Summary

1. I have some intuition that crowd forecasting could be a useful tool for important decisions like cause prioritization but feel uncertain
2. I'm not aware of many example success stories of crowd forecasts impacting important decisions, so I define a simple framework for how crowd forecasts could be impactful:
    a. Organizations and individuals (stakeholders) making important decisions are willing to use crowd forecasting to help inform decision making
    b. Forecasting questions are written such that their forecasts will affect the important decisions of stakeholders
    c. The forecasts are good + well-reasoned enough that they are actually useful and trustworthy for stakeholders
3. I discuss 3 bottlenecks to success stories and possible solutions:
    a. Creating the important questions
    b. Incentivizing time spent on important questions
    c. Incentivizing forecasters to collaborate

Background

1. I've been forecasting actively on Metaculus since Mar 2020, GJO and Foretell since mid-2020
2. I have some intuition that crowd forecasting could be a useful tool for important decisions like cause prioritization but feel uncertain
    a. I feel very uncertain about questions like AI timelines and find it hard to understand how other people seem comparatively confident
    b. I resonate some with The case of the missing cause prioritisation research - EA Forum (albeit I've been following the EA community for less time than OP)
    c. When I say crowd forecasting, I'm referring mostly to platforms like Metaculus, Foretell and Good Judgment but prediction markets like PredictIt share many similar features
3. I've been reflecting recently on what would need to happen to bridge the gap in my mind between "crowd forecasting seems like it could be useful" and "crowd forecasting is consistently useful for important decisions"

Success story

Framework

1. I'll define the impact of forecasting based on how actionable the forecasts are: Forecasts are impactful to the extent that they affect important decisions
2. I'm mostly thinking of important decisions from an EA perspective. Some examples:

    a. What causes should 80,000 Hours recommend as priority paths for people to work in?
    b. How should Open Philanthropy allocate grants within global health & development?
    c. Should AI alignment researchers be preparing more for a world with shorter or longer timelines?
    d. What actions should we recommend the US government take to minimize pandemic risk?
3. Crowd forecasting's success story can be broken down into something like:
    a. Organizations and individuals (stakeholders) making important decisions are willing to use crowd forecasting to help inform decision making
        i. I doubt this is the main bottleneck right now but it may be in the future
            1. e.g. OpenPhil and Rethink Priorities both seem excited about using crowd forecasts to inform important decisions
    b. Forecasting questions get written such that their forecasts will affect important decisions
        i. Bottleneck here: Creating important questions
    c. The forecasts are good + well-reasoned enough that they are actually useful and trustworthy for stakeholders
        i. Bottlenecks here: Incentivizing time spent on important questions and Incentivizing forecasters to collaborate
        ii. The distribution of effort across questions may need to be shifted such that more time is devoted to more important questions.

Examples

Potential positive examples thus far:

1. General sense that COVID predictions on Metaculus made some people take COVID more seriously early, though challenging to find specific sources
2. Metaculus El Paso COVID predictions being useful for El Paso government
3. Good Judgment helping OpenPhil
4. Metaculus + GJO COVID predictions being sent to CDC, see e.g. https://www2.lehigh.edu/news/forecast-the-impacts-of-vaccines-and-variants-on-the-us-covid-trajectory
    a. Did this affect actual decisions?
5. Metaculus tournament supporting the decision-making of the Virginia Department of Health (h/t Charles Dillon):
    a. Would be curious for an impact report here on how concretely decisions were affected

Bottlenecks

Note: These are my current best guesses as to the biggest bottlenecks to the success story. Each bottleneck could and does easily have 1+ posts to itself, I'll aim to give my rough thoughts on each one here.

Creating the important questions

1. In general, I think it's just really hard to create questions on which crowd forecasting will be impactful
2. Current failure modes:
    a. Too much trend extrapolation resulting in questions that are hard to imagine being actionable
        i. See e.g. this complaint regarding Forecasting AI Progress
    b. Not targeting near-term areas of disagreement
        i. Helpful to target near-term areas of disagreement to determine which of various "camp"s' world models are more accurate
        ii. When questions aren't framed around a disagreement between people's models, often hard to know how to update on results
    c. Too long-term / abstract to feel like you can trust without associated reasoning (e.g. AI timelines without associated reasoning)
    d. Requires too much domain expertise / time
        i. I mostly conceptualize crowd forecasting as an aggregator of opinions given existing research for important questions, since ideally independent of crowd forecasting there is rigorous research going on in the area.
            1. Example: People can read about Ajeya and Tom's AI forecasts and aggregate these + other considerations into an AI timeline prediction
            2. But if a question requires someone to do anywhere close to the work of a report from scratch, unrealistic to expect this to happen
        ii. My sense is that for a lot of questions on Metaculus forecasters don't really have enough expertise to know what they're doing for some questions
3. Solution ideas
    a. Experiment with best practices for creating impactful questions. Some ideas:
        i. Early warning signs for e.g. pandemics, nukes
        ii. Orgs decompose decision making then release either all questions or some sub-questions to crowd forecasting
            1. For big questions, some sub-questions will be better fit for crowd forecasting than others; good for orgs to have an explicit plan on how forecasts will be incorporated into decision-making
                a. This may already be happening to some extent but transparency seems good + motivating for forecasters

2. Metaculus Causes like Feeding Humanity in collaboration with GFI is a great step in the right direction in terms of explicitly aiding orgs' decision-making
    a. Ideally there'd be more transparency about which questions will affect GFI's decision-making in which ways

iii. Idea for question creation process: double crux creation
1. E.g. I get someone whose AGI median is 2035, and someone whose AGI median is >2100, and try to find double cruxes as near-term as possible for question creation
2. Get signal on whose world model is more correct, + get crowd's opinion on as concrete a thing as possible
3. Related: decomposing questions into multiple smaller questions to find cruxes
    a. Foretell has done this with stakeholders in order to create an issue campaign which I'm excited to track the impact of

iv. More research into conditional questions like the Possible Worlds Series

b. Write up best practices for creating impactful questions
i. https://www.metaculus.com/question-writing/ is a good start but I'd love to see a longer version with more focus on the **Explain why your question is interesting or important** section, specifically **Explain what decisions will be affected differently by your question**

c. Incentivize impactful question creation
i. Leaderboards
ii. Monetary incentives

Incentivizing time spent on important questions

1. Current failure modes
    a. Intuition: Some questions are OOMs more important than others but don't get nearly OOM more effort on crowd forecasting platforms
    b. [2101.01816] Incentive-Compatible Forecasting Competitions contains relevant content
    c. Metaculus incentive issue: incentive to spend a little time on lots of questions
        i. I spent several months predicting and updating on every Metaculus questions resolving within the next 1.5-2 years to move up the sweet Metaculus leaderboard
            1. On reflection, I realized I should do deep dives into fewer impactful questions, rather than speed forecasting on many questions
        ii. Important questions like Neil's (action relevant for RP) don't get extra attention deserved
    d. Incentive issue on most platforms: performance on all questions is weighted approximately equally, despite some questions seeming much more impactful than others
2. Solution ideas
    a. Give higher weight on leaderboards or cash prizes to more important questions

b. Metaculus moving in the right direction with Forecasting Causes with prizes, OpenPhil AI tournament
    i. I'd like to see things move further in this direction
    ii. Further incentive changes to give both monetary prizes + internet points to people who focus lots of time on important questions
c. Foretell itself has a clearer area / narrower focus than Metaculus and GJOpen
d. Good Judgment model: organizations pay for forecasts on particular questions
    i. Has strengths and weaknesses, willingness-to-pay is a proxy for impact but ideally there can also be focus on questions for which good forecasts is a public good
e. Ideally, want to incentivize people to do detailed forecasts on important questions like read Ajeya's draft timelines report and update their AI timelines predictions based on that, writing up their reasoning for others to see, which brings me to...

Incentivizing forecasters to collaborate

1. Current failure modes
    a. See [2101.01816] Incentive-Compatible Forecasting Competitions for why collaboration isn't incentivized currently
    b. My personal experience with this
        i. Noticed that when I shared my reasoning community would trend toward my prediction, so it was better to be silent
        ii. I like to think of myself as altruistic but am also competitive, and similar to how I got sucked in by the Metaculus leaderboard to predict shallowly on many questions, I got sucked in to not sharing my reasoning
    c. Example: Metaculus AI tournament
        i. No one left any comments despite the importance of the topic, then when they required 3 comments each everyone left very brief, uninformative comments
        ii. See discussion here: an important point is that this problem gets worse as monetary incentives are added, which may be needed as part of a solution to incentivizing time spent on important questions
2. Solution ideas
    a. Collaborative scoring rules such as Shapley values
    b. Prizes for insightful comments
    c. More metrics taking into account comment upvotes
    d. Prizes for comments which affect others' forecasts

Alternative impact model

On my current mental model, crowd forecasting will have something like a power law distribution of impact, where it has the most impact from affecting a relatively small amount of important

decisions. An alternative impact model is that crowd forecasting will “raise the sanity waterline” and its impact will be more diffuse and hard to measure, making many relatively low impact decisions better.

Some reasons I’m skeptical of this alternative model:

1. Some decisions seem so much more important than others.
2. There’s a fairly large fixed cost to operationalizing a question well and getting a substantial amount of reasonable forecasters to forecast on it, that I’m not very optimistic about reducing, which makes me skeptical about the diffuse, tons-of-questions model of impact.
3. I have trouble thinking of plausible ways most Metaculus questions will impact decisions that are even low impact.

Conclusion

These are my best guesses as to the success story for crowd forecasting, biggest current bottlenecks, and solution ideas. I’d be excited to see further work or thoughts on any of these components. Specifically:

1. Does my success story make sense? Is it missing something important?
2. Do my bottlenecks seem like the biggest bottlenecks?
    a. Would be good to hear the perspective of forecast stakeholders in addition to forecasters
3. Are there solution ideas I’m missing?