Some open questions of AI forecasting

A quick write-up by Jaime Sevilla. Jan 2024.

This document outlines some open questions in AI forecasting. I think these are candidates for an entrepreneurial person to spend a day or two thinking about. If you plan to work on them, I'd recommend writing down a methodology and executing a quick analysis based on it, then sharing it with others for comment. I erred on the side of doing this quickly rather than spending a lot of time curating the list.

I marked down with a star ⭐ questions that some of my colleagues think are particularly interesting.

Training requirements

- How can we estimate the training requirements of AI that can substitute humans in ~all economic tasks? Can we do better than the direct approach? ⭐
- What features of tasks help predict automatability? How would one go about building an "automatability score" for different tasks? Relevant background reading.
- How can we relate the inputs to AI development to productivity increases in specific economic tasks, for example the tasks listed in O*NET?
    - In particular, how can we relate benchmark performance to economic productivity?
- What sectors and jobs are more susceptible to automation? How can we predict productivity improvements and displacement in those sectors?
- What conceptual frameworks can help us understand task creation due to AI / faster tech development? We have early work from Acemoglu (2019), but I believe better analysis is possible.
- How much transfer learning should we expect between different tasks and modalities? Eg how much can we improve performance on biology questions by training on physics questions? ⭐
- How can we measure data quality? In which ways can we quantify the quality-adjusted data input to a system? For example, how could we weigh data quality so we could formalize claims of the type "training on X tokens from wikipedia articles is roughly equivalent to training on Y tokens from Reddit"?

Drivers and bottlenecks in AI

- We currently think about the gains from innovations as equivalent to a multiplicative factor on the available compute (the Compute Equivalent Gain or CEG). This assumes that the innovations are scale-invariant, ie they save you a fixed amount of OOMs of compute regardless of the scale of the system. However, some important past innovations, such as Chinchilla, can be better thought as a change in the exponents of

scaling laws, so the gains increase as model sizes increase. How inaccurate is the scale-invariance assumption? ⭐

- What alternates to CEG for conceptualizing algorithmic progress exist? Is there a more practical concept of algorithmic efficiency for predicting future capabilities?
- What were the most important innovations in AI? How important they were, in terms for example of their CEG? How often do they happen? How lasting and transferable are they?
    - In particular, what about post-training enhancements (PTEs)? How do they relate to the release of new frontier models? How well do different PTEs stack with one another? How can we expect them to contribute to performance over time?
- What are the limits to algorithmic innovation? What is for example the least amount of compute one would need to achieve a certain level of performance on a certain benchmark? ⭐
- How important is experimentation and access to large compute budgets for developing innovations in AI? What's the ratio of experimentation to theory behind major breakthroughs?⭐
- Relatedly, it seems like the contributions of hardware and software to scaling are surprisingly close, or at least comparable in magnitude. For example, in CV they seem to have contributed 50/50, while in language modelling hardware contributed 6.5 OOMs and software 5 OOMs. How can we explain or dispel this coincidence? ⭐
- What trends can we observe in inference compute? What are the most important innovations we have seen to decrease inference costs, and by how much? ⭐
- What factors of development other than compute can help predict AI system performance? Can one build a predictive model of performance such as the one here that incorporates factors other than model scale and date and better predicts performance?
    - For example, labor input to a development project seems an important variable to account for. Would models of AI development that incorporate a labor component have significantly more predictive power? How can we measure it in a way that accounts for seniority and talent?
- How large is the pool of talent who could work on large-scale ML R&D? How concentrated has been the origin of innovation in ML in terms of individual authors?
- How much more efficient are different labs at using their resources? For example, it seems that Gemini Ultra might have been trained with significantly more compute than GPT-4, but only achieved modest performance gains. On a smaller scale, Mistral seems to have achieved significantly better performance with their MoE model than the similarly sized Llama2. Which labs are ahead of the curve, and by how much? ⭐
- What are trends of improvement in robotics relative to training compute / onboard processing power / data seen? (and how should we measure performance in the first place?) What manufacturing capacity for robots exists and could exist within a decade?
- We model the returns to RnD as a semi-endogenous process with fixed returns to investment. Is this model flawed? Would it be better to for example model returns as diminishing over time? What alternatives are there?

- There is some interest in developing AI systems that behave more like agents - striving for long-term goals with little supervision rather than following the directions of a user over a short session. How will AI agents be different from current LLMs? Will current scaling laws still give a roughly accurate picture of future capabilities if agents become the dominant paradigm? Will they be better described by new scaling laws? Or will the scaling laws paradigm have much less predictive power for describing agent performance?
- What major paradigm shifts have we seen in AI so far? What are the base rate chances of a paradigm shift happening in the next decade? What promising candidates exist?
- What hardware paradigms might succeed CMOS technology? What types of fundamental limits would these technologies face?

Societal impacts

- How will access to AI development diffuse over time? In particular, how will the costs to achieve a certain level of capabilities diminish over time? How many people will have access to certain capability levels over time?
- How will open source models compare to frontier models in AI? How many years does it take for OS to catch up to the frontier? And will this lag shrink or grow over time? ⭐
- What advantages to scale will exist for AI training and inference? How much cheaper will it be to conduct inference at scale? ⭐
- One intuition about the stakes of AI development is that the drastic speed up of the economy and scientific development will threaten the stability of existing institutions. How can we make this argument in a more rigorous way? What evidence does exist linking faster growth and development to instability? ⭐
- How can we relate existing models of AI development to risk assessments of the main risk categories being studied by governments? What historical evidence and analogies exist? ⭐
- How will widespread automation affect employment and wages? What are the key economic parameters in standard models that govern these issues? What evidence do we have about them?
- Given what we know about returns to RnD, how much should we expect AI to increase scientific productivity?
- How should we expect AIs to be better than humans in qualitatively different ways? For example, how big of a deal is for AI to be able to share weight updates across many instances? How can we adjust standard models of AI trajectories based on the accumulation of hardware and substitution of human labour to better account for these differences?
- What are some important predictions about AI for 2024? Eg which performance will be achieved in relevant benchmarks, how many resources will be invested into AI development, which news about AI will be announced?

Wrapping up

This is a short and opinionated list of some important open questions in AI forecasting. Many of them require important conceptual and methodological innovations, though some could be addressed with hard work and some straightforward data analysis.

If you have a background in ML and are interested in working on questions such as these, I invite you to apply for a job at Epoch (subscribe to our newsletter to learn about future job offers!). For people considering joining the field, I encourage you to write a short investigation on a question of your interest and share it publicly online - producing these writeups is one of the best things you can do to build a portfolio. Feel free to email me the result if you do, though be aware that I might not have time to read it carefully (sorry!).

Thank you to Tamay Besiroglu and Ari Brill for their feedback on the questions.