Hurdles of using forecasting as a tool for making sense of AI progress

Nuño Sempere, Misha Yagudin

September 22, 2023

Introduction

In recent years there have been various attempts at using forecasting to discern the shape of the future development of artificial intelligence:

- The AI progress Metaculus tournament
- The Forecasting Research Institute's existential risk forecasting tournament/experiment
- Samotsvety forecasts on the topic of AI progress and dangers
- INFER, previously CSET-Foretell, has had a number of questions on short-term technological progress

Recently, Open Philanthropy—a large foundation which has been making multi-million dollar bets into AI and AI safety and funding some of the above exercises—commissioned Arb, a consultancy, to look into how to produce informative forecasting questions. As part of that project, here is a list of reasons why using forecasting to make sense of AI developments can be tricky, as well as a suggestion of ways forward.

Excellent forecasters and Superforecasters™ have an imperfect fit for long-term questions

Here are some reasons why we might expect longer-term predictions to be more difficult:

1. No fast feedback loops for long-term questions. You can't get that many predict/check/improve cycles, because questions many years into the future, tautologically, take many years to resolve. There are shortcuts, like this pastcasting app, but they are imperfect.
2. It's possible that short-term forecasters might acquire habits and intuitions that are good for forecasting short-term events, but bad for forecasting longer term outcomes. For example, "things will change more slowly than you think" is a good heuristic to acquire for short-term predictions, but might be a bad heuristic for longer term predictions, in the

same sense that "people overestimate what they can do in a week, but underestimate what they can do in ten years". This might be particularly insidious to the extent that forecasters acquire intuitions which they can see are useful, but can't tell where they come from. In general, it seems unclear to what extent short-term forecasting skills would generalize to skill at longer-term predictions.

3. "Predict no change" in particular might do well, until it doesn't. Consider a world which has a 2% probability of seeing a worldwide pandemic, or some other large catastrophe. Then on average it will take 50 years for one to occur. But at that point, those predicting a 2% will have a poorer track record compared to those who are predicting a ~0%.
4. In general, we have been in a period of comparative technological stagnation, and forecasters might be adapted to that, in the same way that e.g., startups adapted to low interest rates.
5. Sub-sampling artifacts within good short-term forecasters are tricky. For example, my forecasting group Samotsvety is relatively bullish on transformative technological change from AI, whereas the Forecasting Research Institute's pick of forecasters for their existential risk survey was more bearish.

Forecasting loses value when decontextualized, and current forecasting seems pretty decontextualized

Forecasting seems more valuable when it is commissioned with an eye to influencing a specific decision. For instance, suppose that you were thinking of startin a new startup. Then it would be interesting to look at:

- The base rate of success for startups
- The base rate of success for all new businesses
- The base rate of success for startups that your friends and wider social circle have started
- Your personal rate of success at things in life
- The inside view: decomposing the space between now and potential success into steps and giving explicit probabilities to each step
- etc.

With this in mind, you could estimate the distribution of monetary returns to starting a startup, vs e.g., remaining an employee somewhere, and make the decision about what to do next with that estimate as an important factor.

But our impression is that AI forecasting hasn't been tied to specific decisions like that. Instead, it has tended to ask questions that might contribute to an "holistic understanding" of the field. For example, look at Metaculus' AI progress tournament. The first few questions are:

- How many Natural Language Processing e-prints will be published on arXiv over the 2021-01-14 to 2030-01-14 period?

- What percent will software and information services contribute to US GDP in Q4 of 2030?
- What will be the average top price performance (in G3D Mark /$) of the best available GPU on the following dates?

My impression is that these questions don't have the immediacy of the previous example; they aren't incredibly connected to impending decisions. You could draft questions which are more connected to impending decisions, like asking about whether specific AI safety research agendas would succeed, whether AI safety organizations that were previously funded would be funded again, or about how Open Philanthropy would evaluate its own AI safety grantmaking in the future. However, these might be worse qua forecasting questions, or at least less Metaculus-like.

Overall, my impression is that forecasting questions about AI haven't been tied to specific decisions in a way that would make them incredibly valuable. This is curious, because if we look at the recent intellectual history of forecasting, its original raison d'être was to make US intelligence reports more useful, and those reports were directly tied to decisions. But now forecasts are presented separately. In our experience, it has often been more meaningful for forecasters to look in depth at a topic, and then produce a report which contains predictions, rather than producing predictions alone. But this doesn't happen often.

The phenomena of interest are really imprecise

Misha Yagudin recalls that he knows of five different operationalizations of "human level AGI". "Existential risk" is also ambiguous: does it refer to human extinction? or to losing a large fraction of possible human potential? if so, how is "human potential" operationalized?

To deal with this problem, one can:

- Not spend much time on operationalization, and accept that different forecasters will be talking about slightly different concepts
- Try to specify concepts as precisely as possible, which involves a large amount of effort.

Neither of those options is great. Although some platforms like Manifold Markets and Polymarket are experimenting with underspecified questions, forecasting seems to work best when working with clear operationalizations. And the fact that this is expensive to do makes the topic of AI a bit of a bad fit for forecasting.

CSET had a great report trying to address this difficulty: Future Indices. By having a few somewhat overlapping questions on a topic, e.g., a few distinct operationalizations of AGI, or a few proxies that capture different aspects of a domain of interest, we can have a summary index that better captures the fuzzy concept that we are trying to reason about than any one imperfect question.

That approach does make dealing with imprecise phenomena easier. But it increases costs, and a bundle of very similar questions can sometimes be dull to forecast on. It also doesn't solve this problem completely—some concepts, like "disempowering humanity", still remain very ambiguous.

Here are some high-level examples for which operationalization might still be a concern:

- You might want to ask about whether "AI will go well". The answer depends whether you compare this against "humanity's maximum potential" or with human extinction.
- You might want to ask whether any AI startup will "have powers akin to that of a world government".
- You might want to ask about whether measures taken by AI labs are "compentent".
- You might want to ask about whether some AI system is "human level", and find that there are wildly different operationalizations available for this

Here are some lower-level but more specific examples:

- Asking about FLOPs/$ seems like a tempting abstraction at first, because then you can estimate the FLOPs if the largest experiment is willing to spend $100M, $1B, $10B, etc. However, the abstraction ends up breaking down a bit when you look at specifics.
  - Dollars are unspecified: For example, consider a group like Inflection, which raises $1B from NVIDIA and Microsoft, and pays NVIDIA and Microsoft $1B to buy the chips and build the datacenters. Then the FLOPs/ is very underdefined. OpenAI's deal with Microsoft also makes their FLOPS/$ ambiguous. If China becomes involved, their ability to restrict emmigration and the pre-eminent role of their government in the economy also makes FLOPs/$ ambiguous.
  - FLOPS are underspecified. Do you mean 64-bit precision bits? 16-bit precision? 8-bit precision? Do you count a multiply-accumulate operation as one FLOP or two FLOPs?
- Asking about what percentage of labour is automated gets tricky when, instead of automating exactly past labour, you automatize a complement. For example, instead of automatizing a restaurant as is, you design the menu and experience that is most amenable to being automated. Portable music devices don't automate concert halls, they provide a different experience. These differences matter when asking short-term resolvable questions about automation.
- You might have some notion of a "leading lab". But operationalizing this is tricky, and simply ennumerating current "leading labs" risks them being sidelined by an upstart, or that list not including important Chinese labs, etc. In our case, we operationalized "leading lab" as "a lab that has performed a training run within 2 OOM of the largest ever at the time of the training run, within the last 2 years", which leans on the inclusive

side, but requires keeping good data of what the largest training data is at each point in time, like here, which might not be available in the future.

Many questions don't resolve until it's already too late.

Some of the questions we are most interested in, like "will AI permanently disempower humanity", "will there be a catastrophe caused by an AI system that kills >5%, or >95% of the human population", or "over the long-term, will humanity manage to harness AI to bring forth a flourishing future & achieve humanity's potential?" don't resolve until it's already too late.

This adds complications, because:

- Using short-term proxies rather than long-term outcomes brings its own problems
- Question resolution after transformative AI poses incentive problems. E.g., the answer incentivized by "will we get unimaginable wealth?" is "no", because if we do get unimaginable wealth, the reward is worth less.
- You may have "prevention paradox" and fixed-point problems, where asking a probability reveals that some risk is high, after which you take measures to reduce that risk. You could have asked about the probability conditional on taking no measures, but then you can't resolve the forecasting question.
- You can chain forecasts, e.g., ask "what will [another group] predict that the probability of [some future outcome] is, in one year". But this adds layers of indirection and increases operational burdens.

Another way to frame this is that some stances about how the future of AI will go are unfalsifiable until a hypothesized treacherous turn in which humanity dies, but otherwise don't have strong enough views on short-term developments that they are willing to bet on short-term events. That seems to be the takeaway from the late 2021 MIRI conversations, which didn't result in a string of $100k bets. While this is a disappointing position to be in, not sure that forecasting can do much here beyond pointing it out.

More dataset gathering is needed

A pillar of Tetlock-style forecasting is looking at historical frequencies and extrapolating trends. For the topic of AI, it might be interesting to do some systematic data gathering, in the style of Our World In Data-type work, on measures like:

- Algorithmic improvement for [chess/image classification/weather prediction/...]: how much compute do you need for equivalent performance? what performance can you get for equivalent compute?
- Price of FLOPs
- Size of models
- Valuation of AI companies, number of AI companies through time

- Number of organizations which have trained a model within 1, 2 OOM of the largest model
- Performance on various capability benchmarks
- Very noisy proxies: Machine learning papers uploaded to arxiv, mentions in political speeches, mentions in american legislation, Google n-gram frequency, mentions in major newspaper headlines, patents, number of PhD students, number of Sino-American collaborations, etc.
- Answers to AI Impacts' survey of ML researchers through time
- Funding directed to AI safety through time

Note that datasets for some of these exist, but systematic data collection and presentation in the style of Our World In Data would greatly simplify creating forecasting pipelines about these questions, and also produce an additional tool for figuring out "what is going on" at a high level with AI. As an example, there is a difference between "Katja Grace polls ML researchers every few years", and "there are pipelines in place to make sure that that survey happens regularly, and forecasting questions are automatically created five years in advance and included in forecasting tournaments with well-known rewards". Epoch is doing some good work in this domain.

Forecasting AI hits the limits of Bayesianism in general

The worries about Tetlock-style forecasting could be answered by saying: sure, that particular brand of forecasting isn't known to work on long-term predictions. But we have good theoretical reasons to think that Bayesianism is a good model of a perfect reasoner: see for example the review of Cox's theorem in the first few chapters of Probability Theory. The Logic of Science. So the thing that we should be doing is some version of subjective Bayesianism: keeping track of evidence and expressing and sharpening our beliefs with further evidence. See here for a blogpost making this argument informally but more coherently.

But Bayesianism is a good model of a perfect reasoner with *infinite compute* and *infinite memory*, and in particular access to a bag of hypotheses which contains the true hypothesis. However, humans don't have infinite compute, and sometimes don't have the correct hypothesis in mind. Knightian uncertainty and Kuhnian revolutions, Black swans or ambiguity aversion can be understood as consequences of normally being able to get around being approximately Bayesian, but sometimes getting bitten by that approximation being bounded and limited.

So there are some situations where we can get along by being approximately Bayesian, like coin flips and blackjack tables, domains where we pull our hairs and accept that we don't have infinite compute, like maybe some turbulent and chaotic physical systems or trying to predict dreams. Then we have some domains in which our ability to predict is meaningfully improving with time, like for example weather forecasts, where we can throw supercomputers and PhD students at it, because we care.

Now the question is where AI in particular falls into that spectrum. Personally, I suspect that it is a domain in which we are likely to not have the correct hypothesis. For example, observers in general, but also the Machine Intelligence Research Institute in particular, failed to predict the rise of LLMs and to orient their efforts into making such systems safer, or into preventing such systems from coming into existence. I think this tweet, though maybe meant to be hurtful, is also informative about how tricky of a domain predicting AI progress is:

> eliezer has IMO done more to accelerate AGI than anyone else.certainly he got many of us interested in AGI, helped deepmind get funded at a time when AGI was extremely outside the overton window, was critical in the decision to start openai, etc.
>
> — Sam Altman (@sama) February 3, 2023

Now, imagine that instead of being interested in AI progress, we were interested in social science, and concerned that they couldn't arrive at the correct conclusion in cases where it was Republican-flavoured. Then, one could notice that moving from p-values to likelihood ratios and Bayesian calculations wouldn't particularly help, since Bayesianism doesn't work unless your prior assigns a sufficiently high prior probaiblity to the correct hypothesis. In this case, I think one easy mistake to make might be to just shrug and keep using p-values.

Similarly, for AI progress, one could notice that there is this subtle critique of forecasting and Bayesianism, and move to using, I don't know, scenario planning, which arguendo could be even worse, assume even more strongly that you know the shape of events to come, and not provide mechanisms for noticing that none of your hypotheses are worth much. I think that would be a mistake.

Forecasting also has a bunch of other limitations as a genre

Forecasting can be seen as a genre, in which someone writes a forecasting question, that question is deemed sufficiently robust, and then forecasters produce probabilities on it.

As a genre, it has some limitations. For instance, when curious about a topic, not all roads lead to forecasting questions, and working in a project such that you *have* to produce forecasting questions seems oddly limits.

The conventions of the forecasting genre also dictate that forecasters will spend a fairly short amount of time researching before making a prediction. Partly this is a result of, for example, the scoring rule in Metaculus, which incentivized forecasting on many questions. Partly this is that unpaid forecasting platforms determine that forecasting will be a hobby, rather than a full-time occupation, and even when they pay money, they pay comparatively little. If one thinks that some questions require one to dig deep, and that one will otherwise easily produce shitty forecasts, this might be a particularly worrying feature of the genre.

Perhaps also as a result of its unprofitability, the forecasting community has also tended to see a large amount of churn, as hobbyist forecasters rise up in their regular careers and it becomes more expensive for them in terms of income lost to forecast on online platforms. You also see this churn in terms of employees of these forecasting platforms, where maybe someone creates some new project—e.g., Replication Markets, AI Progress Tournament, Ought's Ellicit, etc.—but then that project dies as its principal person moves on to other topics.

Forecasting also makes use of scoring rules, which aim to reward forecasters such that they will be incentivized to input their true probabilities. Sadly, these often have the effect of incentivizing people to not collaborate and share information. This can be fixed by using more capital intensive scoring rules that incentivize collaboration, or by grouping forecasters into teams such that they will be incentivized to share information within a team.

As an aside, here is a casual review of the track record of long-term predictions

If we review the track record of superforecasters on longer term questions, we find that… there isn't that much evidence here—remember that the ACE program started in 2010. In *Superforecasting* (2015), Tetlock wrote:

> Taleb, Kahneman, and I agree there is no evidence that geopolitical or economic forecasters can predict anything ten years out beyond the excruciatingly obvious—"there will be conflicts"—and the odd lucky hits that are inevitable whenever lots of forecasters make lots of forecasts. These limits on predictability are the predictable results of the butterfly dynamics of nonlinear systems. In my EPJ research, the accuracy of expert predictions declined toward chance five years out. And yet, this sort of forecasting is common, even within institutions that should know better.

However, in p. 33 of Long-Range Subjective-Probability Forecasts of Slow-Motion Variables in World Politics: Exploring Limits on Expert Judgment (2023), we see that the experts predicting "slow-motion variables" 25 years into the future attain a Brier score of 0.07, which isn't terrible.

Karnofsky, the erstwhile head-honcho of Open Philanthropy, spins some research by Arb and others as saying that the track record of futurists is "fine". Here is a more thorough post by Dan Luu which concludes that:

> …people who were into "big ideas" who use a few big hammers on every prediction combined with a cocktail party idea level of understanding of the particular subject to explain why a prediction about the subject would fall to the big hammer generally fared poorly, whether or not their favored big ideas were correct. Some examples of "big ideas" would be "environmental doomsday is coming and hyperconservation will pervade everything", "economic growth will

create near-infinite wealth (soon)", "Moore's law is supremely important", "quantum mechanics is supremely important", etc. Another common trait of poor predictors is lack of anything resembling serious evaluation of past predictive errors, making improving their intuition or methods impossible (unless they do so in secret). Instead, poor predictors often pick a few predictions that were accurate or at least vaguely sounded similar to an accurate prediction and use those to sell their next generation of predictions to others.

By contrast, people who had (relatively) accurate predictions had a deep understanding of the problem and also tended to have a record of learning lessons from past predictive errors. Due to the differences in the data sets between this post and Tetlock's work, the details are quite different here. The predictors that I found to be relatively accurate had deep domain knowledge and, implicitly, had access to a huge amount of information that they filtered effectively in order to make good predictions. Tetlock was studying people who made predictions about a wide variety of areas that were, in general, outside of their areas of expertise, so what Tetlock found was that people really dug into the data and deeply understood the limitations of the data, which allowed them to make relatively accurate predictions. But, although the details of how people operated are different, at a high-level, the approach of really digging into specific knowledge was the same.

In comparison with other mechanisms for making sense of future AI developments, forecasting does ok.

Here are some mechanisms that the EA community has historically used to try to make sense of possible dangers stemming from future AI developments:

- Books, like Bostrom's *Superintelligence*, which focused on the abstract properties of highly intelligent and capable agents in the limit.
- Reports by Open Philanthropy. They either try to model AI progress in some detail, like example 1, or look at priors on technological development, like example 2.
- Mini think tanks, like Rethink Priorities, Epoch or AI impacts, which produce their own research and reports.
- Larger think tanks, like CSET, which produce reports like this one on Future Indices.
- Online discussion on lesswrong.com, that typically assumes things like: intelligence gains would be fast and explosive, that we should aim to design a mathematical construction that guarantees safety, that iteration would not be advisable in the face of fast intelligence gains, etc.
- Occasionally, theoretical or mathematical arguments or models of risk
- One-off projects, like Drexler's Comprehensive AI systems

- Questions on forecasting platforms, like Metaculus, that try to solidly operationalize possible AI developments and dangers, and ask their forecasters to anticipate when and whether they will happen.
- Writeups from forecasting groups, like Samotsvety
- More recently, the Forecasting Research Institute's existential risk tournament/experiment writeup, which has tried to translate geopolitical forecasting mechanisms to predicting AI progress, with mixed success.
- Defering to intellectuals, ideologues, and cheerleaders, like Toby Ord, Yudkowsky or MacAskill.

None of these options, as they currently exist, seem great. Forecasting has the hurdles discussed above, but maybe other mechanisms have even worse downsides, particularly the more pundit-like ones. Conversely, forecasting will be worse than deferring to a brilliant theoretical mind that is able to grasp the dynamics and subtleties of future AI development, like perhaps Drexler's on a good day.

Anyways, you might think that this forecasting thing shows potential. Money is not a limitation for you, so…

In this situation, here are some strategies of which you might avail yourself

A. Accept the Faustian bargain

1. Make a bunch of short-term and long-term forecasting questions on AI progress
2. Wait for the short-term forecasting questions to resolve
3. Weight the forecasts for the long-term questions according to accuracy in the short term questions

This is a Faustian bargain because of the reasons reviewed above, chiefly that short-term forecasting performance is not a guarantee of longer term forecasting performance. A cheap version of this would be to look at the best short-term forecasters on the AI categories on Metaculus, and report their probabilities on a few AI and existential risk questions, which would be more interpretable than the current opaque "Metaculus prediction".

If you think that your other methods of making sense of what it's going on are sufficiently bad, you could choose this and hope for the best? Or, conversely, you could anchor your beliefs on a weighted aggregate of the best short-term forecasters and the most convincing theoretical views. Maybe things will be fine?

B. Attempt to do a Bayesianism

Go to the effort of rigorously formulating hypotheses, then keep track of incoming evidence for each hypothesis. If a new hypothesis comes in, try to do some

version of just-in-time bayesianism, i.e., monkey-patch it after the fact. Once you are specifying your beliefs numerically, you can deploy some cute incentive mechanisms and reward people who change your mind.

Hope that keeping track of hypotheses about the development of AI at least gives you some discipline, and enables you to shed untrue hypotheses or frames a bit earlier than you otherwise would have. Have the discipline to translate the worldviews of various pundits into specific probabilities[1], and listen to them less when their predictions fail to come. And hope that going to the trouble of doing things that way allows you to anticipate stuff 6 months to 2 years sooner than you would have otherwise, and that it is worth the cost.

C. Invest in better prediction pipelines as a whole

Try to build up some more speculative and formidable type of forecasting that can deal with the hurdles above. Be more explicit about the types of decisions that you want better foresight for, realize that you don't have the tools you need, and build someone up to be that for you.

[1]Back in the day, Tetlock received a grant to "systematically convert vague predictions made by prominent pundits into explicit numerical forecasts", but I haven't been able to track what happened to it.