware spending might increase by 10-30% per year after "wake up".

   This plot (generated by Claude) suggests that the compounded annual growth rate of spending was actually pretty similar before and after WW2[3]; growth before WW2[4] was already 9.3%[5]. If we only focus on the growth *during* WW2 and recklessly extrapolate that to AI though, we could argue that if "wake up" involves very intense[6] and sustained competition for years[7] then you'd see a sustained speed-up, or only a one-off increase otherwise.

   Takeaway: There are some cases where object-level arguments for discontinuity are pretty predictive. In this example, you see a massive spike in growth during a very intense period, but similar-to-trend growth after despite arguments[8] for moderately increased growth.

[3] With a huge spike during WW2 and a much smaller spike during the Sputnik era.
[4] Over 1915-1940.
[5] Though the data here might be kinda sketchy.
[6] Note that the Sputnik speed-up was much smaller than the WW2 speedup (14.3% vs 117.1%), which implies either that speed-up from a larger base is much smaller or that it took the greater incentives of WW2 to meaningfully increase spending.
[7] The takeoff model says 20 years with 17% growth here. With an approximate doubling every year as in WW2, that would involve 5 years of intense competition on par with WW2 to hit a cap of 1% global GDP. This is a huge overestimate though, as we're starting from a much larger base.
[8] The model mentions a notable effort to expand R&D after WW2, and by the 60s you get the Sputnik era.



What makes a good argument for discontinuity?

Looking at the three models, it seems like events which massively accelerate (or decelerate) growth can cause discontinuous patterns, but events which moderately accelerate (or decelerate) growth are probably priced into the existing pattern. Additionally, progress always has effects in both directions cancelling each other out a bunch of the time, we should expect not-super-rare accelerating events to already have occurred before and been balanced by other factors.

So arguments for discontinuity should, on the outside view, probably describe massively accelerating and/or super-rare events. Automated AI R&D is one such event.

More specifically, consider the AI Futures Model section on why their median parameters are 3.5-5 years slower than AI 2027. These weren't necessarily predictable in advance, but in general I thought the original AI 2027 forecast didn't sufficiently justify why the super-exponential trend wouldn't run into fully balancing negative factors. With every historical advancement that could plausibly speed things up, there were very likely some such factors, which without consideration would've led to similarly short forecasts.

Arguing for why an advancement posited to be trend-breaking is of large enough effect size / otherwise strange enough to overcome such new factors would probably be quite useful for figuring out whether such advancements are actually trend-breaking or are just driving the trend.